Cluster randomised trials with different numbers of measurements at baseline and endline: sample size and optimal allocation

Running head: Baseline data: cluster randomised trials

Word count: main text 3833; main text plus appendices 4780

Authors: Andrew J Copas[1], Richard Hooper[2]

Affiliations

1. MRC Clinical Trials Unit at University College London
2. Centre for Primary Care and Public Health, Queen Mary University of London

Corresponding author:

Andrew Copas

Institute for Clinical Trials Methodology

University College London

90 High Holborn

London

WC1V 6LJ

United Kingdom

Email: a.copas@ucl.ac.uk

Tel: +44 20 7670 4888

Abstract

Background/Aims

Published methods for sample size calculation for cluster randomised trials with baseline data are inflexible and primarily assume an equal amount of data collected at baseline and endline i.e. before and after the intervention has been implemented in some clusters. We extend these methods to any amount of baseline and endline data. We explain how to explore sample size for a trial if some baseline data from the trial clusters have already been collected as part of a separate study. Where such data aren't available we show how to choose the proportion of data collection devoted to the baseline within the trial, when a particular cluster size or range of cluster sizes is proposed.

Methods

We provide a design effect given the cluster size and correlation parameters, assuming different participants are assessed at baseline and endline in the same clusters. We show how to produce plots to identify the impact of varying the amount of baseline data accounting for the inevitable uncertainty in the cluster autocorrelation. We illustrate the methodology using an example trial.

Results

Baseline data provide more power, or allow a greater reduction in trial size, with greater values of the cluster size, intra-cluster correlation, and cluster auto-correlation.

Conclusion

Investigators should think carefully before collecting baseline data in a cluster randomised trial if this is at the expense of endline data. In some scenarios this will increase the sample size required to achieve given power and precision.

Key words: cluster randomised trial; baseline data; sample size; power; efficiency; study design

1. Introduction

Many cluster randomised trials compare an intervention to current practice. If outcome data have already been collected as part of a separate study or routine data collection preceding the trial (*i.e.* on a different sample of individuals but from the same clusters), these 'retrospective baseline data' can be included in the analysis along with the trial data. This will increase power or reduce the required sample size for the trial. Methods are available to calculate these impacts, but only for an equal amount of prior baseline data as data collected in the trial.[1]

However, when retrospective data are not available researchers may choose to collect baseline data prospectively as part of the trial. A proportion of the total data collection can be allocated to baseline data, rather than 'endline' data collected after some clusters have implemented the intervention. Whilst earlier researchers focussed on equal data collection at baseline and endline,[1,2] Green *et al*. recently investigated more flexible choices.[3] They established the circumstances in which collecting some baseline data for a given total cluster size will increase power or equivalently reduce the number of clusters required, and also showed how to calculate the optimal proportion of baseline data to maximise power. However they did not directly provide methodology for sample size calculation. We note that prospective collection of baseline data together with endline data from different individuals (two independent cross-sectional samples) is a design commonly chosen when the trial clusters are large communities subject to in- and out- migration which makes a panel design problematic.

Besides influencing power and sample size, another possible benefit of baseline data is to prevent baseline imbalance between arms through their use in stratified or restricted randomisation.[4-6] Baseline data may also improve the perceived 'face validity' of trials with a small number of clusters, where baseline imbalance can be a particular concern.[7]

In this article we provide design effects to allow sample size to be calculated with any number of baseline measurements, whether part of the trial (prospective) or already collected (retrospective). We then explain when baseline data will be more beneficial and show for prospective baseline data collection how to use plots based on the design effect to choose how much baseline data to collect (if any) given uncertainty in the correlation parameters and the range of total cluster sizes considered.

2. Methods

2.1 Design framework

We consider trials where different participants are assessed at baseline and endline in the same clusters (i.e. cross-sectional), and with a continuous outcome.

The data are assumed to follow this model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} X_i + \beta_2 t_{ij} + (1 - t_{ij}) u_{0i} + t_{ij} u_{1i} + \varepsilon_{ij}$$

for participant *j* in cluster *i*, where X_i denotes the trial arm for cluster *i* (coded 0 or 1), and t_{ij} denotes the period in which participant *j* in cluster *i* is assessed (0 for baseline, 1 for endline). We assume the random terms are Normally distributed. The observation error or variability term is denoted ε_{ij} and is assumed independent of terms u_{0i} and u_{1i} which are the cluster random effects for baseline and endline. We denote the variance and correlations thus

$$Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

$$Var(u_{0i}) = Var(u_{1i}) = \sigma_u^2; Corr(u_0, u_1) = \pi$$

and from these we can define the intracluster correlation (ICC) $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$. The term π is the cluster autocorrelation, the correlation between the underlying cluster population means at baseline and endline.

The two trial arms are assumed to be assessed under identical conditions at baseline, but may differ at endline, so that the model does not include a main effect for X_i. We assume all clusters provide the same number of measurements. We assume the baseline and endline data are each collected within a short period, so that measurements within each period within a cluster can be consider equally correlated (i.e. exchangeable).

2.2 Design effects

Previous authors mainly considered an equal number of baseline measurements (n_b) as endline measurements from each cluster (n_e), with only a few exceptions.[3,8] To allow more general designs we first express the correlation between the cluster sample means from the two periods as (see Appendix A for derivation)

$$r = \frac{\pi\rho\sqrt{n_b n_e}}{\sqrt{1 + (n_b - 1)\rho}\sqrt{1 + (n_e - 1)\rho}}$$

and assume ρ and π are both positive as would generally be the case. Other authors gave an equivalent formula for r when $n_b = n_e$.[1,9]

Suppose N_{ind} is the number of measurements required in an individual RCT without baseline data and N is the number of prospective measurements under an alternative design then we define the design effect (DE) by

$$N = DE \times N_{ind}.$$

The number of clusters required is N divided by the number of prospective measurements (i.e. taken as part of the trial) per cluster, rounded up to the next even integer assuming equal cluster allocation to control and intervention.

We next present two design effects for trials with baseline data which differ in whether the baseline data are part of the prospective trial data collection or alternatively are already collected prior to the trial (for example through a separate preparatory study) and so are not included in the sample size target for the trial.

2.2.1 *Baseline data collected within the trial*

For baseline data collected prospectively within the trial the design effect can be expressed as (see Appendix A for derivation)

$$DE_{within} = [1 + (n_e - 1)\rho][1 - r^2][(n_b + n_e)/n_e].$$

Hemming and Taljaard provide an equivalent design effect for when $n_b = n_e$ and also $\pi = 1$.[2]

In some scenarios it may be helpful to consider the number of baseline measurements to be collected per cluster given a particular total cluster size $m = n_b + n_e$, and this can be repeated where a range of sizes is considered. Consequently in Appendix B1 we re-express r and DE_{within} to show the impact of varying the proportion of baseline measurements $\theta = n_b/m$.

Only in some scenarios will collecting baseline measurements as part of the trial given a total cluster size increase power, relative to the design with no baseline measurements ($\theta = 0$). Green *at al.* show that power is increased by baseline data if and only if $\rho > 1/(1 + m\pi)$,[3] and that if this holds then the optimum proportion of baseline measurements to maximise power is given by

$$\theta_{opt} = (m\rho\pi + \rho - 1)/[\rho m(1 + \pi)].$$

They also show this optimum proportion is always less than one half. Because the maximum possible value of π is 1 and investigating the range of values it may take could be challenging, an initial condition to check for whether collecting baseline data could possibly increase power is $\rho > 1/(1 + m)$.

2.2.2 Baseline data already collected before the trial

Where baseline data have already been collected as part of another study, the design effect when considering the sample size for the trial can be expressed as (see Appendix A for derivation)

$$DE_{add} = [1 + (n_e - 1)\rho][1 - r^2].$$

Others give an equivalent design effect for when $n_b = n_e$.[1,10]

2.3 Selecting ranges for the correlation terms

The design effects require specification of the ICC (ρ) and cluster autocorrelation (π). However, unlike the ICC in cluster randomised trials, estimates of autocorrelation are not routinely reported for trials with baseline data or other designs.[10] Estimates of autocorrelation can be derived from a mixed regression analysis of data from cluster randomised trials with baseline data as shown in Appendix C. Estimates may also be obtained from routine datasets with clustered data collected in multiple periods. Clustered data collected in continuous time could be divided into discrete time periods for this purpose, but note that the implicit assumption that observations from the same cluster within each period are exchangeable may be a slightly artificial one in this case. Estimates derived in this way from clustered observational data have been reported for outcomes related to diabetes and ranged from 0.49 to 0.89.[11] In this article we consider values of π between 0.5 and 0.9. When applying these methods to a particular trial it may be possible to narrow the range, depending on trial characteristics and other information available.

Estimates of π are likely to come from previous studies that do not correspond exactly to the planned trial so it is necessary to also think qualitatively about its likely value. Higher values of π are more likely when, between the baseline and endline, there are only small changes in (i) the distribution of participant characteristics, (ii) the nature of exposure/care, (iii) staff providing care if appropriate, and (iv) the data collection method. Since less change is likely, higher values are more likely when the time interval between baseline and endline is short.

3. Results

3.1 Baseline data collected within the trial

Figure 1 presents plots based on the design effect to show the relative change in the numbers of clusters required as baseline data vary from none to half the total data collected (since the optimal proportion is always less than one half).

We see the benefits of baseline data are greater with greater values of the total cluster size, ICC and π. When the total cluster size is 50 and ICC is 0.01 then any baseline data collection reduces power, if half the data collection is devoted to baseline then the number of clusters must increase by around 60% to compensate. Conversely when the total cluster size is 200 and ICC is 0.05 then reductions in the number of clusters of between 15% (π=0.5) and 52% (π=0.9) are possible. Furthermore the shape of the curves indicates that the optimal baseline data proportion would be between around 25% (π=0.5) and 40% (π=0.9). A good design given uncertainty in π might allocate a third of data collection to baseline. In the other scenarios (total cluster size 50 and ICC 0.05, or size 200 and ICC 0.01) appreciable reductions in the number of clusters are impossible. Trialists may nevertheless consider baseline data to provide other benefits such as its use for restricted randomisation. We see that in both scenarios if 25% data collection were devoted to baseline then even if π=0.5 the loss of power can be compensated by an increase in the number of clusters of only around 5%.

3.2 Baseline data already collected before the trial

Figure 2 presents plots to show the relative change in the numbers of clusters required for a trial due to different amounts of additional baseline data, already collected separately from the trial, varying from none to double the amount of endline data. These plots are based on re-expressing *r* to show how it, and hence DE_{add}, vary with the ratio of baseline to endline measurements as we describe in Appendix B2. The three lines represent π=0.5, 0.7 and 0.9*.* The four graphs represent each combination of what in our work we consider a small endline cluster size (n_e) of 50 and large of 200, and low ICC ρ=0.01 and high ρ=0.05.

As with baseline data collected as part of the trial, we see additional baseline data are more beneficial with greater values of the endline cluster size, ICC, and π. When the size is 50 and ρ=0.01 then even double the amount of endline data permits only a negligible reduction in number of clusters, irrespective of π. Conversely when the endline cluster size is 200 and ρ=0.05 then reductions in the number of clusters of between 20% (π=0.5) and 70% (π=0.9) are possible. Furthermore most of the potential reduction in the number of clusters available can be achieved from additional baseline measurements amounting to half the number of endline measurements.

We provide Stata code to generate these types of plots in Appendix D. The reader can also access an R Shiny App and a Stata program to generate these plots for their own trial by visiting https://github.com/UCL/samplesize-CRTs-baseline.

4. Example

We describe a completed trial and consider variations on its design to illustrate our methodology. The trial assessed the effectiveness of a novel theory-based community mobilization intervention to change harmful gender norms.[12,13] The trial was conducted in South Africa and 22 villages (trial clusters) were randomised, 11 to the intervention and 11 to control. The primary outcome was the score from the Gender Equitable Mens Scale (GEMS), and when designing the trial this was assumed to have an ICC of 0.05. The trial collected outcome data through cross-sectional samples taken at endline and also at baseline as part of the trial (no baseline data were already available for the trial clusters before the trial). The expected sample size was 55 per cluster at each time point, but it was planned for the primary analysis to be conducted separately by participant gender with roughly equal numbers of each gender. Hence the target sample size was roughly 55 men and 55 women per cluster divided equally between baseline and endline, giving an average cluster size of 27.5 at each time point, though for each cluster of course these values are whole numbers.

We illustrate our methodology through focussing on sample size and power for the analysis of data from women. The published power calculation states that the design provides 80% power to detect a mean difference of 3 points in GEMS between arms, conservatively based on endline data alone (baseline data were ignored in the sample size calculation but included in the final analysis) and assuming a standard deviation of 6 at each time point. It seems the calculation may be further conservative because the design effect used was based on the full cluster size rather than for each gender, and was therefore too large, since the primary analysis and sample size calculations are conducted for each gender separately. We calculate that the design effect considering women alone should be approximately 2.33 and the effective sample size is therefore roughly 130 per arm so that (ignoring baseline data) the design provides 80% power to detect a mean difference of 2.1 points in GEMS and we continue our illustration treating this as the target effect size.

We considered a range for the cluster auto-correlation, π, of 0.5 to 0.8. We considered values as high as 0.9 unlikely because the two surveys were over 2 years apart and because there would have been some residential turnover in that period so that small changes in the characteristics of the clusters were possible.

We consider the total cluster size that was used, 55, and in Figure 3 we present the impact on the number of clusters required from varying the proportion of baseline data from zero to one half, which was the value used in the design. Whilst the lines plotted closely reflect Figure 1 for size 50 and ICC 0.05, as expected, here we plot the impact on the number of clusters required per arm. The number required with no baseline data, where the standard design effect takes value 3.70, is 8.8, which for graphical illustration we do not round up to 9 as would be needed in practice. The initial check of whether baseline data could potentially increase power (see section 2.1.1) is satisfied as 0.05 > (1/55). Figure 3 shows that a choice of just under one fifth baseline data (10 participants measured at baseline and 45 endline) would be good because this is optimal for π=0.65 and provides a near optimal reduction in the design effect if π=0.5 or 0.8. The optimal proportions calculated following Green *et al.*,[3] and which match Figure 3 closely, are 0.103, 0.185 and 0.253 for π=0.5, 0.65 and 0.8 respectively.

Table 1 shows sample size calculations comparing the designs with total cluster size 55 and either no baseline measurements, 10 participants measured at baseline and 45 at endline, or equal numbers at baseline and endline as was implemented. The calculations reflect Figure 3 and quantify the modest benefit from including 10 baseline measurements relative to none and the modest reduction in power from allocating half the measurements to baseline. Given the modest number of clusters concerned, in this example it would be more natural to select 10 baseline measurements per cluster in order to increase power rather than to reduce the number of clusters in the trial, as shown in the final column of the table.

Trialists may of course consider a range of total cluster sizes, in this example increasing the size beyond 55 would lead to higher optimal proportions of baseline data and decreasing below 55 would lead to lower optimal proportions conveying lower benefit. At average cluster size 27.5 (basis of original sample size calculation) the optimal design has no baseline data unless π is close to 1 (optimal proportion zero at π =0.65), and even then any benefit conveyed from baseline data is very small. This design with no baseline data, as shown in Table 1, can achieve 80% power with many fewer participants (303) than any of the designs with total cluster size 55. To match the 90% power that could be expected with 11 clusters per arm, 10 baseline and 45 endline measurements per cluster, at average cluster size 27.5 and no baseline measurements 15 clusters per arm are required, which equates to 413 participants.

The final decision on the design to collect baseline data, and specifically an equal number of measurements at baseline and endline, may have been influenced by considering other benefits of baseline data such as face validity given the relatively modest number of clusters.

5. Discussion

We have provided new methodology to calculate power or sample size for cluster randomised trials with any amount of baseline data and endline data per cluster, where baseline and endline data are collected cross-sectionally from different groups of participants. We distinguish between scenarios where the baseline data have already been collected and others where it will be collected as part of the trial. When baseline data collected prior to the trial are not available, and trialists have a particular value or range of values for the total cluster size in mind, then the plots we recommend extend the work of Green *et al.*[3] to guide the choice of the proportion of baseline data given uncertainty in the correlation parameters. Since the optimal proportion will often be zero (no baseline data), we provide a simple inequality in the cluster size and ICC so that researchers can see whether it is worth exploring baseline data collection. In more cases it will be possible to include some prospective baseline data without compromising power and yet conveying other benefits such as allowing its use in restricted randomisation.

We found, as did Green *et al*.,[3] that baseline data provide more power when the cluster autocorrelation is greater which is intuitive because this means baseline data are more predictive of endline outcomes so provide more information for the comparison between arms. We also found that baseline data provide more power when the cluster size and the ICC are larger. This too is intuitive because as these increase the information from each additional participant decreases and the standard cluster randomised trial becomes less efficient than other designs among which introducing baseline data is perhaps the simplest example.[10] In common with these other designs, a further benefit of baseline data is some preservation of trial power if the ICC is higher than anticipated.

In our presentation of sample size calculations where retrospective baseline data are available we have assumed these data are freely available to the researchers. Our methodology could also be used however where retrospective data are available but expensive to obtain, because for example they can only be collected through searching of paper records. In this scenario different amounts of baseline data per cluster could be considered, alongside different amounts of endline data, with the choice taking account of the relative cost of their collection.

Because the cluster autocorrelation plays such an important role in planning trials with baseline data (as shown in Figure 3 for example) we strongly recommend that estimates be routinely reported, at least for cluster crossover trials, stepped wedge trials, and cluster randomised trials with baseline data. We suggest that in the latter design, estimates should also be reported separately by study arm. The methodology we have developed assumes the same cluster autocorrelation (and ICC) in the study arms and does not apply in the event of substantial difference.

We have assumed that outcomes measured within the same time period (baseline or endline) in the same cluster are exchangeable, while outcomes from different periods in the same cluster have an attenuated correlation that depends on the cluster autocorrelation. This is likely if baseline and endline data are collected in surveys administered at two time points. If alternatively the two periods involve continuous recruitment of individuals over extended intervals of time, and particularly if these periods are contiguous, then these assumptions are doubtful. In fact in this case we would expect two participants recruited immediately before and after the cross-over between periods to have more closely correlated outcomes than two participants recruited at either end of the same period, and alternative statistical models may be more appropriate.

We have presented methodology to calculate an optimal proportion of baseline data where this is collected prospectively as part of the trial. In conjunction with a range of total cluster sizes this allows trialists to identify an optimal trial design achieving good power whilst minimising the number of clusters and/or participants. We acknowledge however that trialists will need to consider research costs and in particular that typically including a baseline survey adds additional fixed costs. In other words for a given total cluster size a trial with both baseline and endline survey will typically cost more than a trial with endline survey only. However once it is decided to have a baseline survey then costs will typically be similar for each baseline and endline measurement. Costs will also relate to both the number of clusters and number of participants surveyed, costs and feasibility across different total cluster sizes can be compared to select the best size.[14] When the number of clusters available is modest our methodology is better used to increase power or reduce cluster size than investigate a reduction in the number of trial clusters, because of concerns over face validity and also to permit a full range of analysis methods (some authors have suggested that some methods should not be applied when there are fewer than 20 clusters per arm).[15-17]

For prospective collection of baseline data we have focussed on selecting a proportion between zero and one half because Green *et al*. showed the optimal proportion to maximise precision or power is always less than one half.[3] However our design effect can be used for any proportion of baseline data collection between zero and one. A proportion of baseline data above one half could be considered in particular if the baseline data will be used for other purposes, for example to understand population needs and hence refine aspects of the intervention.

If baseline data are collected prospectively we recommend, to ensure high quality data, randomisation should be delayed until after baseline data collection. If this is not possible then alternatively participants and data collectors should remain blinded to allocation.

In further work our methodology could be examined and possibly adapted for binary outcomes, a small number of clusters,[18] unequal cluster sizes,[19] trials where outcomes are collected gradually over time leading to a more complex correlation structure, and 'open cohort' trials in which the some individuals are measured both at baseline and endline, and other individuals in one period only.

Previous methods for sample size calculation have focussed on an equal number of baseline and endline measurements. Our work has provided the means and Stata code by which to relax this artificial restriction and design trials that efficiently either increase power or provide the other benefits of baseline data. This can lead to trials that are cheaper, more robust, and expose fewer participants to unproven interventions.

Funding

AC was supported by the UK Medical Research Council (MC_UU_12023/29). The funder had no involvement in the development of this article.

ORCID iD

Andrew Copas https://orcid.org/0000-0001-8968-5963

References

1. Teerenstra S, Eldridge S, Graff M, et al. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;31: 2169–2178.

2. Hemming K, and Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol* 2016;69:137-146.

3. Green DP, Lin W, and Gerver C. Optimal allocation of interviews to baseline and endline surveys in place-based randomized trials and quasi-experiments *Evaluation Review* online first.

4. Raab GM, and Butcher I. Balance in cluster randomised trials. *Stat Med* 2001; 20:351-365.

5. Ivers NM, Halperin IJ, Barnsley J, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 2012; 13:120.

6. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials* 2004; 1: 297-305.

7. Wright N, Ivers N, Eldridge S, et al. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015; 68: 603–609.

8. Duffy SW, South MC, and Day NE. Cluster randomization in large public health trials: the importance of antecedent data. *Stat Med* 1992; 11:307-316

9. Hooper R, Teerenstra S, de Hoop E, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med* 2016; 35:4718–4728.

10. Hooper R, and Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*. 2015;350:h2925.

11. Martin J, Girling A, Nirantharakumar K, et al. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. *Trials* 2016; 17:402

12. Pettifor A, Lippman SA, Selin AM et al. A cluster randomized-controlled trial of a community mobilization intervention to change gender norms and reduce HIV risk in rural South Africa: study design and intervention. *BMC Public Health* 2015 15:752.

13. Pettifor A, Lippman SA, Gottert A et al. Community mobilization to modify harmful gender norms and reduce HIV risk: results from a community cluster randomized trial in South Africa. *J Int AIDS Soc* 2018, 21:e25134

14. Connelly LB. Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Control Clin Trials* 2003; 24:544-59

15. Huang S, Fiero MH, and Bell ML Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study *Clin Trials* 2016;13(4):445-9

16. Murray DM, Varnell SP, and Blitstein JL Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments *Am J Public Health* 2004;94(3):423-432

17. Kahan BC, Forbes G, Ali Y et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study *Trials* 2016;17(1):438

18. Hayes R, and Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol* 1999;28**:**319–26.

19. Eldridge SM, Ashby D, and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006; 35**:**1292–300.

20. Borm GF, Franseb J, and Lemmens WAJG. A sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol* 2007;60:1234-1238.

Design	Cluster auto-correlation	Design effect	Number of clusters per arm for 80% power	Number of participants per arm for 80% power	Power from 11 clusters per arm[1], %
Individual RCT	N/A	N/A	N/A	130	N/A
Parallel group cluster RCT, average m=27.5	N/A	2.33	11	303	80
Parallel group cluster RCT, m=55	N/A	3.70	9	495	88
Cluster RCT, m=55: 10 baseline and 45 endline measurements	0.50	3.67	9	495	89
	0.65	3.51	9	495	90
	0.80	3.30	8	440	92
Cluster RCT, m=55: half baseline and half endline measurements	0.50	4.24	11	605	84
	0.65	3.96	10	550	86
	0.80	3.61	9	495	89

1. Power calculation included because the example trial had 11 clusters per arm

Table 1. Sample size for different possible designs to detect an intervention effect of 2.1 units given a standard deviation of 6 units in both arms under two-sided testing at 5% level, and power given 11 clusters per arm. The anticipated ICC is ρ=0.05.

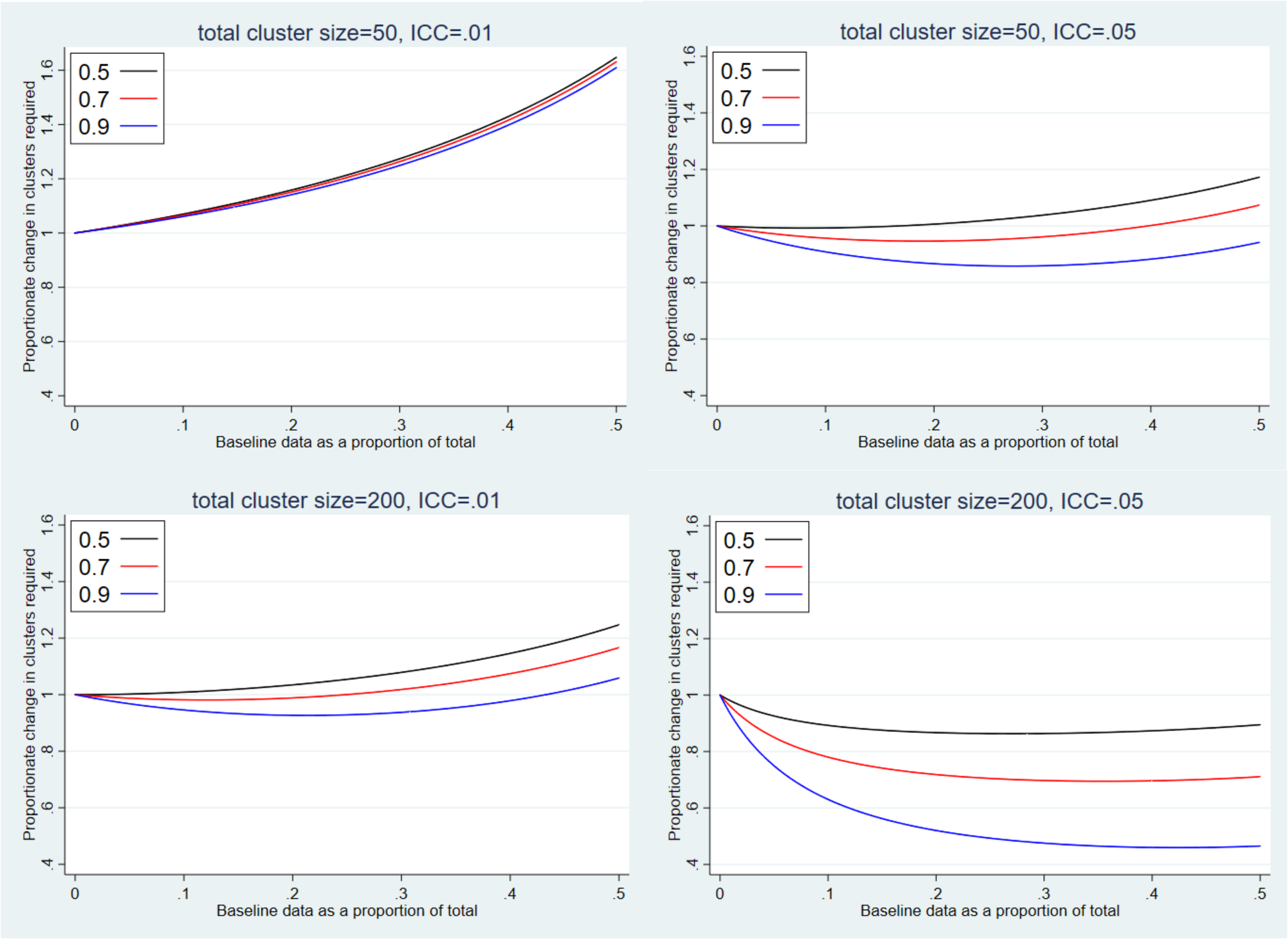

Figure 1. Proportionate change in number of clusters required according to the proportion of baseline data collection (θ), for two different total cluster sizes (m) and ICC values (ρ), and in each case for π=0.5, 0.7 and 0.9

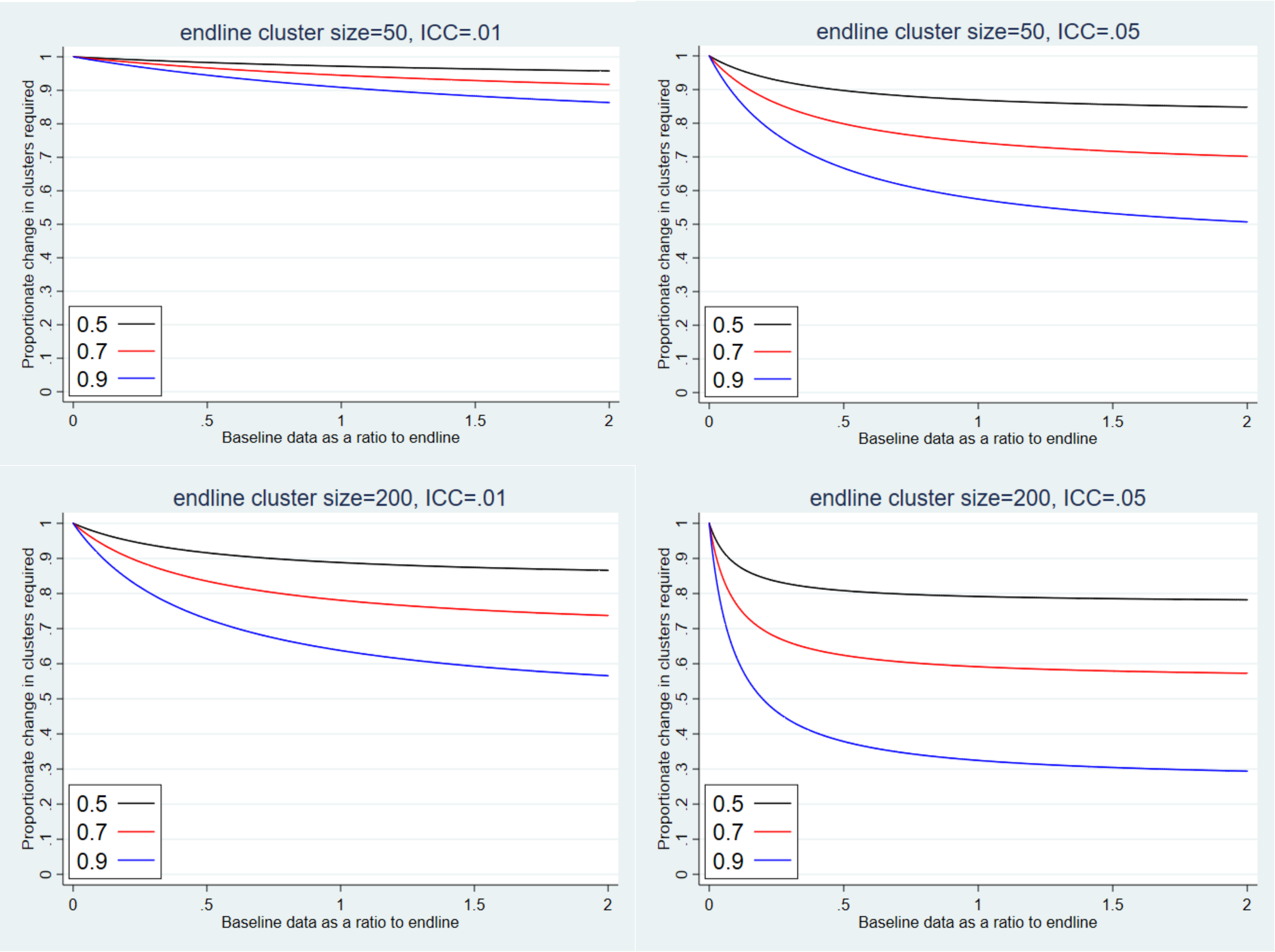

Figure 2. Proportionate change in number of clusters required according to the ratio of additional retrospective baseline data to endline data ranging from 0 to 2, for two different endline cluster sizes (n_e) and ICC values (ρ), and in each case for π=0.5, 0.7 and 0.9

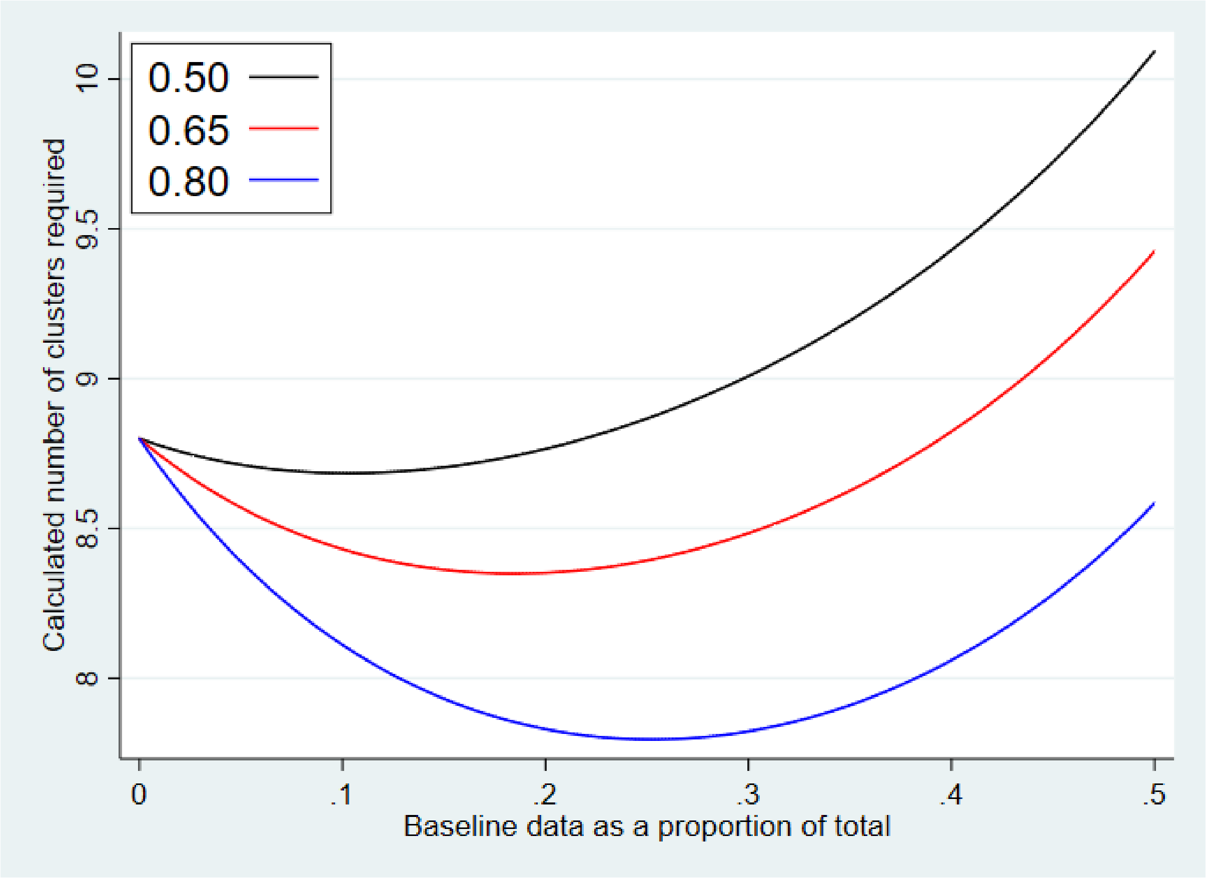

Figure 3. Effect of varying the proportion of data allocated to prospective baseline data collection on number of clusters required per arm (in practice would be rounded up) in example trial for π=0.50, 0.65 and 0.80

Appendix A

A1. Derivation of expression for the sample autocorrelation between cluster means at baseline and endline

Under the model presented in section 2.1, and using the notation of sections 2.1 and 2.2, the variance of a cluster sample mean at baseline conditional on study arm is $Var(\bar{Y}_b) = \sigma_u^2 + \sigma_\varepsilon^2 / n_b$ and similarly at endline is $Var(\bar{Y}_e) = \sigma_u^2 + \sigma_\varepsilon^2 / n_e$

The variance terms above can be re-expressed, noting that

$$\sigma_\varepsilon^2 = \frac{1-\rho}{\rho}\sigma_u^2.$$

For example

$$Var(\bar{Y}_b) = \frac{\sigma_u^2}{\rho n_b}[1 + (n_b - 1)\rho]$$

and a similar expression can be derived for $Var(\bar{Y}_e)$.

The covariance between cluster means at baseline and endline arises only though the cluster random effects as the other random terms are independent in our setting where different individuals are measured at baseline and endline. Specifically

$$Cov(\bar{Y}_b, \bar{Y}_e) = Cov(u_0, u_1) = \pi\sigma_u^2,$$

and hence

$$Corr(\bar{Y}_b, \bar{Y}_e) = \frac{\pi\sigma_u^2}{\sqrt{Var(\bar{Y}_b)}\sqrt{Var(\bar{Y}_e)}}$$

$$= \frac{\pi\rho\sqrt{n_b n_e}}{\sqrt{1 + (n_b - 1)\rho}\sqrt{1 + (n_e - 1)\rho}}$$

A2. Derivation of design effects

Our derivation is based on the work of Borm *et al.*,[20] who show that for an individually randomised trial if baseline data are available from the same individuals then the number of individuals required can be reduced by multiplying by a factor $1-r^2$, where r is the correlation between the baseline and follow-up measurements. This assumes an ANCOVA analysis, comparing the values of the outcome at endline between trial arms whilst adjusting for the baseline values of the outcome through a linear regression model.

For our trial setting we consider a cluster summary ANCOVA analysis, comparing the cluster means at endline between trial arms and adjusting for cluster means at baseline. This corresponds directly to the individually randomised trial setting, but now the interpretation is that the number of clusters required can be reduced due to the baseline data by multiplying by $1-r^2$, where r is now the correlation between the cluster sample means at baseline and endline.[1,10] Note that in Appendix A1 we see that the variance of the sample means may not be the same at baseline and endline, but Borm *et al*. acknowledge the possibility of unequal variance in their derivation. This leads directly to a design effect if the baseline data have already been collected and do not therefore need to be included within the target sample size for the trial. The overall design effect is a product of the design effect for a standard cluster randomised trial with endline data only, and then the reduction factor due to the baseline data:

$$DE_{add} = [1 + (n_e - 1)\rho][1 - r^2].$$

Next we consider the setting in which the baseline data do need to be collected as part of the trial. Noting that design effects are used to calculate a total number of individuals required (and subsequently to calculate the number of clusters by dividing by the cluster size) then it follows that the design effect here can be derived from DE_{add} by adding a simple multiplication by a factor of $[(n_b + n_e)/n_e]$ to recognise the proportionate increase in the number of individuals now that those measured at baseline need to be included:

$$DE_{within} = [1 + (n_e - 1)\rho][1 - r^2][(n_b + n_e)/n_e].$$

Appendix B1

To help consider the impact of varying the proportion of baseline measurements $\theta = n_b / m$, where $m = n_b + n_e$, we can express the DE equivalently

$$DE_{within} = [1 + ((1 - \theta)m - 1)\rho][1 - r^2][1/(1 - \theta)]$$

and also re-express *r* thus

$$r = \frac{\pi\rho m\sqrt{\theta(1 - \theta)}}{\sqrt{1 + ((1 - \theta)m - 1)\rho}\sqrt{1 + (\theta m - 1)\rho}}$$

Appendix B2

To see the impact of the ratio of baseline measurements to endline measurements, $\lambda = n_b / n_{e,}$ we can re-express *r* thus

$$r = \frac{\pi\rho n_e\sqrt{\lambda}}{\sqrt{1 + (\lambda n_e - 1)\rho}\sqrt{1 + (n_e - 1)\rho}}.$$

Appendix C

Suppose we have cross-sectional data collected at two time points from a trial, each row is a measurement (equivalently an individual) and the following variables:

`idclus` Cluster ID number

`baseline` Indicator that measurement is taken at baseline

1=baseline; 0=endline

`endline` Indicator that measurement is taken at endline

1=endline; 0=baseline

`group` Group to which the cluster is randomised

0=randomised to be in the control condition at baseline and endline;

1= randomised to cross over from the control condition at baseline to the intervention condition at endline.

`treat` Whether the outcome was assessed under the control or intervention condition

This can be calculated from group *and* endline*: if* group *and* endline *are both 1 then* treat *is 1, otherwise 0.*

`y` Outcome (continuous)

Then Stata code to estimate the cluster autocorrelation is as follows:

```
mixed y baseline endline treat || idclus: baseline endline,
cov(exch) noconstant stddev
```

where the term cov() specifies the covariance structure for the two cluster random terms at baseline and endline. The selected structure is “exch” denoting exchangeable which implies the two random terms have the same variance and are correlated with each other.

Appendix D

Stata code to generate plots to examine the impact of varying amount of prospective baseline data collection. Here is the code for a trial in which the total cluster size $(n_b + n_e)$ is set to 200, the ICC to 0.01, and the cluster autocorrelation is considered at values 0.5, 0.7 and 0.9

```
local n 200

local rho 0.01


twoway function y = (1-`rho'+(`n'*`rho'*(1-x)))*(1/(1-x))*(1-
((0.5*0.5*`rho'*`rho'*`n'*`n'*x*(1-x))/((1+(((`n'*(1-x))-1)*`rho'))*(1+(((`n'*x)-
1)*`rho')))))/(1+((`n'-1)*`rho')), range(0 0.5) lcolor(black) || ///
function y = (1-`rho'+(`n'*`rho'*(1-x)))*(1/(1-x))*(1-
((0.6*0.6*`rho'*`rho'*`n'*`n'*x*(1-x))/((1+(((`n'*(1-x))-1)*`rho'))*(1+(((`n'*x)-
1)*`rho')))))/(1+((`n'-1)*`rho')), range(0 0.5) lcolor(red) || ///
function y = (1-`rho'+(`n'*`rho'*(1-x)))*(1/(1-x))*(1-
((0.7*0.7*`rho'*`rho'*`n'*`n'*x*(1-x))/((1+(((`n'*(1-x))-1)*`rho'))*(1+(((`n'*x)-
1)*`rho')))))/(1+((`n'-1)*`rho')), range(0 0.5) lcolor(blue) ///
ytitle("Proportionate change in clusters required") xtitle("Baseline data as a
proportion of total") ///
legend(label(1 "0.5") label(2 "0.6") label(3 "0.7") pos(10) ring(0) forcesize symxsize(8)
symysize(1) rowgap(1) size(large) colgap(1) symplacement(left) textfirst cols(1)
colfirst)
```

Next the code to generate plots to examine the impact of the amount of retrospective data. Here is the code for a trial, where the prospective (i.e. endline) cluster size (n_e) is set to 200, the ICC to 0.01 and the cluster autocorrelation considered at values 0.5, 0.7 and 0.9

```
local n 200

local rho 0.01


twoway function y = 1 - ((0.5*0.5*`rho'*`rho'*`n'*`n'*x)/((1+((`n'-
1)*`rho'))*(1+((((`n'*x))-1)*`rho')))), range(0 2) yscale(range(0)) ylabel(0(0.1)1)
lcolor(black) || ///
function y = 1 - ((0.7*0.7*`rho'*`rho'*`n'*`n'*x)/((1+((`n'-1)*`rho'))*(1+((((`n'*x))-
1)*`rho')))), range(0 2) yscale(range(0)) ylabel(0(0.1)1) lcolor(red) || ///
function y = 1 - ((0.9*0.9*`rho'*`rho'*`n'*`n'*x)/((1+((`n'-1)*`rho'))*(1+((((`n'*x))-
1)*`rho')))), range(0 2) yscale(range(0)) ylabel(0(0.1)1) lcolor(blue) ///
title("size=`n', ICC=`rho'") ytitle("Proportionate change in clusters required")
xtitle("Baseline data as a ratio to endline data") ///
legend(label(1 "0.5") label(2 "0.7") label(3 "0.9") pos(7) ring(0) forcesize symxsize(8)
symysize(1) rowgap(1) size(large) colgap(1) symplacement(left) textfirst cols(1)
colfirst)
```