Cluster randomised trials with repeated cross-sections: some alternatives to parallel group designs

HOOPER, Richard and BOURKE, Liam <http://orcid.org/0000-0002-6548-4603>

Cluster randomised trials with repeated cross sections: alternatives to parallel group designs

Richard Hooper,[1] Liam Bourke[2]

[1]Centre for Primary Care & Public Health, Queen Mary University of London, London E1 2AB, UK

[2]Health and Wellbeing Research Institute, Sheffield Hallam University, Sheffield, UK

Correspondence to: R Hooper r.l.hooper@qmul.ac.uk

Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmj.h2925)

Cite this as: *BMJ* 2015;350:h2925
doi: 10.1136/bmj.h2925

Accepted: 9 March 2015

Clinical trials need not use parallel group designs to assess the effect of an intervention. This article considers alternative designs for cluster randomised trials with repeated cross sections that could reduce the number of clusters and participants required

In 1948, the UK Medical Research Council's streptomycin trial established the principles of the modern clinical trial,[1] and for longer still the idea of a comparison group recruited concurrently to the intervention group has been recognised as essential to obtaining sound evidence for clinical effectiveness.[2] But must a trial proceed by running an intervention and comparator in parallel? In this article, we focus on trials where participants are randomised in clusters such as general practices or schools. This format is common when evaluating interventions applied at cluster level.[3] We look at trials where the comparator is routine care, which effectively ask how individuals' outcomes would compare before and after introducing the intervention in a cluster. We discuss efficient alternatives to parallel group trial designs in this case—made possible by delaying introduction of the intervention in some clusters after randomisation, with these clusters continuing in the meantime to receive routine care.

Example study

Murphy and colleagues evaluated a free school breakfast programme in Wales using a cluster randomised trial with schools as clusters.[4] At each of the 111 schools, a baseline sample of about 50 children aged 9-11 years completed assessments of behaviour, cognitive performance, and diet. Half the schools were randomised to receive the intervention. One year later, a new sample of children aged 9-11 years was taken at each school. Because a different sample of children was taken on each occasion, we would describe the design as involving repeated cross sections.[5] We refer to the particular design used by Murphy and colleagues as a parallel group design with a baseline assessment (fig). It is analogous to an analysis of covariance design for an individually randomised trial.[6 7]

In this article, we consider sample size requirements for cluster randomised trials with a variety of designs involving repeated cross sections. In particular, we focus on designs where the introduction of the intervention is delayed for longer in some arms than others. In this case, we make the fundamental assumption that the effect of the intervention—that is, the difference in outcome between individuals in clusters who receive the intervention and individuals in clusters continuing to receive routine care—is the same regardless of that delay. Clusters must be able to provide a fresh sample of eligible individuals at each new cross section. We assume a multilevel model where the levels are clusters, cross sections, and individuals.

SUMMARY POINTS

- Clinical trials need a control, but if handled correctly this need not run in parallel with the intervention
- There are various designs for cluster randomised trials involving more than one cross section
- Multiple cross sections mean fewer clusters are required, but could result in a heavy burden of individual recruitment
- Nevertheless, it is possible to add a cross section to a design with a single follow-up and reduce both the number of clusters and the number of individual participants needed

Designs for cluster randomised trials with repeated cross sections

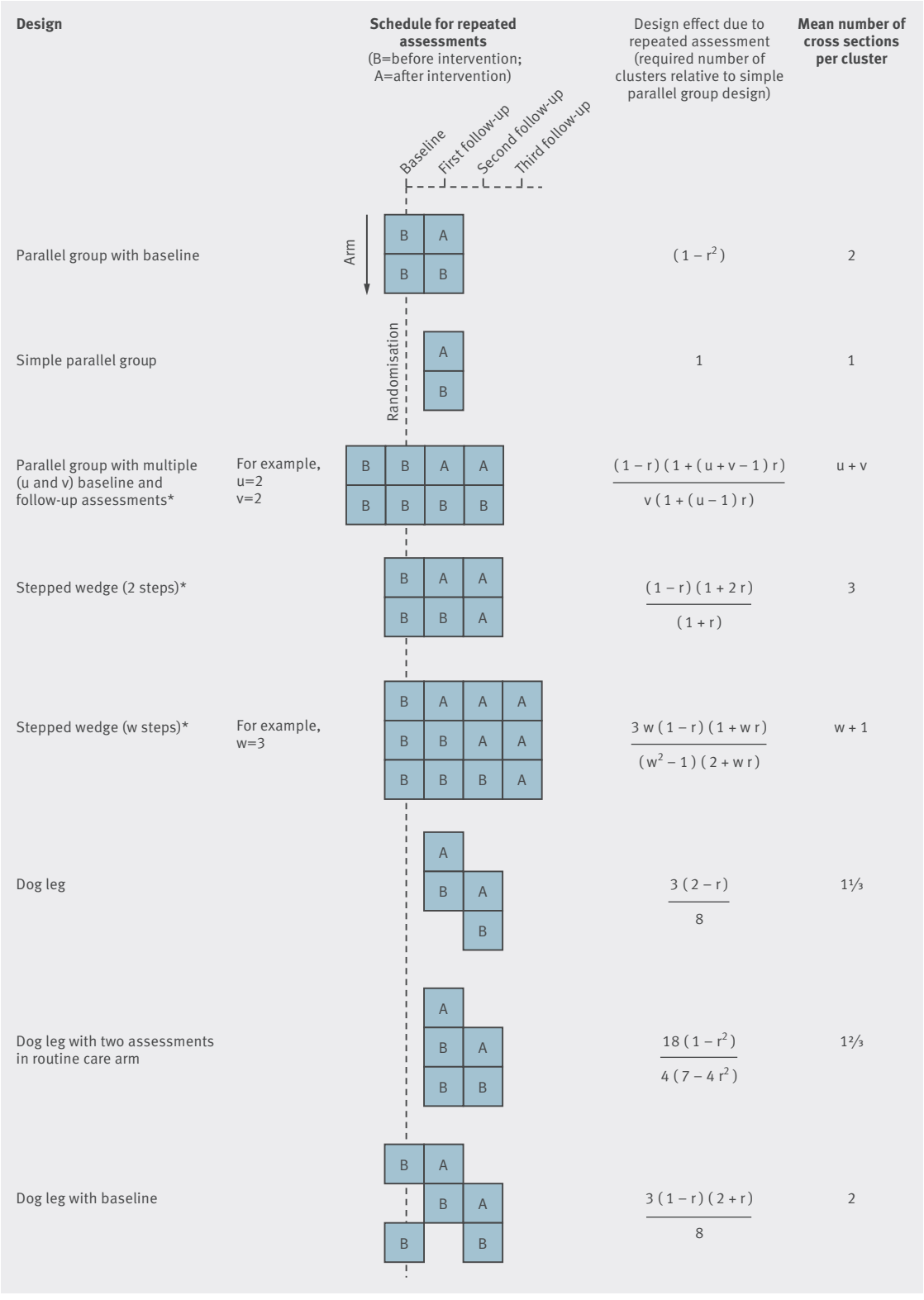
Crossover versus cross forward designs

An alternative to a parallel group design is a crossover design in which intervention and control arms swap over at the end of the first intervention period, with clusters in the control arm then receiving the intervention, and clusters in the intervention arm returning to routine care.[8] This design would not be appropriate if there is a risk of "carry over" within clusters, whereby the clusters have supposedly been returned to routine care, but in fact continue to pass on some of the effects of the intervention to individuals. Nor would it be appropriate if the intervention is being rolled out as part of a policy change, such as in the school breakfast example. In this article, we consider designs where the transition is in one direction only, from routine care to the new intervention. These are sometimes known as one way crossover or unidirectional crossover designs. We suggest a more simple description: cross forward designs.

Parallel group designs

The figure illustrates several cross forward designs. At the top is the parallel group design with a baseline assessment. Using "B" for "before intervention" and "A" for "after intervention," we code the schedule of assessments in this case as BA in one arm and BB in the other (the control arm receives routine care throughout, and is therefore still "before intervention" at follow-up).

The baseline assessment can be discarded, leaving a simple parallel group design. Sample size calculations for this design are particularly straightforward.[9] We use this design as a reference for comparing sample size requirements of other designs with repeated assessments of clusters. In the figure, each design shows the



Designs for cluster randomised trials comparing outcomes before and after the introduction of a new intervention. Figure shows the schedule for repeated assessments and design effect due to repeated assessment (assuming equal numbers of clusters in each arm) according to the correlation, *r*, between two sample means from the same cluster at different times. *Design effect assumes the effect of the intervention is maintained at the same level once it has been introduced

relative number of clusters needed to achieve the same statistical power as a simple parallel group design. We call this the "design effect" due to repeated assessment. In each case, it depends only on the correlation, *r*, between two sample means from the same cluster in different cross sections (web appendix). Design effects are derived for normally distributed, continuous outcome measures but can also be applied to binary outcomes.[10]

Stepped wedge designs

If, in a parallel group design with baseline, we give the control clusters the intervention after the first follow-up and then follow up both arms a second time, we end up with another kind of cross forward design: the stepped wedge—in this case one with two steps (fig). Stepped wedge designs can have any number of steps up to and including the total number of clusters, and deliver the

doi: 10.1136/bmj.h2925 | *BMJ* 2015;350:h2925 | the bmj

intervention to all clusters according to a staggered timetable that varies with trial arm, where "arm" now simply refers to a randomised group.[11 12] Stepped wedge designs are relatively new: a recent systematic review found only 25 stepped wedge trials, with all but two published since 2000.[13]

Incomplete cross forward designs

Stepped wedge designs need fewer clusters than parallel group designs with a single follow-up, simply because they assess the same clusters repeatedly. Alternatives such as parallel group designs with multiple baseline or follow-up assessments offer a similar advantage. (Among designs with a fixed number of repeated cross sections, the particular design which minimises the required number of clusters depends on the circumstances,[12] and further research is needed.) The advantage of multiple cross sections is offset, however, by having to recruit a new sample of individuals each time. We quantify some of these sample size problems in the next section, but in this section we consider how we might reduce both the number of clusters and the number of individual participants required while increasing the number of cross sections. This will be worthwhile when trial costs are determined mostly by the numbers of individuals and clusters involved rather than by the duration of follow-up.

Incomplete cross forward designs leave gaps in the assessment schedule in some trial arms, requiring fewer individuals to be recruited.[14] The simplest incomplete cross forward design is the dog leg, named after the shape made by the assessment schedule (fig).[15] This design has no baseline assessments (that is, assessments at or before randomisation). Clusters in the first arm are assessed after receiving the intervention, but are not assessed again (they may or may not continue to receive the intervention, depending on the context). Clusters in the second arm are assessed after a period of routine care, and assessed again after receiving the intervention. Clusters in the third arm receive routine care throughout, and are assessed once, at the second follow-up.

Elaborations to the dog leg might also be worth investigating (fig). The most obvious place in the schedule for an additional assessment is at the first follow-up in the third arm.[15] A less obvious modification, ensuring that each cluster is assessed twice, is to add a baseline assessment (before randomisation) to the first and third arms. Dog leg designs are a recent methodological development, and have not been used for trials so far.

Table 1 | Steps in the calculation of sample size for a cluster randomised trial with repeated cross sections

Method	School breakfast programme example[4]
1. Specify values for:	
Sample size in each cluster at each cross section (m)	50
Intracluster correlation (ρ)	0.02
Cluster autocorrelation (π)	0.8
2. Determine from tables or by calculation:	
No of participants required for an individually randomised, simple parallel group design (n_0)	2600
3. Choose a design for the repeated assessments	Simple parallel group with baseline
4. Determine for this design:	
No of trial arms (K)	2
Mean number of cross sections per trial arm (s)*	2
5. Calculate the following:	
Design effect due to cluster randomising (d_c)†	1.98
Correlation between two sample means from the same cluster at different times (r)‡	0.4040
Design effect due to repeated assessment (d_r)§	0.8368
6. Required sample size is:	
No of clusters ($n_0 \times d_c \times d_r \div m$)	88 (rounded up to a multiple of K)
No of participants ($m \times s \times$ no of clusters)	8800

*As shown in the final column in the figure.
†$d_c = 1+(m-1)\rho$.
‡$r = m\rho\pi \div d_c$.
§Calculated using the relevant formula from the figure, with r calculated above.

Sample size calculation

The steps involved in calculating required sample size for a cluster randomised trial with repeated cross sections are described in table 1. In the school breakfast trial,[4] Murphy and colleagues wanted 80% power at the 5% significance level to detect an effect size (ratio of mean difference to standard deviation) of 0.11 for their continuous outcomes. They planned to assess the outcomes of 50 children in each cross section at each school, and assumed an intracluster correlation of 0.02.

We start with the sample size for an individually randomised trial with simple parallel group design; using standard methods or tables, this is determined to be about 2600.[16 17] The relative adjustment required for a cluster randomised trial—the design effect due to cluster randomising—is well known in this case,[9] and evaluates to 1.98 (table 1).

We follow Murphy and colleagues[4] and assume a parallel group with baseline design. The next step is to calculate the correlation, r, between two sample means from the same cluster at different times. This correlation depends on the sample size in each cluster at each cross section, on the intracluster correlation,[9] and on the reliability of the cluster population mean over time, also known as the cluster autocorrelation.[7] Murphy and colleagues did not estimate a cluster autocorrelation. In the methodological literature, this number is sometimes assumed to be 1,[11] but this assumption will underpower repeated cross section trials when individuals sampled from the same cluster at *different* times are more heterogeneous than individuals sampled from the same cluster at the *same* time. As with the intracluster correlation, we could use different values for the cluster autocorrelation to see its effect on the required sample size, or calculate a value using similar data from completed trials. Here, we assume a cluster autocorrelation of 0.8, which gives r=0.4040, and the design effect due to repeated assessment is then 0.8368 (table 1).

We multiply our initial sample size of 2600 by the respective design effects due to cluster randomising and repeated assessment, and divide by the cluster size, giving 88 clusters in all (rounded up to a multiple of two since there are two arms). In each arm, we take two repeated cross sections of 50 children each, so that the total number of participants required is 8800. Table 2 shows sample size requirements for other trial designs.

Comparison of designs

Using the school breakfast programme example,[4] the dog leg design requires fewer schools and fewer participants than a simple parallel group design or parallel group design with baseline. Murphy and colleagues proposed a trial of 11 100 children in 111 schools; a dog leg design requires just 4200 children in 63 schools. In fact, a dog leg design always requires fewer clusters and fewer participants than a simple parallel group design, and likewise a dog leg with baseline always requires fewer clusters and fewer participants than a parallel group design with baseline (web appendix). Modifying the dog leg by adding another follow-up in the routine care group confers little advantage when r is moderate, as our example illustrates.

Risk of bias should be considered carefully with any design. The fundamental assumption of cross forward designs—that the effect of the intervention is the same if there is a delay following randomisation—may not always hold. For example, if clusters have to perform poorly in some sense to be eligible for the trial, they may show a natural improvement over time, and thus offer less room for the intervention to show its effect. In incomplete designs, there is a risk of differential attrition of clusters in different arms: in a dog leg trial, for example, clusters in the second and third arms are followed for longer than clusters in the first arm, and clusters in the third arm may have little contact with researchers during the first follow-up period, making them more readily lost to follow-up. Because incomplete designs involve different patterns of assessments in different arms, they also assume implicitly that the pattern or frequency of previous assessments at cluster level does not influence subsequent individual outcomes. Such an influence would be unlikely since different individuals are assessed each time, but could arise if, for example, staff at a cluster changed their behaviour after observing assessments. In addition, if outcome data from multiple cross sections can be obtained at little cost—such as from a pre-existing anonymised database—then incomplete designs lose their appeal.

Conclusions

Clinical trial designs where the same clusters of participants are assessed in more than one cross section (allowing intervention clusters to be compared not only with parallel controls but also with themselves under an earlier control condition) need fewer clusters than a trial with a single cross section, but might also need more participants overall. If investigators want to minimise the overall number of participants and are willing to increase the number of cross sections, an incomplete design could be worth considering. A dog leg design run over two repeated cross sections, for example, needs fewer clusters and fewer participants in total than a trial with a single cross section.

Sample size calculations for cluster randomised trials with repeated cross sections require a cluster autocorrelation to be specified in addition to the intracluster correlation, and allows for variation over time within a cluster in addition to variation between clusters. Calculations that ignore the cluster autocorrelation (such as those ignoring the intracluster correlation) risk underpowering a trial. Routinely collected time series data, as they become more widely available, should help researchers quantify cluster autocorrelations and intracluster correlations, as well as highlighting secular trends in outcomes under routine care.

Despite methodological challenges and risk of bias, efficient trial designs such as incomplete cross forward designs have an important role. These designs can help researchers meet ethical and financial requirements to limit numbers of participants in research,[18] as well as create opportunities for research in small populations or rare conditions.[19]

Table 2 | Comparison of sample size requirements for different trial designs in the school breakfast programme example[4]*

Design for repeated assessments†	Schools	Participants	No of follow-up times
Parallel group with baseline	88	8800	1
Simple parallel group	104	5200	1
Stepped wedge (2 steps)	80	12 000	2
Stepped wedge (3 steps)	48	9600	3
Stepped wedge (4 steps)	36	9000	4
Dog leg	63	4200	2
Dog leg with two assessments in routine care arm	63	5250	2
Dog leg with baseline	57	5700	2

*To achieve 80% power at the 5% significance level to detect a mean difference equal to 0.11 standard deviations, assuming cluster size 50, intracluster correlation 0.02, and cluster autocorrelation 0.8.
†As shown in the figure.

We are indebted to an anonymous reviewer of a previous article who suggested modifying the dog leg design by adding a baseline assessment in two of the groups.

Contributors: Both authors drafted the manuscript. RH is the guarantor.

Funding: The work summarised in this article was not funded by a specific grant. RH is supported by the UK National Institute for Health Research as part of its funding for the Pragmatic Clinical Trials Unit at Barts & the London School of Medicine & Dentistry. LB is supported by the Higher Education Funding Council for England.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and declare we have no competing interests.

Data sharing: A technical appendix is available through the BMJ website or on request from the corresponding author.

1 Medical Research Council. Streptomycin treatment of tuberculous meningitis. *Lancet* 1948;1:582-96.
2 Hrobjartsson A, Gotzsche PC, Gluud C. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. *BMJ* 1998;317:1243-5.
3 Donner A, Klar N. Cluster randomization trials. *Stat Methods Med Res* 2000;9:79-80.
4 Murphy S, Moore GF, Tapper K, et al. Free healthy breakfasts in primary schools: a cluster randomised controlled trial of a policy intervention in Wales, UK. *Public Health Nutr* 2010;14:219-26.
5 Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Wiley, 2012:85-7.
6 Frison L, Pocock SJ. Repeated measures ion clinical trials:analysis using mean summary statistics and its implications for design. *Stat Med* 1992;11:1685-704.
7 Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;31:2169-78.
8 Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. *Trials* 2009;10:27.
9 Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998;316:1455.
10 Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999;55:137-48.

11 Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28:182-91.
12 Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;350:h391.
13 Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;64:936-48.
14 Hemming K, Lilford R, Girling AJ. Stepped wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2015;34:181-96.
15 Hooper R, Bourke L. The dog-leg: an alternative to a cross-over design for pragmatic clinical trials in relatively stable populations. *Int J Epidemiol* 2014;43:930-6.
16 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;311:1145-8.
17 Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies. 3rd ed. Wiley-Blackwell, 2009.
18 Altman DG. Statistics and ethics in medical research III: how large a sample? *BMJ* 1980;281:1336-8.
19 Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: methodological review. *BMJ* 2014;349:g6802.

Web appendix: Design effects due to repeated assessment in longitudinal trial designs