
Article:
Gao, F., Earnest, A., Matchar, D.B. et al. (2 more authors) (2015) Sample size calculations for the design of cluster randomized trials: A summary of methodology. Contemporary Clinical Trials, 42. 41 - 50.

https://doi.org/10.1016/j.cct.2015.02.011

Sample size calculations for the design of cluster randomized trials: a summary of current methodology

Fei Gao[a,b], Arul Earnest[b], David B Matchar[c], Michael J Campbell[d] and David Machin[d,e,*]

[a] National Heart Research Institute Singapore, National Heart Centre Singapore, 5 Hospital Drive, Singapore 169609: gao.fei@nhcs.com.sg

[b] Center for Quantitative Medicine, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857: arul.earnest@hotmail.com

[c] Health Services & Systems Research, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857: davidmatchar@duke-nus.edu.sg

[d] Medical Statistics Group, School of Health and Related Research, University of Sheffield, Regents Court, 30 Regent Street, Sheffield S1 4DA, UK: m.j.campbell@sheffield.ac.uk

[e] Department of Cancer Studies and Molecular Medicine, Clinical Sciences Building, University of Leicester, Leicester Royal Infirmary, Leicester LE2 7LX, UK: dm113@le.ac.uk

*All Pre-Publication Correspondence to:

David Machin, Poachers Cottage, Southover, Frampton, Dorset DT2 9NQ, UK. dm113@le.ac.uk Telephone: (+44) 1300 321 113.

*All Post-Publication Correspondence to Senior author:

Gao Fei, National Heart Research Institute Singapore, National Heart Centre Singapore, 5 Hospital Drive, Singapore 169609. E-mail: gao.fei@nhcs.com.sg; Telephone: (+65) 6704 2245; Fax: (+65) 6844 9056

Abstract

Cluster randomized trial designs are growing in popularity in, for example, cardiovascular medicine research and other clinical areas and parallel statistical developments concerned with the design and analysis of these trials have been stimulated. Nevertheless, reviews suggest that design issues associated with cluster randomized trials are often poorly appreciated and there remain inadequacies in, for example, describing how the trial size is determined and the associated results are presented. In this paper, our aim is to provide pragmatic guidance for researchers on the methods of calculating sample sizes. We focus attention on designs with the primary purpose of comparing two interventions with respect to continuous, binary, ordered categorical, incidence rate and time-to-event outcome variables. Issues of aggregate and non-aggregate cluster trials, adjustment for variation in cluster size and the effect size are detailed. The problem of establishing the anticipated magnitude of between- and within-cluster variation to enable planning values of the intra-cluster correlation coefficient and the coefficient of variation are also described. Illustrative examples of calculations of trial sizes for each endpoint type are included. [Word count: ~~184~~175]

Key Words: cluster randomized trial; sample size;

INTRODUCTION

In contrast to clinical trials in which individual subjects are each randomized to receive one of the therapeutic options or interventions under test, the distinctive characteristic of a cluster trial is that specific groups or blocks of subjects (the clusters) are first identified and these units are assigned at random to the interventions. The term "cluster" in this context may be a household, school, clinic, care home or any other relevant grouping of individuals. When comparing the interventions in such cluster randomized trials, account must always be made of the particular cluster from which the data item is obtained.

A large and ever increasing number of cluster randomized trials have been conducted or are underway covering many aspects of cardiovascular related medicine. These include trials of cardiovascular guidelines **[1]**, prescribing practice **[2]**, community health awareness **[3]**, breast feeding promotion on cardiometabolic risk factors in childhood **[4]**, the effectiveness of a multifactorial intervention to improve both medication adherence and blood pressure control and to reduce cardiovascular events **[5]**, and improving outcomes in patients with left ventricular systolic dysfunction **[6].** In the TEACH trial of local pharmacy support **[7]**, the clusters were the local pharmacists of patients with heart failure (HF) who had been hospitalised and then discharged into the community. The plan was that clusters were each randomized to one of the two interventions on a 1:1 basis: CONTROL or PHARM. Those pharmacists allocated PHARM would give their patients additional educational (motivational) support. Hence, all the patients within a particular cluster received the same intervention. A patient experiencing any one of a readmission, emergency room visit or mortality due to HF was regarded as a failure.

There are numerous publications describing design, analysis and reporting issues concerned with cluster randomized trials, including text books **[8-11]**. However much of the literature is fragmented and some quite old (though still relevant). Further some of the articles are quite technical in nature so investigators may find it difficult to determine best practice. A review of cluster trials **[12]**, published subsequent to the 2004 extension of the CONSORT guidelines **[13-14]**, concluded that the methodological quality of cluster trials often remains suboptimal.

To facilitate and improve this situation, we focus on methods of determining the number of subjects (and clusters) required with the aim to provide a compact but comprehensive reference for those designing cluster trials. [Word count: 390]

GENERAL DESIGN CONSIDERATIONS

Individually randomized trials – continuous outcome measure

At the close of a clinical trial, and once all the data collection is complete, a comparison will be made between the interventions with respect to the primary endpoint. For the case of two interventions, Standard (S) and Test (T), with n_S and n_T patients respectively randomized individually to each, the statistical process for a continuous outcome measure, y, is made by comparing the corresponding means $\bar{y}_S$ and $\bar{y}_T$ by use of Student's t-test. This tests the null hypothesis that the difference $\delta = \mu_T - \mu_S = 0$, where μ_S and μ_T are the true or population means of interventions S and T. If the null hypothesis is rejected then we conclude μ_S and μ_T differ.

However, prior to this analysis, the trial must first be designed and conducted. In general, critical decisions to be made by the design team are the choice (and number) of interventions to compare and the endpoint measure which will be used for the evaluation.

A vital detail is the difference in the outcome (the effect size or δ_{Plan}) between the randomized interventions which might be anticipated. Such a difference should be one (if established) of sufficient clinical importance to justify the expense of conducting the planned trial and likely to lead to changes in clinical practice. Also required is the standard deviation (SD), σ_{Plan}, of the endpoint variable of concern. A further design option is the choice of the ratio of subjects 1: φ allocated to S and T respectively (see below).

Once these aspects are provided, the numbers of subjects to be randomized to each intervention for a continuous endpoint is **[15]**:

$$n_S = \left(\frac{1+\varphi}{\varphi}\right)\frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{(\delta_{Plan}/\sigma_{Plan})^2} + \left[\frac{z_{1-\alpha/2}^2}{2(1+\varphi)}\right], n_T = \varphi n_S \qquad (1)$$

giving a total $N = n_S + n_T$.

Here, α is the Type I error which is the required probability of rejecting the null hypothesis when falsely declaring 'A difference' between the interventions and ~~when~~ 'No difference' is present. The Type II error β corresponds to the probability of falsely accepting the null hypothesis of 'No difference' when the actual size of the difference is indeed δ_{Plan}. The quantity $1 - \beta$ is known as the power.

Further, $z_{1-\alpha/2}$ and $z_{1-\beta}$ are values with probabilities of $\alpha/2$ and β respectively in the upper tail of the standard Normal distribution. Typically $\alpha = 0.05$ leading to $z_{1-0.05/2} = z_{0.975} = 1.9600$ while $\beta = 0.2$ or 0.1 leading to $z_{0.8} = 0.8416$ and $z_{0.9} = 1.2816$ respectively.

The final term $\left[\frac{z_{1-\alpha/2}^2}{2(1+\varphi)}\right]$ in equation (1) applies only when the sample size is small. However, when $\alpha = 0.05$ and $\varphi = 1$, this implies adding $\left[\frac{1.96^2}{2\times(1+1)}\right] = \frac{3.8416}{4} \approx 1$ unit extra to each intervention group.

An alternative is first to assume $\varphi = 1$ in (1), to obtain n subjects for each intervention and then calculate the final numbers per intervention using

$$n_S = \frac{n(1+\varphi)}{2\varphi}, n_T = \frac{n(1+\varphi)}{2}. \qquad (2)$$

This increases the initial total number of subjects N from 2n to $\frac{n(1+\varphi)^2}{2\varphi}$ which, if $\varphi = 0.5$, implies N = 2.25n. If, as we will be concerned with later, it is the number of clusters that is being calculated then k, k_S, k_T and K replace the corresponding n's.

Cluster randomized trials

When the randomised allocation applies to the clusters, the basic principles for sample size calculation still apply although modifications are required. To illustrate these we first describe the t-test, for comparing two means from a non-cluster design, using linear regression terminology with intercept μ_S and slope δ, that is,

$$y_j = \mu_S + \delta x + \varepsilon_j, \qquad (3)$$

where the subjects concerned are j = 1, 2, ... , N; x = 0 for S and x = 1 for T. Further ε_j is a random variable with mean zero and, within each intervention group, assumed to have the same variance, σ^2. If this regression model is fitted to the data then $d = \bar{y}_T - \bar{y}_S$ estimates $\delta = \mu_T - \mu_S$ and the null hypothesis remains $\delta = 0$. However the analysis must now take account of the cluster to which an individual subject belongs. When we compare two interventions N (= $n_S + n_T$) patients will be recruited who will come from clusters of size m with therefore $k_S = n_S/m$, $k_T = n_T/m$ and $K = k_S + k_T$ clusters in total.

To allow for the clusters, model (3) is extended to:

$$y_{ij} = \mu_S + \delta x + \gamma_i + \varepsilon_{ij}. \qquad (4)$$

Here the clusters are i = 1, 2, … , K and the subjects j = 1, 2, … , m in each cluster. The cluster effects, γ_i, are assumed to vary at random within each intervention about a mean

of zero, with a variance, $\sigma^2_{Between-Cluster}$. Further the ε_{ij} are also assumed to have mean zero but with variance, $\sigma^2_{Within-Cluster}$ and both random variables, γ and ε, are assumed to be Normally distributed. The combined sum of the within- and between-cluster variances, $\sigma^2_{Total} = \sigma^2_{Between-Cluster} + \sigma^2_{Within-Cluster}$.

Although the format of the (random-effects) model (4) will change depending on the outcome measure of concern, all will contain random terms accounting for the cluster design.

The sample size formula (1) is essentially determined as a consequence of model (3) while the formulae which follow for cluster trials, are based on (4).

CLUSTER DESIGN CONSIDERATIONS

Intra-cluster correlation coefficient (ICC) – continuous outcome

A feature of all cluster trials is that subjects recruited from within the same cluster cannot be regarded as acting independently of each other in terms of their response to the intervention received. The magnitude of this within-cluster dependence, which ultimately influences the eventual trial size, is quantified by the intra-cluster correlation coefficient (ICC), ρ, which is interpreted in a similar way to Pearson correlation.

With each subject in every cluster providing an outcome measure, the ICC is the proportion of ~~the combined sum of the within- and between-cluster variances~~ σ^2_{Total} accounted for by the between-cluster variation, that is

$$\rho = \frac{\sigma^2_{\text{Between-Cluster}}}{\sigma^2_{\text{Total}}} = \frac{\sigma^2_{\text{Between-Cluster}}}{\sigma^2_{\text{Between-Cluster}} + \sigma^2_{\text{Within-Cluster}}}. \qquad (5)$$

Thus, since variances cannot be negative, the ICC cannot be negative. A major challenge in planning the sample size is identifying an appropriate value for ρ. In practice, estimates of ρ are usually obtained from previously reported trials using similar

randomization units and outcome measures. The values of ρ arising in a primary care setting tend to vary from 0.01 to 0.05 with a median value quoted as 0.01 **[16].** Larger ICCs have been reported **[17]** although for community intervention trials they are typically < 0.01 **[8].** ~~Our suggestion is to try different values of the ICC and investigate how sensitive the sample size estimate is to these changes.~~

The Design Effect (DE)

The impact of the ICC, on the planned trial size, will depend on its magnitude and on the number of subjects recruited per cluster, m, through the so-called design effect (DE),

$$DE = 1 + (m - 1)\rho. \qquad (5X)$$

The DE is then multiplied by the sample size obtained from (say) equation (1) to give that required for a cluster design. In practice there may be substantial variation in m from cluster to cluster and to allow for such variation DE becomes: **[18]**

$$DE = 1 + \left(\bar{m} + \frac{[SD(m)]^2}{\bar{m}} - 1\right)\rho, \qquad (6)$$

where $\bar{m}$ is the anticipated mean cluster size, and $SD(m)$ the corresponding standard deviation. As the value of DE depends on ρ, whose value may not be firmly established, our suggestion is to try different values of the ICC and investigate how sensitive the sample size estimates are to these changes. In all situations, DE will be ≥ 1 since m and $\bar{m} > 1$ and $\rho \geq 0$.

~~Another option~~ Van Breukelen and Candel [19] have suggested that if m varies the above approach is conservative and a better method is to adjust the total number of clusters initially planned, K, to:

$$\mathrm{K}_{\mathrm{Adjusted}} = \frac{\mathrm{K}}{\{1-[\mathrm{CV(m)}]^2\,\xi(1-\xi)\}}, \qquad (7)$$

where the coefficient of variation, $CV(m) = \frac{SD(m)}{\bar{m}}$ and $\xi = \frac{\bar{m}\rho}{\bar{m}\rho + (1-\rho)}$. ~~[19].~~

However, since $0 < \xi < 1$, the maximum possible value of $\xi(1-\xi) = \frac{1}{4}$ and so the largest adjustment to K corresponds to $1/\{1 - [CV(m)]^2/4\}$. Further since the CV(m) is usually less than 0.7 the inflation of K necessary to allow for varying m is at most about 14%. If CV(m) = 0.35 then the inflation is at most 3%.

In general ~~practice~~, variation in m results in an increase in total sample size for the trial and equation (6) tends to overestimate the required sample size . ~~Hence whether to use equation (6) or (7) requires some judgement. For example, if SD(m) is large it might be quite difficult to raise the number of subjects recruited in each cluster to ensure $\bar{m}$ is increased and thereby the requisite (new) total sample size achieved. In which case, increasing K through equation (7) may be the most practical option. In contrast~~, In situations where increasing $\bar{m}$ is feasible, equation (6) may suffice.

Control hypertension and hypercholesterolemia [20]

To illustrate the impact of varying cluster size on the DE, we use the results from STITCH2 trial which includes the precise number of clusters within each intervention and the number of subjects recruited per cluster. Table 1 shows that cluster size varied considerably from 2 to 47. For both interventions combined, CV(m) = 15.29/26.43 = 0.59 and this magnitude is not atypical.

Table 1. Number of clusters and the corresponding CV(m) of cluster size by ~~care~~ intervention group of the STITCH2 trial (data from Dresser, et al. (2013) [20])

	~~Care~~Intervention	N(m)	Min(m)	$\bar{m}$	Max(m)	SD(m)	CV(m)	

	Guideline	20	2	28.75	47	15.59	0.54	
	STITCH2	15	2	23.33	45	14.83	0.64	
	Total	35	2	26.43	47	15.29	0.59	

If m = 26 is taken as the anticipated cluster size in a trial, then from equation (5X), DE = 1 + 25ρ, whereas if information from Table 1 is used equation (6) gives $DE = 1 + \left(26.43 + \frac{15.29^2}{26.43} - 1\right)\rho$ = 1 + 34.3ρ which is clearly larger.

Community Based Exercise Programme [21]

In contrast, a community based exercise programme in over 65 year olds involving K = 12 practices recruited a mean of $\bar{m}$ = 535 individuals from each with SD(m) = 139.9 to give a much lower figure of CV(m) = 0.26.

When planning a new trial, investigators may be guided by results such as these.

Potential attrition

For many different reasons, the eventual numbers of clusters and/or subjects recruited may be less than ~~that~~ those planned.

Clearly, the loss of all information from a cluster has greater impact than the loss of (few) patients within a cluster. Thus, as a precaution, the initial number of clusters, K, indicated by the preliminary sample size calculations may need to be increased. Relevant experience of the design group, or reference to published studies reporting such losses,

may provide guidance on the extent of potential loss. Any change in K may lead the design team to reconsider m, N or both

~~So,~~ In anticipation of ~~this~~ possible subject attrition, the initial plan for the trial size, N, can be inflated by the division of the proportion, θ, of subjects recruited for whom the endpoint measure is likely to be recorded. However this tends to overestimate the sample size required. An alternative is first to modify the DE using mθ in place of m, obtain the resulting sample size by the appropriate means, and then divide by θ again. This process tends to underestimate the sample size required since it is unlikely that the drop-out rate is equal in each cluster and so the cluster size will vary, which results in loss of power [11, p67]. Consequently, a compromise sample size mid-way between the two approaches may be sought. Any change in N may lead the design team to ~~alter~~ reconsider m, K or both.

The allocation ratio, φ

Although the majority of clinical trials involve equal allocation to each of the interventions, there may be circumstances in which the proportions may differ. Thus, in the case of two interventions, n_S patients may be allocated to S, while n_T ($\neq n_S$) are allocated to T. In ~~which~~ that case, the allocation ratio, φ, is set so $n_T = \varphi n_S$ and $N = n_S(1+\varphi)$ is the planned total trial size. In general, as φ moves away from unity, the required sample size will increase.

Community Based Exercise Programme [21]

In this trial, 8 clusters for Control and 4 for Test were used for evaluating the programme as budgetary constraints limited the number of Test facilities (clusters) available whereas:

" … the relative costs of including controls were very small … ". Thus instead of using a 1:1 design, with 4 clusters, per intervention, a 2:1 allocation using 12 clusters enabled a larger trial with greater power to be conducted without increasing the number of T clusters, k_T.

Non-aggregate and Aggregate designs

A 'non-aggregate' design uses the individual observations as the unit of analysis and so ~~the regression models may be extended to account for individual covariate values such as, for example, subject gender, age, disease severity or a pre-randomisation (baseline) measure of the chosen endpoint for the trial. T~~ the analysis will be based on ~~an extension of~~ the random effects regression model (4). The objective of the sample size calculation is to determine the appropriate number of subjects required per intervention, n_S and n_T and the number of clusters is determined on division by the anticipated number of subjects per cluster.

An 'aggregate' ~~or 'field'~~ design is one in which a summary measure from each cluster is obtained. For example with continuous data, and n = mk subjects in each intervention, the trial will provide k cluster means $\bar{y}_{01}, \bar{y}_{02}, \ldots, \bar{y}_{0k}$ (each based on m observations) with the mean of these means for intervention x = 0 compared with those from x = 1. In the situation when k is small, the analysis may use the t-test (rather than the z-test) and so the sample size calculations will be based on equation (1) but replacing the anticipated SD, σ_{Plan}, by the anticipated standard deviation of the summary measure, such as the mean. This will usually be available from prior studies and obviates the need for a design effect.

In this situation, m is fixed by the design team and the calculation provides the required number of clusters, K = 2k.

However, although the analysis for both designs appears to be the same the number of degrees of freedom, df, differ. For an aggregate design, with fixed cluster size m, $df_{Aggregate}$ = {N/m} – 2 while for a non-aggregate design, $df_{Non\text{-}Aggregate}$ = {N/DE} – 2 = {N/[1+(m – 1)ρ]} – 2 which is ~~smaller~~ larger except in the improbable situation of ρ = 1. This means that correcting for small sample size has a greater effect on aggregate designs since the df will also be smaller.

SAMPLE SIZE FOR CLUSTER TRIALS

Continuous endpoint

Non-aggregate design

For the model of equation (4) the variance for individuals in each cluster within each intervention group is assumed the same, and of the form:

$$Var(y_{ij}) = \sigma^2_{Between-Cluster} + \sigma^2_{Within-Cluster} = \sigma^2_{Total} \qquad (8)$$

If planning values for $\sigma_{\text{Between-Cluster}}$ and $\sigma_{\text{Within-Cluster}}$ can be provided by the design team, then the planned values for σ_{Total} (denoted σ_{Plan}) and the ICC can be obtained from equation (5). Alternatively values for the ICC may be obtained from previous experience. In either case, although m needs to be pre-specified, DE can be determined and equation (1) is modified to become:

$$n_S = DE \times \left(\frac{1+\varphi}{\varphi}\right)\frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{(\delta_{Plan}/\sigma_{Plan})^2},\ n_T = \varphi n_S. \qquad (9)$$

The total sample size, N = n_S + n_T = Km and it follows that the required numbers of clusters are $k_S = \frac{K}{1+\varphi}$ and $k_T = \frac{\varphi K}{1+\varphi}$. Note that the final term of equation (1), $\left[\frac{z^2_{1-\alpha/2}}{2(1+\varphi)}\right]$, is omitted here as N is likely to be large in most circumstances.

Daily exercise and Quality of Life [22]

Suppose a design team is planning a confirmatory non-aggregate cluster randomized trial, based on the one previously published, to see if a daily exercise regime delivered by personal trainers at the suggestion of their general practitioner for a year (T) would lead to improved quality of life compared to no intervention (S) in older men. The primary outcome is the Physical Function (PF) score of the SF-36 measure at 1-year. Previous experience from a cross-sectional (non-cluster) study suggests that such men have a mean score of 66.4 units, with σ_{PF} = 29.5 units and the effect of the daily exercise regime would be considered important if it increased the PF by at least 10 units.

These lead to planning values δ_{Plan} = 10 and σ_{Plan} = 29.5. However, due to the high cost of providing personal trainers, the new design team decide on a 3:2 randomization in favour of the S group, that is φ = 2/3 = 0.6667. This implies that K will have to be a multiple of 5 if the clusters are to be randomized in this ratio. Previous experience suggests that an achievable cluster size is m = 30 and ρ = 0.01 so that DE = 1 + [(30 – 1) × 0.01] = 1.29. Further the investigators set a two-sided α = 0.05, β = 0.1, and so the sample size required from equation (9) is:

$$n_S = 1.29 \times \left(\frac{1+0.6667}{0.6667}\right) \times \frac{(1.96+1.2816)^2}{(10/29.5)^2} = 1.29 \times 228.60 = 294.89$$ or 295 and n_T = 0.6667 × 295 = 196.67 or 197. The planned total sample size is N = 295 + 197 = 492 men. Then with m = 30, this implies K = 492/30 = 16.4 or 17 clusters. If the investigators set k_S = 10 and k_T = 7 then the ratio 10:7 is not dissimilar to 9:6 the stipulated randomization ratio of 3:2.

If the approach of (2) had been used then, for equal allocation n = 236, so that for φ = 0.6667, $n_S = \frac{236\times(1+0.6667)}{2\times0.6667}$ = 236 ×1.250 = 295 and $n_T = \frac{236\times(1+0.6667)}{2}$ = 236 ×0.833 = 197 as previously obtained.

Aggregate design

If an aggregate design is considered, then the summary mean, $\bar{y}_i$, calculated from the m subjects within the cluster i, is the endpoint of concern. In this case,

$$Var(\bar{y}_i) = \sigma^2_{Between-Cluster} + \frac{\sigma^2_{Within-Cluster}}{m}, \qquad (10)$$

which can be alternatively expressed more compactly as

$$Var(\bar{y}_i) = \frac{DE\sigma^2_{Total}}{m}. \qquad (11)$$

The sample size now refers to the number of clusters required. However, equation (9) is still used but with k_S, k_T and K replacing n_S, n_T and N. Further, as the resulting number of clusters may be small, the comparison of means between the interventions will be made using the t-test. In which case, $z_{1-\alpha/2}$ and $z_{1-\beta}$ from the Normal distribution should be replaced by the corresponding quantities for the t-distribution. However, these values depend on the degrees of freedom (df) which are K – 2. At the preliminary stage we do not know K. So the process begins by estimating k_S (and k_T) using equation (9) to obtain an initial value say K_0 for the required total number of clusters, so $df_0 = K_0 - 2$. Tables of the t-distribution give the values $t_{df0,1-\alpha/2}$ and $t_{df0,1-\beta}$ which are substituted for $z_{1-\alpha/2}$ and $z_{1-\beta}$ in equation (9) to obtain a revised total number of clusters, K_1. If this is different from K_0, the degrees of freedom are recalculated as $df_1 = K_1 - 2$ and the process repeated until the value of K stabilizes.

In practice in 1:1 (φ = 1) randomization designs, if $K_0 \geq 20$, the process is unlikely to be required while if $K_0 < 20$ the refinement generally leads to 2 more clusters being added. Thus, the process of refining the sample size in this way is effectively the same as the simpler one of including the final term of equation (1), which is $\left[\frac{z^2_{1-\alpha/2}}{2(1+\varphi)}\right]$.

Daily exercise and Quality of Life [22]

Had the previous example been designed as an aggregate cluster trial, then with the information provided as σ_{Total} = 29.5 and ρ = 0.01, equation (5) can be used to obtain $\sigma^2_{Between-Cluster} = \rho \times \sigma^2_{Total}$ = 0.01 × $(29.5)^2$ = 8.70 and $\sigma^2_{Within-Cluster} = \sigma^2_{Total} - \sigma^2_{Between-Cluster}$ = $(29.5)^2$ – 8.70 = 861.6. Hence, if we assume m = 30, equation (10) gives $\sigma_{Plan} = \sqrt{8.70 + \frac{861.6}{30}} = 6.12$ and, from equation (1) with φ = 2/3, $k_S = \left(\frac{1+0.6667}{0.6667}\right) \times \frac{(1.96+1.2816)^2}{(10/6.12)^2} = 9.84$ or 10 clusters and $k_T = \varphi k_S = 0.6667 \times 10 = 6.7$ or 7 clusters giving a total of K = 17 in all as in the non-aggregate design.

However, as the number of clusters is relatively small, including the final term of equation (1), adds $\frac{z^2_{1-\alpha/2}}{2(1+\varphi)} = \frac{1.96^2}{2(1+0.6667)} = 1.15$ or about 1 cluster per intervention to give K = 19.

If the approach of (2) had been used then, with equal allocation k = 8.82, so that for φ = 0.6667, $k_S = \frac{8.82 \times (1+0.6667)}{2 \times 0.6667}$ = 11.03 or 12 and $k_T = \frac{8.82 \times (1+0.6667)}{2}$ = 7.35 or 8 to give K =20.

Binary outcome

Non-aggregate design

For a binary outcome the dependent variable y_{ij} of equation (4) only takes the values 0 or 1 with the probability π_i that $y_{ij} = 1$ and which is assumed constant for each subject within a cluster. Such data are analysed using a random effects logistic regression model in which, because of the clusters, γ is retained but is assumed to come from either a Normal or Beta distribution, while ε is assumed to come from a Binomial distribution. **[23]**

In order to calculate a sample size, the anticipated proportions responding in each intervention group, π_S and π_T need to be anticipated, from which $\delta_{Plan} = \pi_T - \pi_S$. It is usual to assume that

$$\sigma_{Plan} = \sqrt{\bar{\pi}(1-\bar{\pi})}\,, \qquad (12)$$

where $\bar{\pi} = \frac{\pi_S + \pi_T}{2}$.

Also required is the intra-class correlation, ρ_{Binary}, for use in the expression DE of equation (5). This can be obtained from

$$\rho_{Binary} = \frac{\sigma^2_{Between-Cluster}}{\sigma^2_{Total}} = \frac{\sigma^2_{Between-Cluster}}{\bar{\pi}(1-\bar{\pi})}\,. \qquad (13)$$

Finally the sample size for this situation is calculated from equation (9) but using the effect size, $\delta_{Plan} = \pi_T - \pi_S$, and σ_{Plan} as specified in equation (12).

Control hypertension and hypercholesterolemia [20]

If a ~~similar~~ non-aggregate trial is planned on the basis of STITCH2, assuming m = 50 subjects will be included from each general practitioner (GP), then planning values for S, the proportion achieving target, might be assumed as 0.40 while that for T as 0.52. From

these, $\delta_{\text{Plan}} = 0.52 - 0.40 = 0.12$, $\bar{\pi} = \frac{0.52+0.40}{2} = 0.46$ and $\sigma_{Plan} = \sqrt{0.46(1-0.46)} =$ 0.4984. Further assuming $\rho_{\text{Binary}} = 0.0620$~~0.071~~ (derived from results of the STITCH trial using equation (14) below) equation **(5X)** gives DE = [1 + (50 – 1) × 0.062] = 4.038. Finally from equation (9), with two-sided $\alpha = 0.05$ and $\beta = 0.2$, $n_S = n_T =$ 4.038~~[1 + (50 – 1) × 0.071]~~ × $\frac{(1+1)}{1}\frac{(1.96+0.8416)^2}{(0.12/0.4984)^2}$ = ~~1212.9~~1093.5 implying N = ~~2426~~2188 subjects in total. The corresponding number of clusters K = ~~2426~~2188/50 = ~~49~~44 which ~~may be increased to 50 to~~ allows a 1:1 randomization.

The preliminary estimate of sample size may have to be revised to account for possible variation in cluster size, non-participation of some of the clusters, and/or reduced numbers of individuals completing the assessments.

Control hypertension and hypercholesterolemia [20]

Thus, as was the case in the original STITCH2, if the number recruited per GP was likely to vary then a conservative application of equation (7) would lead to increasing the number of GPs by 14%. Hence, the number of clusters becomes K = ~~49~~44 × 1.14 = ~~55.9~~50.2 or ~~56~~52 ~~in practice~~. The number of patients is thereby increased to N = ~~56~~52 × 50 = ~~2,800~~2,600. l

Equally, although 52 GP were identified for STITCH2, as only 44 (85%) eventually participated in the trial this implies that in future trials the initial planning number of clusters might be increased ~~by 15% suggesting here that to 56 × 1.15 ≈ 64~~ to include 52 / 0.85 ≈ 62 GPs and hence 62 × 50 = 3,100 patients.

Further, as a precautionary measure to account for potential patient (not cluster) loss, ~~an approximate 15% (θ = 0.85) increase in this number may be adopted, in which case N = 2,800/0.85 ≈ 3,300 may be targeted.~~, it might be assumed that only 90% of patients will comply, so a commensurate increase in the number to recruit to N = 3,100/0.90 ≈ 3,444.4 may be considered.

Alternatively, the DE itself may be adjusted to give a reduced value as $DE = \{1 + [(50 \times 0.9) - 1] \times 0.062\} = 3.728$ replacing 4.038 of the preliminary calculations. Thus the revised number of subjects becomes $3{,}100 \times \frac{3.728}{4.038} = 2{,}862.0$ which is smaller than the first revised method estimate of 3,444.4. A compromise suggests that about 3,150 patients are required.

Further discussion by the design team may then suggest 52 GPs should be approached, from which 55 patients would be recruited to the trial.

~~Should all three possibilities have to be accounted for then the planned trial size may be designed to recruit up to (say) 4,000 individuals with consequent increases in either, m, K or both.~~

Aggregate design

In an aggregate design, each cluster within each intervention provides a single proportion, p_{xi}, calculated from the m patients within that cluster. These proportions correspond to the $\bar{y}_{xi}$ of the continuous measure situation albeit now confined to values between 0 and 1.

The corresponding variance of each cluster proportion, p_{xi}, is:

$$Var(p_{xi}) = \sigma^2_{Between-Cluster} + \frac{\bar{\pi}(1-\bar{\pi})}{2m}. \qquad (14)$$

Control hypertension and hypercholesterolemia [20]

The report of the STITCH2 states: "The primary analysis compared the proportion of participants achieving targets [for example, specified blood pressure levels] between the two treatment groups using a two-sample t-test at the level of the cluster … ". This clearly indicates an aggregate design was planned and that the individual cluster proportion is regarded as a continuous outcome. In addition, they specify, $\sigma_{\text{Total}} = 0.15$ which is taken as σ_{Plan} and by including the simple adjustment in equation (1), the number of clusters required are $k_T = k_S = \frac{1+1}{1}\frac{(1.96+0.8416)^2}{(0.12/0.15)^2} + \frac{1.96^2}{2(1+1)} = 25.49$ or 26 per intervention and K = 52.

Further, using equation (14), this leads to $\sigma^2_{Between-Cluster} = 0.15^2 - \frac{0.46(1-0.46)}{2\times 50} = 0.0200$ ~~0.017. Further~~ and from equation (13), $\rho_{\text{Binary}} = 0.0200$~~0.017~~$/(0.46 \times 0.54) = 0.062$ ~~0.071~~ as we had noted earlier. ~~By including the simple adjustment in equation (1), the number of clusters required are $k_T = k_S = \frac{1+1}{1}\frac{(1.96+0.8416)^2}{(0.12/0.15)^2} + \frac{1.96^2}{2(1+1)} = 25.49$ or 26 per intervention and K = 52.~~

Ordinal outcome

Non-aggregated design

In some situations a binary outcome may be extended to comprise an ordered categorical variable of G (>2) levels. In which case the two interventions are compared using ordered logistic regression. Further, only non-aggregate designs are likely as individual cluster summaries (needed for an aggregate design) take the form of a G-level tabulation rather than a single measure.

In principle, if the underlying measure is categorical then any comparisons between groups will be more sensitive than if a binary outcome is chosen. Consequently for given Type I and Type II errors the numbers of patients required will usually be smaller. In practice, there is little statistical benefit to be gained by having more than G = 5 ordered categories **[24]**. Thus, although there are G = 23 categories in the Hospital Anxiety and Depression Scale (HADS), with a low score as a desirable outcome **[25]**, the data might be reduced to five for planning purposes (Table 2). When considering a new trial that is aimed at reducing HAD scores, investigators could use these data to provide planning values for S.

Although an ordinal scale outcome is envisaged, at the initial planning stages, investigators may first think in binary terms and, for example, consider the (cumulative) proportion with HADS ≤ 7 of 60.39% with S might be improved by 10% to 70.39% using T. This $\delta_{\text{Plan}} = 0.1$ is then expressed as the planning odds ratio of $\text{OR}_{\text{Plan}} = \dfrac{0.7039/(1-0.7039)}{0.6039/(1-0.6039)} = 1.56$.

The basic assumption when considering the range of categories is that, wherever the investigators make the binary cut (in this example at ≤3, ≤7, ≤10 or ≤15), the same

planning OR applies [26]. Thus had the cut been made at HADS ≤ 10, then the cumulative proportion with S is 0.8182 would imply that, if the OR_{Plan} = 1.56, this proportion would increase to 0.8753 with T.

The process of calculating the sample size begins by using the observed proportions for S as the planning values π_{S0}, π_{S1}, π_{S2}, π_{S3}, then with the design OR_{Plan} calculate those anticipated for T as π_{T0}, π_{T1}, π_{T2}, and π_{T3}.

In general, if the categories are dichotomized by including the categories 1, 2, …, g, in one group, and the remainder g + 1, g + 2, … , G categories in the other, then a general expression for the odds ratio is

$$OR_g = \frac{C_{Tg}/(1-C_{Tg})}{C_{Sg}/(1-C_{Sg})}, \text{ g} > 0. \qquad (15)$$

where C_{Sg} and C_{Tg} are the cumulative proportions in the S and T groups respectively. If all the OR_g are assumed to be equal to OR_{Plan} then equation (15) can be rearranged to give

$$C_{Tg} = \frac{OR_{Plan}C_{Sg}}{OR_{Plan}C_{Sg}+(1-C_{Sg})}. \qquad (16)$$

Once all the C_{Tg} are obtained from equation (16), the corresponding π_{Tg} are calculated by subtraction as in Table 2, for example, $\pi_{T1} = C_{T1} - C_{T0} = 0.7039 - 0.3766 = 0.3274$.

Table 2 Deriving the anticipated proportions within each category for the Test intervention calculated from that anticipated for patients in the Standard with an assumed planning odds ratio, OR_{Plan} = 1.56.

	Outcome					
	HADS score					
	0-3	4-7	8-10	11-15	16-22	
g	0	1	2	3	4	
Standard	$s_0 = 43$	$s_1 = 50$	$s_2 = 33$	$s_3 = 24$	$s_3 = 4$	$n_S = 154$
$p_{Sg} = s_g/n_S$	$p_{S0} = 0.2792$	$p_{S1} = 0.3246$	$p_{S2} = 0.2143$	$p_{S3} = 0.1558$	$p_{S4} = 0.0260$	
Planning	$\pi_{S0} = 0.2792$	$\pi_{S1} = 0.3246$	$\pi_{S2} = 0.2143$	$\pi_{S3} = 0.1558$	$\pi_{S4} = 0.0260$	
C_{Sg}	$C_{S0} = \pi_{S0}$ = 0.2792	$C_{S1} = \pi_{S0}+\pi_{S1}$ = 0.6039	$C_{S2} = \pi_{S0}+\pi_{S1}+\pi_{S2}$ = 0.8182	$C_{S3} = \pi_{S0}+\pi_{S1}+\pi_{S2}+\pi_{S3}$ = 0.9740	C_{S4} = 1	
Test						
C_{Tg}	C_{T0} = 0.3766	C_{T1} = 0.7039	C_{T2} = 0.8753	C_{T3} = 0.9832	C_{T4} = 1	
π_{Tg}	$\pi_{T0} = 0.3766$	$\pi_{T1} = 0.3274$	$\pi_{T2} = 0.1713$	$\pi_{T3} = 0.1079$	$\pi_{T4} = 0.0168$	
$(\pi_{Sg}+\pi_{Tg})^3$	$(0.2792 + 0.3766)^3$ = 0.2820	$(0.3246 + 0.3274)^3$ = 0.2772	$(0.2143 + 0.1713)^3$ = 0.0573	$(0.1558 + 0.1079)^3$ = 0.0183	$(0.0260 + 0.0168)^3$ = 0.0001	Total 0.6349
					$\Gamma = 1 - (0.6349/8) = 0.9206$	

The expression for calculating sample sizes for comparing two interventions using clusters with individual subject responses from G ordered categories is [24]:

$$n_S = DE \times \left\{\frac{3}{\Gamma}\left(\frac{1+\varphi}{\varphi}\right)\frac{(z_{1-a/2}+z_{1-\beta})^2}{(\log OR_{Plan})^2}\right\}, n_T = \varphi n_S, \quad (17)$$

where

$$\Gamma = \left[1 - \frac{1}{8}\sum_{g=0}^{G-1}(\pi_{Sg} + \pi_{Tg})^3\right]. \quad (18)$$

In certain circumstances the calculation for Γ can be simplified. Thus if G > 5, Γ ≈ 1, while if all $\pi_{Sg} + \pi_{Tg}$ are approximately equal then $\Gamma \approx 1 - 1/G^2$.

Improvement in HADS score [25]

Assuming investigators plan a cluster trial on the basis of the information of Table 2 with anticipated effect size OR_{Plan} = 1.56, then log OR_{Plan} = 0.4447 and equation (18) gives Γ = 0.9206. Further assuming m = 30 with sd(m) = 0, ρ = 0.001, then DE = 1 + (29 × 0.001)] = 1.029. Finally the investigators set φ = 1, two-sided α = 0.05, β = 0. 2 and obtain from equation (17) $n_S = n_T = 1.029 \times \left\{\frac{3}{0.9206}\left(\frac{1+1}{1}\right)\frac{(1.96+0.8416)^2}{(0.4447)^2}\right\}$ = 1.029 × 258.68 = 266.17. To be divisible by m = 30 this is rounded to 270 to give N = 2 × 270 = 540 subjects and K = 540/30 = 18 clusters with 9 per intervention.

Incidence Rate outcome

Aggregate design

In some situations, all the m individuals within each cluster are followed-up for a fixed period (say F years) and the number of occurrences of a specific event is recorded among the individuals within that time. If r_i individuals in cluster i experience the event then the event rate per-person-years is estimated by $\lambda_i = r_i / (m \times F)$. In other situations, each of the m subjects within cluster i may have different follow-up times, say, f_{ij}, in which case

the incidence rate for cluster i is $\lambda_i = r_i/Y_i$, where $Y_i = \sum_j f_{ij}$ is the anticipated total follow-up time recorded for the cluster. In practice, the incidence rate may be expressed as per-person, per-100- or per-1000-person days, years or other time frames depending on the context.

An aggregate design is the usual option as it is the rate provided from each cluster which will be the unit for analysis. In this case, an alternative to the ICC as a measure of how close individuals responses are within a cluster is the coefficient of variation, cv(Rate) = SD(Cluster Rates within an Intervention)/Mean(Rate for that Intervention), for the aggregated outcome of concern **[10]**. This should not be confused with CV(m) of the individual cluster sizes defined previously and used in equation (7).

In designing a cluster trial, the investigators would need to specify planning values λ_S and λ_T for the mean incidence rates of the interventions, the corresponding cv_S and cv_T (often assumed equal), as well as the maximum duration of follow-up, F, of the individuals in the clusters. The number of clusters required is: **[10, 27]**

$$k_S = \left\{\left(\frac{\lambda_S+\varphi\lambda_T}{mF\varphi}\right) + (cv_S^2)\lambda_S^2 + (cv_S^2)\lambda_S^2\right\}\frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{(\lambda_S-\lambda_T)^2} + \left[\frac{z_{1-\alpha/2}^2}{2(1+\varphi)}\right],\ k_T = \varphi k_S. \quad (19)$$

The $\left[\frac{z_{1-\alpha/2}^2}{2(1+\varphi)}\right]$ of equation (19) is the adjustment for when the number of clusters is small.

If subjects are only followed up to when the event of interest occurs then mF in equation (19) may be replaced by Y, the anticipated cumulative follow-up time that will be recorded in every cluster.

Left ventricular systolic dysfunction [6]

The results from a trial concerned with attempts to improve outcome for patients with left ventricular systolic dysfunction suggested that Usual care (S) was associated with a 7.2

deaths per 100 years (λ_S = 0.072). It is hoped that Enhanced care (T) might reduce this by 20% (λ_T = 0.0576). A 1:1 cluster trial is planned with two-sided α = 0.05, β = 0.2, m = 12, F = 5 years, and $cv_S = cv_T = 0.1$. Use of equation (19) results in $k_S = k_T = 86$, so that a total of K = 172 primary care units are required. Had more variation been anticipated, perhaps $cv_S = cv_T = 0.2$, then K = 192.

Time-to-event outcome

Non-aggregate design

Rather than merely counting the number of events (as for the incidence rate) if the individual times to the event (often termed survival times) are recorded and used in the analysis the usual summary for each intervention is the Kaplan-Meier survival curve. The comparison between interventions is then made using Cox proportional hazards regression model **[23]** including a random effects term to account for the cluster design. This analysis provides an estimate of the corresponding hazard ratio (HR) which summarises the relative survival difference between the groups. A HR = 1 corresponds to the null hypothesis of no difference.

For planning purposes, it is usual to specify γ_S and γ_T which are the anticipated proportions of subjects alive at a fixed time-point beyond the date their cluster was randomised. Once the design team has specified these, then the planning HR can be calculated from

$$HR_{Plan} = \frac{\log\gamma_T}{\log\gamma_S} \qquad (20)$$

The number of subjects required is **[28]**

$$n_S = DE \times \left(\frac{1}{\varphi}\right)\left(\frac{1+\varphi HR}{1-HR}\right)^2 \frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{[(1-\gamma_S)+\varphi(1-\gamma_T)]} \text{ and } n_T = \varphi n_S, \qquad (21)$$

to give a total sample size $N = n_S + n_T$. The ICC is difficult to estimate for survival data, but one suggestion [11, p64] is to treat the data as binary to get the ICC. In general we would recommend using a range of ICCs as the authors in the following example did.

Heart dysfunction

In the Trial of Education And Compliance in Heart (TEACH) dysfunction trial **[7]** the investigators anticipated that the 1-year rate of re-hospitalisation following earlier hospital admission for heart problems (S) would be about 75%. It was further anticipated that this could be reduced to 60% using enhanced education on their condition from their home pharmacist (T). Thus, with $\gamma_S = 0.75$ and $\gamma_T = 0.60$ representing the anticipated proportions re-admitted at 1-year, equation (20) gives $HR = \frac{\log 0.60}{\log 0.75} = 1.7757$. Further, if we take, as the investigators did, m = 2, $\varphi = 1$, $\alpha = 0.05$ and $\beta = 0.2$ then, for a non-aggregate design, equation (21) becomes $n_S = [1 + (2-1)\rho] \times \left(\frac{1}{1}\right)\left(\frac{1+1.7757}{1-1.7757}\right)^2 \frac{(1.96+0.8416)^2}{[(1-0.60)+(1-0.75)]} = 154.63 \times (1+\rho)$.

Setting ρ equal to 0.05 and 0.10, as the investigators did, gives the respective values of $N = n_S + n_T = 2 \times 163 = 326$ and 342 with the corresponding total number of pharmacies (clusters) required as K = 326/2 = 163 and 171 respectively. To allow a 1:1 randomisation, these are then increased to 164 and 172 to give either 82 or 86 pharmacies per intervention.

Aggregate design

For an aggregate design, the endpoint will be the survival rate at a fixed time following randomisation, say at the 1-year follow-up. The planning values for these rates, say γ_S

and γ_T, are then taken as π_S and π_T and used in the same way as for sample size calculations of a binary endpoint cluster design.

Matched designs

In the preceding sections the individual clusters participating in the trial have been identified and then randomised (say) in equal numbers to receive the S or T intervention. However, if the clusters themselves are of variable size, then an alternative method of allocation is first to rank these clusters in terms of their size, and then create cluster pairs of a similar size. Once these 'matched' pairs are identified, the allocation of T is made at random to one of the pair and the other is then automatically assigned to S. Options, other than size, may be used to create the matched pairs. The choice being perhaps related to features of the clusters concerned; such as their location in Rural or Urban areas.

Once the trial is complete, the difference in summary measure from each matched pair of clusters will be calculated. Thus a matched design implies an aggregate design. Thus, for example, if the endpoint is continuous this measure will be the difference, $d_i = \bar{y}_{iT} - \bar{y}_{iS}$. for each of the cluster pairs, K_{Pairs}. From these values the mean difference $\bar{d}$ is obtained and this estimates the true difference between the interventions, δ. The paired t-test then tests the null hypothesis $\delta = 0$.

However the values of $\bar{y}_{iT}$ and $\bar{y}_{iS}$ from the matched clusters may be themselves associated. If, in a completed trial involving K_{Pairs}, the individual values of $\bar{y}_{iT}$ and $\bar{y}_{iS}$ are available then their correlation, η, may be calculated and used for future planning purposes. It is recommended **[10]** that, η, replaces the ICC, ρ, in the DE of equation (6).

Nevertheless it should be recognised that, unlike for the ICC, we know of no published values of η. In this situation, the number of cluster pairs required is:

$$K_{Pairs} = DE \times \left\{ \frac{2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{\left(\delta_{Plan}/\sigma_{Plan}\right)^2} + \left[\frac{z_{1-\alpha/2}^2}{2}\right] \right\}. \qquad (22)$$

Here σ_{Plan} is the anticipated standard deviation of the differences, d_i, obtained from the cluster pairs. As the number of cluster pairings is likely to be relatively small the correction term $\left[\frac{z_{1-\alpha/2}^2}{2}\right]$ is added **[10]**.

With little or no prior knowledge of either η, σ_{Plan} or both the design team would need to consider a range of options before deciding on the number of cluster pairs to include.

Daily exercise and Quality of Life [22]

If we suppose this planned trial was to involve communities with very diverse socio-economic characteristics, then the design time might wish to create cluster pairs with similar features. The anticipated improvement in PF with T over S is assumed the same with δ_{Plan} = 10 units. However the previous trial provided a planning value of 6.12 units whereas for a matched design σ_{Plan} might be anticipated to be smaller than this to an extent depending on the numbers to be recruited per cluster, m. Thus a range of values for σ_{Plan}, as well as η, are investigated.

As a first step, the investigators take η = 0.01 and $\sigma_{Plan} = 6.12$ and, from equation (22), with m = 30, two-sided α = 0.05 and β = 0.1 obtain, $K_{Pairs} = [1 + (30 - 1) \times 0.01] \times \left\{\frac{2(1.96+1.2816)^2}{(10/6.12)^2} + \left[\frac{1.96^2}{2}\right]\right\}$ = 7.55 or 16 clusters which are then matched in pairs. If η = 0.05 then the number of cluster pairs increases to K_{Pairs} = 15. A reduced $\sigma_{Plan} =$ 3.06 results in K_{Pairs} = 4 and 8 for η = 0.01 and 0.05respectively.

1

Conclusion

Cluster trials consume considerable logistical and other resources, so a critical factor is to determine the appropriate size for the trial in question. As subjects are not individually randomised to the interventions but are allocated in clusters then this feature needs to be accounted for in both the planning and the statistical analysis. In some instances, the number of clusters available may be fixed, in others the number of subjects per cluster is fixed or possibly both may be open to choice. Although 1:1 allocation of interventions is usual, there is nevertheless a decision to be made with respect to this ratio.

As in individually randomized trials, the two-sided test-size (α) is conventionally set at 0.05, whereas the power ($1 - \beta$) is often set at 0.8 although a higher value (say 0.9) is more desirable. Further each design team will have to decide on the anticipated effect size (the difference between S and T) which will be very context specific but should reflect a realistic and clinically important difference between the groups. However, in the cluster trial situation an ICC (or some other measure of the lack of independence of the subjects within a cluster) will need to be specified. In some situations, cluster trials may have been done in similar circumstances to that in planning, so that the magnitude of such measures may be well documented. However in most situations some (often considerable) judgment is required. In either case the design team will need to consider the impact on sample size of a range of options for this (and other design features) before deciding the final trial size. The investigators too will need to verify what will be required by CONSORT **[29]** for reporting their trial to ensure all these requirements are in place before the trial commences. Of particular relevance here is the need for a clear but succinct justification of trial size. Thus it is important to retain details of the way in

which, at the planning stage, the eventual trial design and size were determined. Further discussion of recent issues in the design and analysis of cluser trials is given in [30] [Conclusion word count: 340]

References

[1] Etxeberria A, Pérez I, Alcorta I, Emparanza JI, Ruiz de Velasco E, Iglesias MT, Orozco-Beltrán D, Rotaeche R. The CLUES study: a cluster randomized clinical trial for the evaluation of cardiovascular guideline implementation in primary care. BMC Health Serv Res 2013;13:438.

[2] Fretheim A, Oxman AD, Håvelsrud K, Treweek S, Kristoffersen DT, Bjørndal A. Rational prescribing in primary care (RaPP): a cluster randomized trial of a tailored intervention. PLoS Med. 2006;3(6):e134.

[3] Kaczorowski J, Chambers LW, Dolovich L, Paterson JM, Karwalajtys T, Gierman T, Farrell B, McDonough B, Thabane L, Tu K, Zagorski B, Goeree R, Levitt CA, Hogg W, Laryea S, Carter MA, Cross D, Sabaldt RJ. Improving cardiovascular health at population level: 39 community cluster randomised trial of Cardiovascular Health Awareness Program (CHAP). BMJ 2011;342:d442.

[4] Martin RM, Patel R, Kramer MS, Vilchuck K, Bogdanovich N, Sergeichick N, Gusina N, Foo Y, Palmer T, Thompson J, Gillman MW, Smith GD, Oken E. Effects of promoting longer-term and exclusive breastfeeding on cardiometabolic risk factors at age 11.5 years: A cluster-randomized, controlled trial. Circulation. 2014;129:321-9.

[5] Pladevall M, Brotons C, Gabriel R, Arnau A, Suarez C, de la Figuera M, Marquez E, Coca A, Sobrino J, Divine G, Heisler M, Williams LK; Writing Committee on behalf of the COM99 Study Group. Multicenter cluster-randomized trial of a multifactorial

intervention to improve antihypertensive medication adherence and blood pressure control among patients at high cardiovascular risk (The COM99 Study). Circulation 2010;122:1183-91.

[6] Lowrie R, Mair FS, Greenlaw N, Forsyth P, Jhund PS, McConnachie A, Rae B, and McMurray JJV on behalf of the Heart Failure Optimal Outcomes from Pharmacy Study (HOOPS) Investigators. Pharmacist intervention in primary care to improve outcomes in patients with left ventruicular systolic dysfunction. European Heart Journal, 2012;33:314-24.

[7] Gwadry-Sridhar F, Guyatt G, O'Brien B, Arnold JM, Walter S, Vingilis E and MacKeigan L. TEACH: Trial of Education And Compliance in Heart dysfunction chronic disease and heart failure (HF) as an increasing problem. Contemporary Clinical Trials 2008; 29: 905-18.

[8] Donner A and Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.

[9] Eldridge S and Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. Chichester: Wiley-Blackwell; 2012.

[10] Hayes RJ and Moulton L. Cluster Randomised Trials. Boca Raton: Chapman and Hall/CRC; 2009.

[11] Campbell MJ and Walters SJ. How to design, analyse and report cluster randomized trials in medicine and health related research. Chichester: Wiley-Blackwell; 2014.

[12] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, Grimshaw JM, Weijer C, Zwarenstein M and Donner A. Impact of CONSORT extension for cluster randomized trials on quality of reporting and

study methodology: review of random sample of 300 trials, 2000-8. BMJ 2011; 343: d5886.

[13] Campbell MK, Elbourne DR and Altman DG. CONSORT statement: extension to cluster randomized trials. BMJ 2004; 328: 702-8.

[14] Campbell MJ. Extending CONSORT to include cluster trials. BMJ 2004, 328, 654-5.

[15] Machin D, Campbell MJ, Tan SB and Tan SH. Sample Size Tables for Clinical Studies. 3 ed. Chichester: Wiley-Blackwell; 2009.

[16] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol, 2004; 57: 785-94.

[17] Thompson DM, Fernald DH and Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood prescription health projects. Ann Fam Med 2012; 10: 235-40.

[18] Batistatou E, Roberts C and Roberts S. Sample size and power calculations for trials and quasi-experimental studies with clustering. The Stata Journal 2014; 14: 159-175.

[19] Van Breukelen GJP and Candel MJJM. Comments on "Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests". Stat Med 2012; 31: 397-400.

[20] Dresser GK, Nelson SA, Mahon JL, Zou G, Vandervoort MK, Wong CJ, Feagan BG and Feldman RD. Simplified therapeutic intervention to control hypertension and hypercholesterolemia: a cluster randomized controlled trial (STITCH2). J Hypertens; 2013; 31: 1702-13.

[21] Munro JF, Nicholl JP, Brazier JE, Davey R and Cochrane T. Cost effectiveness of a community based exercise programme in over 65 year olds: cluster randomised trial. J Epid Comm Hlth 2004; 58: 1004-10.

[22] Walters SJ, Munro JF and Brazier JE. Using the SF-36 with older adults: a cross-sectional community-based survey. Age and Aging 2001; 30, 337-343

[23] Tai BC and Machin D. Regression Methods for Medical Research. Chichester: Wiley-Blackwell; 2014.

[24] Whitehead J. Sample size calculations for ordered categorical data. Stat Med 1993; 12: 2257-72.

[25] Fayers PM and Machin D. Quality of life. 3 ed. Chichester: Wiley-Blackwell; 2015.

[26] Campbell MJ, Julious SA and Altman DG. Sample size for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ 1995; 311, 1145-8.

[27] Hayes RJ and Bennett S. Simple sample size calculations for cluster-randomized trials. Int J Epidemiol 1999; 28: 319-26.

[28] Xie T and Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. Stat Med 2003; 22: 2835-46.

[29] Campbell MK, Piaggio G, Elbourne DR and Altman DG. Consort 2010 statement: extension to cluster randomized trials. BMJ 2012; 345: e5661. doi:10.1136/bmj.e5661.

[30] Campbell MJ (2014) Challenges of cluster randomised trials Journal of Comparative Effectiveness Research 3(3), 271-281.

Acknowledgements

1 This research was funded by the affiliated Institutions of the authors. No specific grants
2 (only core funding) supported this work.