The dog-leg: an alternative to a cross-over design for pragmatic clinical trials in relatively stable populations

Richard Hooper,
Senior Lecturer in Medical Statistics

Liam Bourke,
Lecturer in Public Health Research

Centre for Primary Care & Public Health,
Queen Mary University of London

Address for correspondence:

Dr Richard Hooper,
Centre for Primary Care & Public Health,
Yvonne Carter Building,
58 Turner Street,
London E1 2AB,
UK

phone: 020 7882 7324
fax: 020 7882 2552
email: r.l.hooper@qmul.ac.uk

Word count: 2891 (manuscript) + 236 (abstract)

Key messages

We propose a novel clinical trial design called a dog-leg design which, like a cross-over design, runs over two consecutive intervention periods.

The dog-leg design does not require the effect of the experimental intervention in the first period to have disappeared in the second.

In some realistic situations the dog-leg design is more efficient than a parallel groups design with a baseline assessment.

The dog-leg design also requires fewer assessments in total than a parallel groups design where participants are only assessed once, at follow-up.

Summary

Background

A cross-over trial design is more powerful than a parallel groups design, but requires that treatment effects do not carry over from one period of the trial to the next. We focus here on interventions in chronic disease populations where the control is routine care: in such cases we cannot assume the intervention effect is easily washed out in crossing over from the experimental intervention back to the control.

Methods

We introduce an alternative trial design for these situations, and investigate its performance. One group is assessed before and after the experimental intervention, while two other groups provide respective, independent treatment comparisons in each period. We call this a dog-leg design because of the pattern of assessments in the three groups. The dog-leg design is reminiscent of a stepped-wedge design, but with a reduced schedule of assessments, and with the notable difference that not all groups receive the intervention.

Results

If the correlation between baseline and follow-up is <0.72, the dog-leg design is more efficient than a parallel groups design with a baseline assessment. The dog-leg design also requires fewer assessments in total than a parallel groups design where participants are only assessed once, at follow-up.

Conclusions

The dog-leg design is simple, and has some attractive properties. Though there is a risk of differential attrition in the three arms, the design's good performance relative to alternatives makes it a useful addition to the methodologist's toolkit.

Keywords: clinical trial design, cross-over trials, stepped wedge, dog-leg

Introduction

Cross-over trials – clinical trials in which participants act as their own controls and are randomised to the order in which they get the experimental and control treatments – are contraindicated when the effects of treatment could carry over from one period of the trial to the next.[1] They can work well for pharmacokinetic trials,[2] but are problematic for interventions whose effects are intended to be durable.

In this article we focus on pragmatic trials in chronic disease populations, or other populations with relatively stable health, and where the comparator is routine care. We are interested in this case in how patients would improve with the introduction of the new intervention. While it would be tempting to run a trial where we simply assess the same group of participants pre- and post-intervention, the flaw in such a design is that patient outcomes could change naturally over time following recruitment into the trial, even in a relatively stable population. A comparison group is needed who are assessed at the same time following recruitment but who have not had the intervention.

The simplest approach is to randomise participants into two groups – one receiving the experimental intervention, the other routine care – and assess both at the end of the intervention period. This is the standard, parallel groups design. Such a trial can be done with or without a baseline assessment of outcome. The degree of advantage in including a baseline assessment depends on the correlation between baseline and follow-up assessments: the greater the correlation, the greater the advantage.[3] The most efficient way of analysing a parallel groups design with a baseline assessment is with an analysis of covariance (ANCOVA), though an analysis of change scores is also common.[4]

A cross-over trial, which uses participants as their own controls, achieves more power than a parallel groups trial, but assumes that the effect of the first period of the trial has disappeared by the time that the measurement of the effect in the second period is taken. Other designs allow for the possibility of carry-over of treatment effects. Balaam's design, which consists of four arms with one for every combination of intervention and control in the first and second periods, is optimal under a simplifying assumption that the carry-over effect of having had the experimental intervention in the first period is the same whether you have the experimental intervention or the control in the second period, and similarly for the carry-over effect of the control.[5] This simplifying assumption is, however, suspect for an intervention whose effect we think may largely be maintained once it has been introduced.

In this article we present a design which makes no assumptions about the carry-over effect of the experimental intervention. We are interested in the difference in outcome between intervention and routine care in a population previously treated with routine care. We will assume this difference does not depend on the duration of previous routine care, or equivalently that there is no carry-over effect of routine care, though we will allow the possibility of an overall effect of time since randomization, that is a natural shift in outcomes over time in the population of interest. The asymmetry in our assumptions about carry-over of intervention and control may be reasonable when the control is routine care and the population is relatively stable, but may be harder to justify otherwise.

Methods

Design

Figure 1 illustrates the design. Figure 1(a) shows the timing of the intervention and assessments post-randomisation. Routine care is assumed to be given up to the point that any intervention is indicated. Figure 1(b) is a schema for the schedule of assessments, distinguishing between assessments conducted before and after the intervention

Participants are randomised into three arms: in the central arm (group 2) participants act as their own controls, being assessed both before and after the experimental intervention. So that effects of treatment and time since randomisation can be disentangled, two additional arms provide a treatment comparison in each period of the trial: these are not repeated in the same individuals, but rather in independent groups in order to avoid having to assume the intervention effect can be washed out. We refer to this as a dog-leg design because of the pattern formed by assessments in the three groups ("dog-leg" in English refers to something that is bent or crooked like a dog's hind leg).

Like a cross-over trial, a dog-leg trial must be run over two consecutive periods of intervention and follow-up. The dog-leg design is reminiscent of a stepped wedge design (a form of pre-post design in which the introduction of the intervention is staggered over a number of groups and time intervals),[6] but with a reduced schedule of assessments, and with the notable difference that not all of the groups receive the intervention. We assume from the symmetry of the design that it is optimal to assign equal proportions of participants to groups 1 and 3, but we do not make any prior assumption as to what this proportion should be.

[Figure 1 about here]

Statistical model

Suppose there are N participants in total, with the same number n in each of groups 1 and 3. We consider a random-effects model for a Normally-distributed outcome y_{ijk} for individual j in group i at time k, with baseline mean μ, effect of time since randomisation β, treatment effect θ, and variances between and within participants σ_b^2 and σ_w^2, respectively.

Then

$$\begin{aligned} y_{1j1} &= \mu + \theta + \eta_{1j} + \varepsilon_{1j1}, & j &= 1, \ldots, n; \\ y_{2j1} &= \mu + \eta_{2j} + \varepsilon_{2j1}, & j &= 1, \ldots, N\text{-}2n; \\ y_{2j2} &= \mu + \beta + \theta + \eta_{2j} + \varepsilon_{2j2}, & j &= 1, \ldots, N\text{–}2n; \\ y_{3j2} &= \mu + \beta + \eta_{3j} + \varepsilon_{3j2}, & j &= 1, \ldots, n; \end{aligned}$$

where

$$\eta_{ij} \sim \mathrm{N}(0, \sigma_b^2),$$
$$\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma_w^2),$$

or in vector notation

$$\mathbf{y} = \mathbf{Z}\begin{pmatrix} \mu \\ \beta \\ \theta \end{pmatrix} + \mathbf{e},$$

where **e** is multivariate Normal with mean **0** and block-diagonal variance-covariance matrix **V**.

The standard error of the treatment effect could be derived using the theory of generalised least squares estimators, whereby the variance-covariance matrix of the estimator $\left(\hat{\mu}, \hat{\beta}, \hat{\theta}\right)'$ is given by $(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$, but for the model specified above the standard error can be derived without recourse to matrix inversion by directly determining the minimum variance unbiased estimator (see Appendix). By comparing this standard error with the standard error from an ANCOVA in a parallel groups design (more specifically by calculating the ratio of the variances)[3] we determined the relative efficiency of the two designs – that is, the inverse ratio of sample sizes required for the two designs to achieve the same statistical power. We also used the standard error of the treatment effect to construct sample size tables for dog-leg trials.

Correlation between baseline and follow-up – a systematic review

The relative advantages of different trial designs depend on the correlation between assessments before and after intervention in the same individual. Sample size calculations for trials have tended to assume this correlation is moderate.[3] To provide information for this article on typical correlations we systematically reviewed clinical trials published in the *Lancet* or *New England Journal of Medicine* between 01 January 2012 and 31 March 2013. Details of methods and results are given in the online supplement.

Results

Relative efficiency

The estimate of the treatment effect from a dog-leg design is simply

$$\hat{\theta} = (\bar{y}_{11} - \bar{y}_{21} + \bar{y}_{22} - \bar{y}_{32})/2,$$

where $\bar{y}_{ik}$ is the mean outcome in group *i* in period *k* (see Appendix). The standard error of this estimate depends not only on the correlation between baseline and follow-up, but on the proportions of participants allocated to the three arms. If the proportion of participants allocated to each of groups 1 and 3 is *p*, and the correlation between assessments before and after the intervention in the same individual is *r*, then the standard error is

$$\sqrt{\frac{\sigma^2(1-p(1+r))}{2Np(1-2p)}},$$

where σ is the standard deviation of the outcome before or after intervention, and N is the total sample size. Note that using our earlier notation for the relevant variance components, $\sigma^2 = \sigma_b^2 + \sigma_w^2$ and $r = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2)$.

The standard error of the treatment effect from an ANCOVA in a parallel groups design with total sample size N and 1:1 allocation ratio is[3]

$$\sqrt{\frac{4\sigma^2(1-r^2)}{N}}.$$

Figure 2(a) shows the relative efficiency of the dog-leg design compared with a parallel groups design analysed with ANCOVA, plotted against p for various values of r. The best choice of p depends on r, but close-to-optimal performance can be achieved over a wide range of r-values by choosing p=1/3, that is by allocating participants to arms in the ratio 1:1:1.

[Figure 2 about here]

Figure 2(b) shows the relative efficiency plotted against r for allocation ratios 1:1:1 (p=1/3), 1:2:1 (p=0·25) and 2:1:2 (p=0·4). This confirms p=1/3 as a sensible choice in general. As long as r<0·72 the dog-leg design with allocation ratio 1:1:1 is more efficient than a parallel groups design analysed with ANCOVA, and may be up to 43% more efficient (*i.e.* a parallel groups design analysed with ANCOVA could need up to 43% more participants than a dog-leg design to achieve the same statistical power).

Compared to a parallel groups design with *no* baseline assessment,[3] the relative efficiency of a 1:1:1 dog-leg design is even better, being always at least 133%. Since participants in a dog-leg trial each get an average of 1·33 assessments, this means the total number of assessments required in a dog-leg design is smaller even than the number required for a parallel groups design where participants are assessed only once, at follow-up.

Sample size

Using the above expression for the standard error of the treatment effect, we constructed a table of the sample size per group needed to achieve 80% or 90% power at the 5% significance level with a dog-leg design, assuming a variety of effect sizes and correlations between baseline and follow-up, and with a 1:1:1 allocation ratio to the three arms (Table 1). Sample size tables for parallel groups trials with and without baseline assessments are widely available.[7]

[Table 1 about here]

For example, an investigator wants to conduct a trial of a new intervention for people with Parkinson's disease, and would like 80% power to detect a difference of 6 scale

points on a quality of life measure with a standard deviation of 15 scale points (an effect size of 0·4). Suppose other work with this measure suggests that the correlation between quality of life before and after treatment in the same individual is 0·6. A parallel groups design with no baseline would require 200 participants in total (100 of whom would receive the intervention) and 200 assessments in total.[7] A parallel groups design with baseline assessment would require 128 participants in total (64 of whom would receive the intervention) and 256 assessments in total.[7] From Table 1, a dog-leg design requires just 108 participants in total (72 of whom receive the intervention) and just 144 assessments in total.

Correlation between baseline and follow-up

Of 250 clinical trials published in the *Lancet* and *New England Journal of Medicine* between 01 January 2012 and 31 March 2013, 36 met the criteria specified in our systematic review. We were able to calculate 30 correlations from 23 of these trials (see online supplement): these correlations had a median of 0·71, with interquartile range 0·59 to 0·80.

Discussion

Adding further assessments

As with any design, the power of the dog-leg could potentially be increased by adding assessments. Not all such augmentations are helpful, however. A second assessment in group 1 (period 2) could not be used to estimate the treatment effect because we are assuming nothing about the carry-over effect of the experimental intervention. A baseline assessment (at the time of randomisation) in group 1 is also of little help by itself if no other group is assessed at randomisation for comparison, since we are assuming a possible effect of time since randomisation.

The most obvious choice for an additional assessment is in group 3, period 1 (Figure 3). Using the generalised least squares, matrix algebra approach described in the Methods, we determined the standard error of the treatment effect in this augmented dog-leg design to be

$$\sqrt{\frac{\sigma^2(p_2+p_3)(1-r^2)}{N((p_2+p_3)(p_2p_3+p_1(p_2+p_3)(1-r^2))+p_1p_2p_3)}},$$

where p_1, p_2 and p_3 are the proportions of participants allocated to groups 1, 2 and 3, respectively. If we assume directly from the symmetry in p_2 and p_3 that the optimal allocation has $p_2 = p_3 = p$, say, then the standard error becomes

$$\sqrt{\frac{2\sigma^2(1-r^2)}{N(4p(1-2p)(1-r^2)+p)}}.$$

Figure 4 shows the relative efficiency of the augmented dog-leg design compared with a parallel groups design analysed with ANCOVA, for different p and r. Again a choice of 1:1:1 for the allocation ratio seems sensible for r>0·8, say (for larger r one

might as well use the ANCOVA design, which is essentially equivalent to p=0.5). Figure 5 shows the relative efficiency of the 1:1:1 augmented dog-leg design compared with the 1:1:1 dog-leg design. Interestingly, when r=0·5 there is no advantage to the augmented design. At other r-values there is some advantage, but for r<0·8 the relative efficiency never exceeds 125%, meaning that the total number of assessments needed for the dog-leg design is still less than the number needed for the augmented design. When assessments are expensive, invasive or painful, the number of assessments required may have at least as much practical and ethical importance as the number of participants. With invasive assessments, participants in group 3 in particular, who receive no active intervention, may question whether two assessments are really needed.

[Figure 3 about here]
[Figure 4 about here]
[Figure 5 about here]

Practicalities of running a dog-leg trial

The dog-leg design is unconventional in having a different schedule of assessments in each randomised arm. Group 1 participants are not assessed in the second period of the trial, nor is the analysis affected by the treatment they receive in this period. The trial protocol should specify how participants in group 1 are to be treated in the second period; a choice that can be made according to ethical and scientific considerations. Group 3 participants, meanwhile, are not assessed until the second period, but this is straightforward to achieve: if the duration of each period is 3 months, say, then a participant randomised into group 3 should have their first (and only) assessment scheduled at 6 months post-randomisation, just as if they were in the control arm of a parallel groups trial with 6-month follow-up.

Data from a dog-leg trial could be analysed with standard statistical software that allows a mixed regression model with a random effect of participant and fixed effects of treatment and time since randomization. This would also enable adjustment for additional covariates if required.

Dependence on the correlation

A curious feature of the dog-leg design is that its efficiency relative to a parallel groups design analysed with ANCOVA first increases, and then decreases with increasing correlation between baseline and follow-up assessments. This is because for small r the ANCOVA is able to take relatively little advantage of the baseline assessment, while the dog-leg design is able to benefit from some participants being their own controls (a benefit which increases as the correlation increases). For larger r the ANCOVA is increasingly able to adjust for individual participant differences in the whole sample, out-performing the dog-leg which assesses most of its participants only once.

Conclusions

The dog-leg design requires fewer participants for the same power than a parallel groups design with a baseline assessment if the correlation between baseline and

follow-up is not too high. From our review we conclude that correlations below 0·7 are not uncommon in studies of chronic diseases. The dog-leg design also requires even fewer assessments than a parallel groups design where participants are assessed only once, making it particularly attractive for minimising expensive or invasive assessments. In addition, the intervention is offered to the majority (2/3) of participants in a 1:1:1 dog-leg design, which may be an incentive to recruitment.

On the negative side, a dog-leg trial, like a cross-over trial, must be run over two consecutive intervention periods, increasing the risk of attrition. Moreover there is a risk of differential attrition in the three arms, which differ in the timing and nature of contact with participants. The dog-leg design can only be used if we are confident that making an assessment does not alter subsequent outcomes (some participants assessed in a given period have a previous assessment, and some do not).

Dog-leg trials seem not to have been described previously, at least not explicitly. Methodological reviews of cross-over trials do not discuss them,[1,8–9] indeed the dog-leg design violates principles of balance and uniformity usually considered optimal for cross-over designs.[10] Reviews of stepped wedge designs have asked whether designs with small numbers of steps should be considered methodologically distinct,[11–12] but this seems to apply particularly to studies with two steps rather than three: in a recent review of 25 stepped wedge trials, nine had two steps but only one had three (and this was not a dog-leg design).[12] Stepped wedge designs are often motivated by practical concerns – limited resources requiring staggering of the intervention; clustering of the participants; an intervention which was due to be rolled out anyway – whereas the dog-leg design is simply motivated by a desire to increase statistical power. Given its interesting properties and its simplicity, we think the dog-leg design is a valuable addition to the methodologist's toolkit.

Acknowledgements

We thank three reviewers of an earlier draft for their very valuable input.

Appendix

Let $\bar{y}_{ik}$ be the mean outcome in group *i* at time *k*. If we assume a treatment effect estimate of the form

$$\hat{\theta} = a\bar{y}_{11} + b\bar{y}_{21} + c\bar{y}_{22} + d\bar{y}_{32},$$

then for $\hat{\theta}$ to be unbiased we need

$$\begin{gathered} a + b + c + d = 0; \\ c + d = 0; \\ a + c = 1. \end{gathered}$$

Hence the estimate has the form

$$\hat{\theta} = a\bar{y}_{11} - a\bar{y}_{21} + (1-a)\bar{y}_{22} - (1-a)\bar{y}_{32}.$$

This has variance

$$\frac{\sigma^2}{N}\left(\frac{1}{p} + \frac{1}{1-2p} - 2a(1-a)\left(\frac{1+r}{1-2p} + \frac{1}{p}\right)\right),$$

where

$$\begin{gathered} \sigma^2 = \sigma_b^2 + \sigma_w^2, \\ r = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2), \\ p = n / N, \end{gathered}$$

which is minimised when $a = ½$ with the value

$$\frac{\sigma^2(1 - p(1+r))}{2Np(1-2p)}.$$

References

[1] Mills EJ, Chan A-W, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis and presentation of cross-over trials. *Trials* 2009; **10**: 27.

[2] Senn S, Ezzet F. Clinical cross-over trials in phase I. *Stat Methods Med Res* 1999; **8**: 263–278.

[3] Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med* 1992; **11**: 1685–1704.

[4] Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow-up measurements. *BMJ* 2001; **323**: 1123–1124.

5 Jones B, Kenward MG. *Design and analysis of cross-over trials*. 2nd edn. London: Chapman & Hall/CRC, 2003.

[6] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Clinical Trials* 2007; **28**: 182–191.

[7] Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies. 3rd edn. Chichester: Wiley-Blackwell, 2009.

[8] Huitson A, Poloniecki J, Hews R, Barker N. A review of cross-over trials. *J Royal Stat Soc D* 1982; **31**: 71–80.

[9] Senn S. Cross-over trials in Statistics in Medicine: the first '25' years. *Stat Med* 2006; **25**: 3430–3442.

[10] Matthews JNS. Multi-period cross-over trials. *Stat Methods Med Res* 1994; **3**: 383–405.

[11] Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodology* 2006; **6**: 54.

[12] Mdege ND, Man M-S, Taylor 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**: 936–948.

Figure legends

Figure 1. Schematic representations of the dog-leg design: (a) time-line from randomisation in each of the three trial arms; (b) schedule of assessments (a shaded square indicates no assessment).

Figure 2. Relative efficiency (defined as the inverse ratio of sample sizes required to achieve the same power) of the dog-leg design compared with an analysis of covariance design: (a) according to the proportion, *p*, allocated to each of groups 1 and 3; (b) according to the correlation, *r*, between outcomes assessed in the same individual before and after intervention ("ratio" on the graphs refers to the allocation ratios to the three trial arms).

Figure 3. Schematic representations of the augmented dog-leg design: (a) time-line from randomisation in each of the three trial arms; (b) schedule of assessments (the shaded square indicates no assessment).

Figure 4. Relative efficiency of the augmented dog-leg design compared with an analysis of covariance design: (a) according to the proportion, *p*, allocated to each of groups 2 and 3; (b) according to the correlation, *r*, between outcomes assessed in the same individual in different periods.

Figure 5. Relative efficiency of the 1:1:1 augmented dog-leg design compared with the 1:1:1 dog-leg design, according to the correlation, *r*, between outcomes assessed in the same individual in different periods.

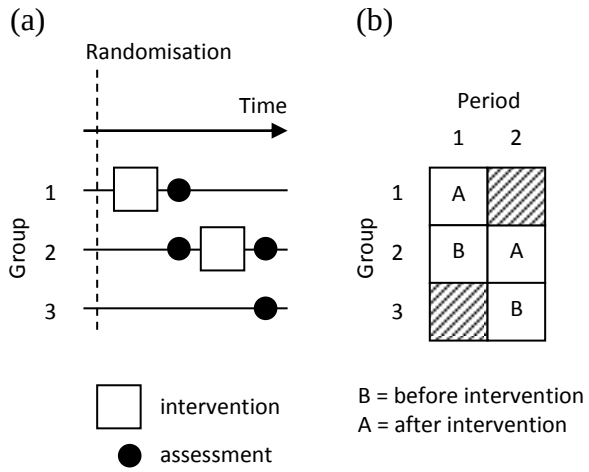
(a)
Randomisation
Time
Group
1
2
3
intervention
assessment
(b)
Period
1
2
Group
1
2
3
A
B
A
B
B = before intervention
A = after intervention

Figure 1

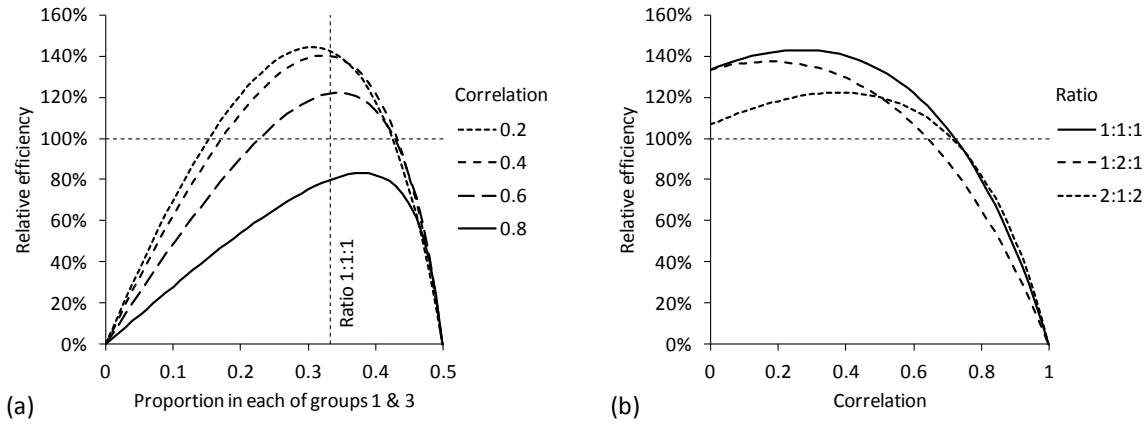
160%
140%
120%
100%
80%
60%
40%
20%
0%
Relative efficiency
0
0.1
0.2
0.3
0.4
0.5
Proportion in each of groups 1 & 3
Ratio 1:1:1
Correlation
0.2
0.4
0.6
0.8
(a)
0
0.2
0.4
0.6
0.8
1
Correlation
Ratio
1:1:1
1:2:1
2:1:2
(b)

Figure 2

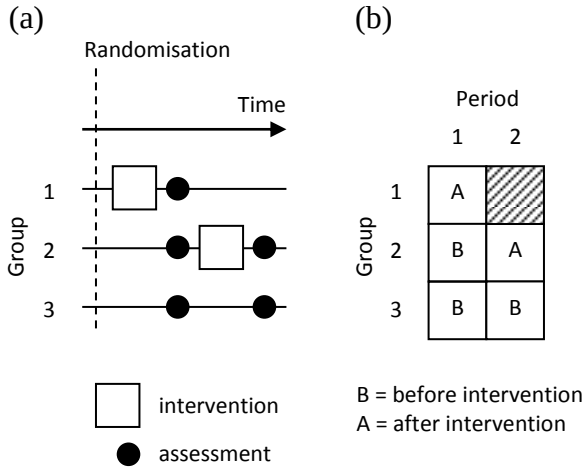
(a)
Randomisation
Time
Group
1
2
3
intervention
assessment
(b)
Period
1
2
Group
1
2
3
A
B
A
B
B
B = before intervention
A = after intervention

Figure 3

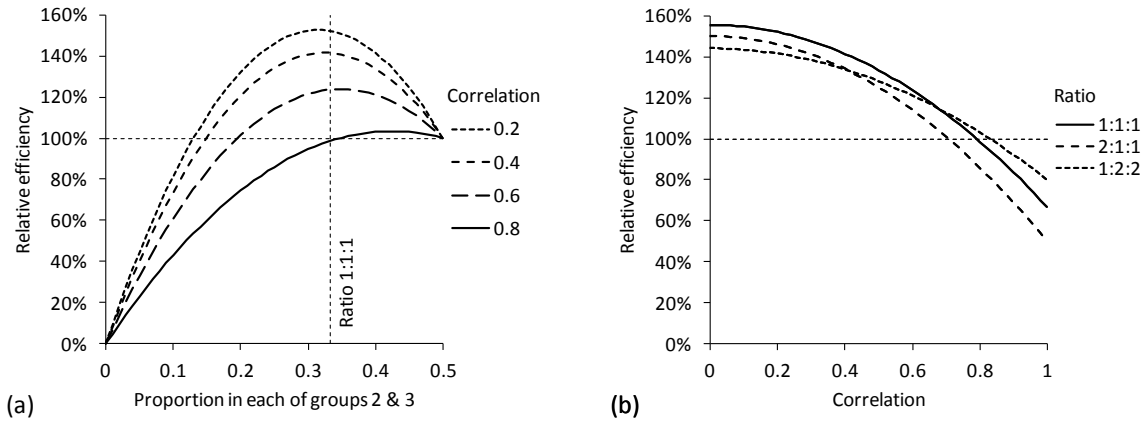
160%
140%
120%
100%
80%
60%
40%
20%
0%
Relative efficiency
0
0.1
0.2
0.3
0.4
0.5
Proportion in each of groups 2 & 3
Ratio 1:1:1
Correlation
0.2
0.4
0.6
0.8
(a)
0
0.2
0.4
0.6
0.8
1
Correlation
Ratio
1:1:1
2:1:1
1:2:2
(b)

Figure 4

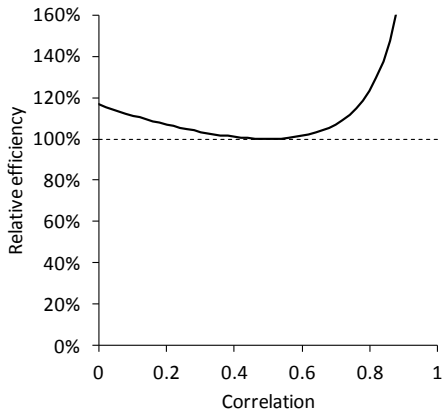
160%
140%
120%
100%
80%
60%
40%
20%
0%
Relative efficiency
0
0.2
0.4
0.6
0.8
1
Correlation

Figure 5

Table 1. Sample size needed in each arm to achieve 80% or 90% power at the 5% significance level, for different assumed effect sizes (treatment difference divided by standard deviation of outcome), using a dog-leg design with 1:1:1 allocation to the three arms.

		Correlation between baseline and follow-up						
Power	Effect size	0·1	0·2	0·3	0·4	0·5	0·6	0·7
80%	0·1	747	708	669	630	590	551	512
	0·2	188	178	169	159	149	139	129
	0·3	85	80	76	72	67	63	58
	0·4	48	46	43	41	39	36	34
	0·5	32	30	28	27	25	24	22
90%	0·1	1000	947	895	842	790	737	685
	0·2	251	238	225	212	199	186	173
	0·3	113	107	101	95	89	84	78
	0·4	64	61	58	54	51	48	44
	0·5	42	40	38	35	33	31	29