BOOTSTRAPPING ECONOMIC DATA IN A CLUSTER RCT

Terry N. Flynn, Elise Whitley, Tim J. Peters

Dept Social Medicine, University of Bristol

1. Introduction and issues in economic analysis

"The cluster randomization trial is one in which intact social units, or clusters of individuals, rather than individuals themselves, are randomized to different intervention groups. Trials randomizing clusters, sometimes called group randomization trials, have become particularly widespread in the evaluation of non therapeutic interventions, including lifestyle modification, educational programmes and innovations in the provision of health care." [1]

The most fundamental issue in any analysis of such data is the potential violation of the independence assumption implicit in most standard statistical models. Subjects within a cluster are usually not independent and the variability in responses from individuals within a cluster is usually less than that from individuals in different clusters. The use of traditional methods of analysis such as ordinary least squares will typically underestimate standard errors leading to spurious significance. Such issues in the analysis of continuous *clinical* data have largely been dealt with. However a particular area of analysis that has received little attention to date is that of the analysis of economic data from cluster RCTs. This is of particular interest because:

1. Many trials randomise only a small number of clusters, leading to potential problems with a bad randomisation and difficulties with making inference on cluster-level covariates and between-cluster variability. Although in theory this problem is as relevant to clinical data, in practice it may not be as severe: with continuous clinical data, methods of analysis based on the *t*-test are fairly robust to moderate violations of the assumptions of normality and homogeneity of variances.[2] Economic data may prove more troublesome because:
2. Highly skewed cost data, at the cluster-level or the individual-level, are potentially very problematic. It is not clear to what extent appeals to normality or even symmetry of data can be made with regard to the distribution of cluster means, given the limited number of clusters in the population (for example, hospitals in the country), the variation in medical practice, and the social and geographical factors that may influence these.

Economic evaluations alongside cluster randomised controlled trials are not yet common. This is understandable, given that the principles of the design and analysis of the cluster RCT itself are not yet widely understood and formalised. Nevertheless an awareness of the likely issues in their design is crucial because:

1. Economic data are increasingly being used by policy-makers to inform priority setting. Cost-effectiveness data from cluster RCTs will be required at some point.
2. Cluster RCTs are often expensive to set up and run. Mistakes are costly to rectify.
3. Adequately powered economic evaluations have often been observed to be larger than their clinical counterparts. In conjunction with point 2 above, this problem is likely to be particularly acute.

Given the strengths of the bootstrap in analysing skewed economic data from individually randomised trials, it is natural to address the issue of how well it performs in clustered data. The relative merits of the bootstrap compared with the simple box method are well known for individually randomised trial data but a comparison of the two methods in clustered data is of interest because:

1. The simplest methods of bootstrapping that can handle clustered data do not have optimal properties in terms of matching second and higher moments. Good performance is not guaranteed.
2. Even such simple bootstrapping methods require programming and are not yet available in standard statistical packages. The box method is easily understood and such confidence intervals are easily estimated.
3. Given the potential for more than one level of correlation in a cluster RCT, there will be combinations of intracluster correlation coefficient (ICC) and correlation for which the box is likely to perform acceptably.

Therefore it is important to establish the conditions under which different methods of analysis of economic data perform best. These conditions can be established through simulation studies that attempt to assess the relative merits of methods of analysis under various plausible trial scenarios where the underlying nature of the data is known. The aims of this work are twofold:

1. To compare confidence interval coverage of cluster-adjusted bootstrap methods with that of a traditional robust method in estimating a mean cost difference.
2. To compare confidence interval coverage of cluster-adjusted bootstrap methods with that of the box method in estimating a cost-effectiveness ratio.

2. The cluster-adjusted bootstrap

The flexibility of the bootstrap has been praised but it has suffered from a serious limitation: bootstrapped data must be independently and identically distributed to ensure that the parameter of interest is identically distributed for each bootstrap replication. When stratification, cluster sampling or probability weights are introduced this assumption is violated, leading to incorrect inference. Work has been carried out in the 1980s and 1990s to generalise the bootstrap to non i.i.d data.

Table 1: Analysis of variance table for random effects model.

ANOVA table for a random effects model				
Source of variation	Sum of squares	Degrees of freedom	Mean square	Expected mean square
Between clusters	$SS_B = \sum_{i=1}^{k} n\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2$	k-1	$\frac{SS_B}{k-1}$	$n\sigma_B^2 + \sigma_W^2$
Within clusters	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_i\right)^2$	N-k	$\frac{SS_W}{k(n-1)}$	σ_W^2
Total	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \overline{\overline{Y}}\right)$	N-1		

Consider table 1, the analysis of variance table for the simplest type of random effects model. The jth individual in the ith cluster (j=1,…,n ; i=1,…,k) has an outcome

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where μ represents the overall mean outcome in the population, α_i are the cluster effects which are identically distributed with mean zero and variance σ_B^2, ε_{ij} are the individual effects which are identically distributed with mean zero and variance σ_W^2 and the group and individual effects are independent. The total number of individuals is $N = \sum_{i=1}^{k} n = kn$, the estimated cluster means are $\overline{Y}_i = \sum_{j=1}^{n} Y_{ij} / n$ and the grand mean is $\overline{\overline{Y}} = \sum_{i=1}^{k} \overline{Y}_i / k$.

All the bootstrap methods that have been proposed in the literature have a common feature, namely the need to bootstrap whole clusters. Davison and Hinkley identified two methods of selecting individuals within-clusters.[3]

Method 1: 'Single bootstrap' selection.

Under this procedure, clusters are bootstrapped and each resampled cluster is kept intact (in other words if a particular cluster is picked then all individuals within that cluster are selected). Under this scenario the expected value of the between-cluster mean square (of each bootstrap sample) is $n\sigma_B^2 + \sigma_W^2$. Therefore when clusters are bootstrapped the sample mean is identically distributed across bootstrap samples and inference can be made about the distribution of the sample mean. The bootstrapped data show that:

$E^*(Y_{ij}^*) = \bar{y}$;

$$E\left\{\text{var}^*(Y_{ij}^*)\right\} = \frac{k-1}{k}\sigma_B^2 + \frac{kn-1}{kn}\sigma_W^2$$

$$E\left\{\text{cov}^*(Y_{ij}^*, Y_{il}^*)\right\} = \frac{k-1}{k}\sigma_B^2 - \frac{1}{kn}\sigma_W^2$$

Thus, there is approximate matching of the second moments:

$\text{var}(Y_{ij}) = \sigma_B^2 + \sigma_W^2$;

$\text{cov}(Y_{ij}, Y_{il}) = \sigma_B^2$, $j \neq l$.

It should be noted that the expected variance and covariance of the resampled outcome data are slightly biased downwards. However an estimator such as the sample mean is strongly consistent (in that its bias is zero and is variance tends to zero as the total sample size approaches infinity); the level of bias is small unless the number of clusters becomes very small. Furthermore, compared to a traditional robust variance estimator, the disadvantage of this bias should be weighed against potential advantages. For instance if a confidence interval is required for the sample mean and there is considerable skewness in the distribution then a bootstrap confidence interval around the mean, such as the BC_a interval, might exhibit better coverage than a symmetric one.

Method 2: 'Double bootstrap'selection.

An alternative method involves resampling individuals as well as resampling whole clusters. The approach uses a first stage bootstrap applied to the estimated cluster means (sampling with replacement). Individuals are then bootstrapped. Under this second stage, the deviations from the estimated cluster means are resampled. Hence, within each bootstrap cluster, a bootstrap sample of the deviations from the estimated mean is performed. This bootstrap sample of the deviations is then added on to the estimated mean.

More formally, define:

$\hat{x}_i = \bar{y}_i$, the estimated cluster means;

$\hat{z}_{ij} = y_{ij} - \bar{y}_i$, the deviations of individual outcomes from the estimated cluster means.

Resampling is performed as follows:

STAGE 1: $x_1^*,...,x_k^*$ are obtained by randomly resampling with replacement from $\hat{x}_1,...,\hat{x}_k$ (the estimated cluster means)

STAGE 2: Randomly resample with replacement from $\hat{z}_{i1},...,\hat{z}_{in}$ (either a randomly selected cluster or the cluster corresponding to x_i^*) to give $z_{i1}^*,...,z_{in}^*$

STAGE 3: Set $y_{ij}^* = x_i^* + z_{ij}^*$, $i = 1,.....,k; j = 1,.....,n.$

However there is a problem with this approach. The estimated cluster means incorporate both within and between-cluster variability (see table 1). By adding the deviations from these estimated cluster means back in we have, in effect, double counted the within-cluster variance and the expected variance of the parameter being estimated, e.g. the sample mean, will exceed that from the original sample of data. Davison and Hinkley propose shrinking the cluster means to eliminate this problem, as follows. The shrinkage estimates, $\bar{Y}_i'$, are calculated :

$\bar{Y}_i' = c\bar{\bar{Y}} + (1-c)\bar{Y}_i$ where c is given by

$(1-c)^2 = \frac{k}{k-1} - \frac{SS_W}{(n-1)SS_B}$; if the right hand side is negative, it is reset to zero.

The variance of the adjusted cluster means, $\bar{Y}_i'$, is then $\frac{k}{k-1}\sigma_B^2$. Unless the cluster size is large, the deviations from the estimated cluster means should also be standardised to

$\hat{y}_{ij} = \frac{y_{ij} - \bar{y}_i}{\sqrt{(1-1/n)}}$.

3. Methods (cost data)

The first aim was to compare the performance of standard (cluster-adjusted) confidence intervals with those obtained from the above cluster-adjusted bootstrap procedures when estimating the difference in mean costs between two treatment groups. ‘Standard’ has been taken to mean a confidence interval quoted for continuous data in packages such as Stata. In other words, a

symmetric confidence interval utilising a Huber-White (robust) cluster-adjusted standard error. The comparisons of performance were in terms of coverage of the various confidence intervals. For all parameter sets a nominal 95% confidence interval was estimated. Miscoverage was split according to whether the true cost difference was below or above the estimated confidence interval. The conditions captured here certain characteristics that may be common to many cluster RCTs and which, if present, could have implications for the validity of traditional methods of analysis:

1. A relatively small number of clusters
2. Skewed distributions at both the individual and the cluster-level
3. Differences in variances, and hence differences in ICCs, across the two treatment groups.

Conceptualising costs in a cluster RCT

In attempting to construct realistic scenarios for cost data from a cluster RCT a number of factors were considered:

1. What is the nature of the distribution of individual patient costs expected to be in the population of patients normally eligible for treatment?
2. How representative of this population distribution are the cost distributions within clusters likely to be? This has implications for the ICC and the assumptions regarding the distributions.
3. How might the introduction of the intervention affect 2. above?

The nature of the distribution of costs in the population

When conceptualising costs in a clustered framework, similar concerns to those usually outlined must be considered (for example the setting, the possibility of very high unit costs for some patients). Community intervention trials might not exhibit such characteristics because of, for example, the rarity of high cost additional interventions such as a stay in intensive care, or the impracticality of collecting individual-level data at all (thus, 'hiding' the skewness). However, this assumption of a skewed distribution of the costs at the population level might be reasonable if we are considering hospital-based pragmatic RCTs when interventions would not be administered to a homogenous group of patients. In other words, in reality there may always be a group of patients at the extreme bounds of entry criteria, who, for certain medical conditions, are very costly to treat due to co-morbidities or complications.

Implications for the ICC and assumed distributions

Suppose the distribution of costs for a population of patients is expected to be highly positively skewed. Suppose further that clusters of individuals are to be sampled rather than individuals themselves. Depending upon the nature of the clustering this skewness at the population level may arise from two sources:

1. Every possible cluster that could be selected from the population is equally representative of the population. This may come about if all clusters contain a similar proportion of high-cost patients and unit costs are similar across all clusters. The between-cluster variance is likely to be small and the distribution of the cluster means will contribute little to the overall distribution at the population level. The individual costs (within clusters) are highly skewed and it is these that are driving the overall distribution.

2. Most of the skewness in the population of individuals might be attributable to large heterogeneity in cluster mean costs. In other words, for a positively skewed distribution, whilst many clusters are grouped towards the left-hand (short) tail of the distribution, a few clusters exhibit very high costs on average. This could arise for several reasons including the inclusion of teaching hospitals or London hospitals (hence high unit costs), geographical variations in the utilisation of high cost interventions in some clusters or variations in medical practice (e.g. greater use of intensive care units). Assuming a positively skewed distribution for the population of individuals, the effect of such factors is likely to be a relatively high value for the cost ICC. However, the positive skewness at the population level is unlikely to be driven purely by the distribution of cluster means, but is likely to be attributable, at least in part, to skewness in the individual-level data.

Distributional and ICC differences between treatment arms as a result of the intervention

The factors identified above may have differential effects in the two intervention arms. Whether or not there is a non-zero treatment effect, the introduction of treatment may act to standardise practice among health professionals. The resulting direction of movement in the ICC may not necessarily be predictable *a priori* since one or both of the two variance components may change. One possible result is that greater standardisation of treatment leads to homogeneity in success rates among clusters that would otherwise have performed poorly. Thus, treatment may not only reduce the ICC but may pull in the left tail of the distribution leading to a more positively skewed distribution. Another possibility is that the intervention is a new one resulting in differential adherence to treatment protocols among clusters in the intervention arm. Under this scenario success rates in the

intervention arm might be much less homogenous than those in the control arm. Better adherence to treatment protocols among some clusters would lead to a more positively skewed distribution coupled with a larger ICC in the intervention arm.

Parameters and their values

Two potential extreme scenarios have been described. Under the first, all clusters were equally representative of the population, leading to a high degree of skewness within most, if not all, clusters. Under such a scenario, the ICC is likely to be small and results would be expected to be robust to any incorrect assumptions made regarding the between-cluster distribution. Under the second scenario, within-cluster costs are expected to be more homogenous and much of the skewness in the cost data at the population level is attributable to differences in the cluster mean costs. As a result, the ICC would be expected to be much larger, and the between-cluster distribution might be expected to exhibit considerable skewness. This section sets out the parameters which, when varied, attempted to capture the two extreme scenarios above as well as situations in between. Table 2 shows the 15 parameter combinations investigated for each value of the control arm ICC. The ICC is defined as the proportion of total variation accounted for by between-cluster variation and the variance components for ICC=0.1 (10/(10+90)) are shown as an example. These are discussed in more detail below.

Table 2: 15 Parameter combinations run for control ICC=0.1

Control		**How Does the ICC Change as a Result of the Intervention?**		**Intervention**		**Between-cluster Distribution Combination**
σ_B^2	σ_W^2			σ_B^2	σ_W^2	
10	90	Stay Same	-	10	90	N,N LN,LN N,LN
		Double	$\sigma_W^2 \downarrow$	10	40	N,N LN,LN N,LN
			$\sigma_B^2 \uparrow$	22.5	90	N,N LN,LN N,LN
		Halve	$\sigma_W^2 \uparrow$	10	190	N,N LN,LN N,LN
			$\sigma_B^2 \downarrow$	4.74	90	N,N LN,LN N,LN

The between-cluster and within-cluster distributions

There were two distributional assumptions used for the cluster means. The first was the normal distribution. The second required a positively skewed distribution. The lognormal distribution was chosen to be consistent with previous work.[4, 5] For each of the ICC combinations given below, there were three possible combinations of the between-cluster distributions:

1. The cluster means in both groups were normally distributed (N,N).
2. The cluster means in both groups were lognormally distributed (LN,LN).
3. The cluster means in the control group were normally distributed, whilst in the intervention group they were lognormally distributed (N,LN).

The same distribution was used for individual-level data in both treatment arms. Lognormal data were used in order to provide a scenario applicable to individual-level cost data.

The ICCs

Given that the likely range of values for a cost ICC is largely unknown, values between zero and 0.5, a figure that is high (and, in clinical data rarely observed) were used. The ICCs used were 0.01, 0.1 and 0.25 for the control group. For a given value of the ICC in the control group, the intervention ICC can do one of three things. It can remain the same, it can increase or decrease. Moreover, since the ICC consists of two variance components, it can change in one of two ways. Thus the intervention ICC was specified in one of five ways :

1. Remained the same as the control ICC,
2. Doubled, as a result of the appropriate increase in the between-cluster variance,
3. Doubled, as a result of the appropriate decrease in the within-cluster variance,
4. Halved, as a result of the appropriate decrease in the between-cluster variance,
5. Halved, as a result of the appropriate increase in the within-cluster variance.

In justifying why, for instance, a comparison of 2 and 3 is necessary, consider the case in which there are normal distributions at the cluster-level but lognormal distributions at the individual-level. Any increase in the variance of a lognormal distribution increases the kurtosis and skewness (and vice-versa). This is not so for a normal distribution. Hence, changes in σ_B^2 and σ_W^2 might be expected to have different effects upon estimated confidence intervals. Similarly, in justifying why

doubling and halving of ICCs is needed (2 and 3 versus 4 and 5), consider the case in which the between-cluster distribution is normal in the control group and lognormal in the intervention group. This asymmetry in distributions means one cannot simply reverse intervention and control labels.

The number of clusters in each group and number of individuals in each cluster

The number of potential trial sizes was kept relatively small in order to economise on computing time. The number of clusters in each group was 6, 12 or 24. These figures reflect the small numbers of clusters recruited in many cluster RCTs and, coupled with cluster sizes of 25,50 or 100, they allowed alternative combinations of cluster size and number of clusters to be investigated for a given total trial size. Mean cluster sizes between 50 and 100 are not unusual for trials conducted for cardiovascular variables (for example) at a UK District Health Authority level but there is enormous variation in cluster size, depending upon treatment area and type of cluster.[6] Thus, given three possible ICCs in the control group, this made 45 possible combinations in total. These 45 parameter combinations were used for each combination of cluster size and number of clusters. Table 3 shows the trial size combinations run.

Table 3:Trial size combinations run

		Cluster Size		
		25	**50**	**100**
Number of clusters per arm	**6**	X	X	X
	12	X	X	
	24	X		

Model generating process

The data were constructed by assuming there were *n* individuals in each of *2k* clusters. Half of these were randomised to a hypothesised intervention group, whilst the other half were randomised to a control group. A random effects model incorporating a treatment dummy variable was used:

$$C_{hij} = \alpha_{hi} + \beta T_h + \varepsilon_{hij} \qquad h = 0,1; i = 1,....k; j = 1,...n;$$
$$E(\alpha_{hi}) = 0; E(\varepsilon_{ij}) = 0$$
$$Var(\alpha_{hi}) = \sigma_B^2; Var(\varepsilon_{ij}) = \sigma_W^2$$

Thus individual *j* in cluster *i* received treatment *h*. Each individual's outcome, C_{hij}, comprised three elements: the effect of treatment, βT_h, the cluster-specific effect, α_{hi}, and the individual-specific effect, ε_{hij}. The effect of treatment was set at zero, cluster-level data were all lognormally distributed whilst the total variance, $\sigma_B^2 + \sigma_W^2$, was arbitrarily fixed at 100.

Data analysis

For the set of data generated for each simulation, three methods of analysis were performed: the robust method of confidence interval estimation and two methods of bootstrap confidence interval estimation. For the robust method a Huber-White standard error of the data was multiplied by a *t*-deviate and added to/subtracted from the estimated treatment effect to estimate the upper/lower confidence limits**.** For the two bootstrap methods, bootstrapping was performed independently within each treatment group and the statistic of interest became the difference in mean outcome between the two treatment arms. Under method 1 all individuals within a resampled cluster were then selected. Under method 2 a second level of bootstrap was performed on individuals within-clusters selected at level one. The difference between the two group means was then calculated. This was repeated to give B bootstrap estimates of the treatment effect. A BC_a confidence interval was then estimated at the same nominal percentage level (e.g. 95%) as that under the robust method. Given the nature of the BC_a method, the resulting confidence interval need not be symmetric.

4. Results (cost data)

Table 4 – Table 6 below show coverage rates for the three methods of analysis for each sample size combination when averaged over the 15 parameter combinations for a given control group ICC. Within each box the first number represents the coverage of the robust confidence interval, the second represents the coverage of the BS1 method whilst the third figure represents that of the BS2 method. In this format it permits easy comparison of the relative abilities of the methods of analysis to cope with larger degrees of clustering. Furthermore for a given ICC and given total sample size, combinations of cluster size and number of clusters per arm can be compared for each method of analysis. This is facilitated by reading entries along the diagonal lines in each table.

Table 4: Observed coverage (%) for robust, BS1 and BS2 methods of analysis: ICC=0.01

		Cluster Size		
		25	**50**	**100**
Number of clusters per arm	**6**	94.1	93.8	94.0
		88.2	88.2	88.3
		93.2	93.1	92.7
	12	94.6	94.5	
		91.1	91.3	
		93.4	93.3	
	24	94.8		
		92.7		
		93.7		

Table 5: Observed coverage (%) for robust, BS1 and BS2 methods of analysis: ICC=0.1

		Cluster Size		
		25	**50**	**100**
Number of clusters per arm	**6**	93.9	93.9	93.8
		87.6	87.3	86.9
		90.9	90.6	90.3
	12	94.5	94.4	
		90.5	90.2	
		91.8	91.6	
	24	94.8		
		92.1		
		92.6		

Table 6: Observed coverage (%) for robust, BS1 and BS2 methods of analysis: ICC=0.25

		Cluster Size		
		25	**50**	**100**
Number of clusters per arm	**6**	93.6	93.5	93.5
		86.9	86.6	86.4
		90.1	89.9	89.6
	12	94.3	94.2	
		89.9	89.6	
		91.3	91.1	
	24	94.6		
		91.5		
		92.1		

A number of results are immediately apparent from the above three tables:

- All three methods produce coverage of less than 95%, the nominal level, but appear to be consistent with respect to the impact of the number of clusters per arm. In other words, as the number of clusters per arm increases, the observed coverage approaches 95% for each method.
- The coverage of the robust method is always close to 95%, the BS2 is next best and the BS1 method is always outperformed by the other two methods.
- As the ICC in the control group increases (that is, across tables) the robust method performs slightly worse whilst the performance of the bootstrap methods is noticeably poorer.
- When examining numbers along the diagonal in each figure, the bootstrap methods perform much better for a large number of clusters and small cluster size rather than vice-versa. This is probably due to the slight downward bias in the second moments: the degree of bias is an inverse function of the number of clusters.

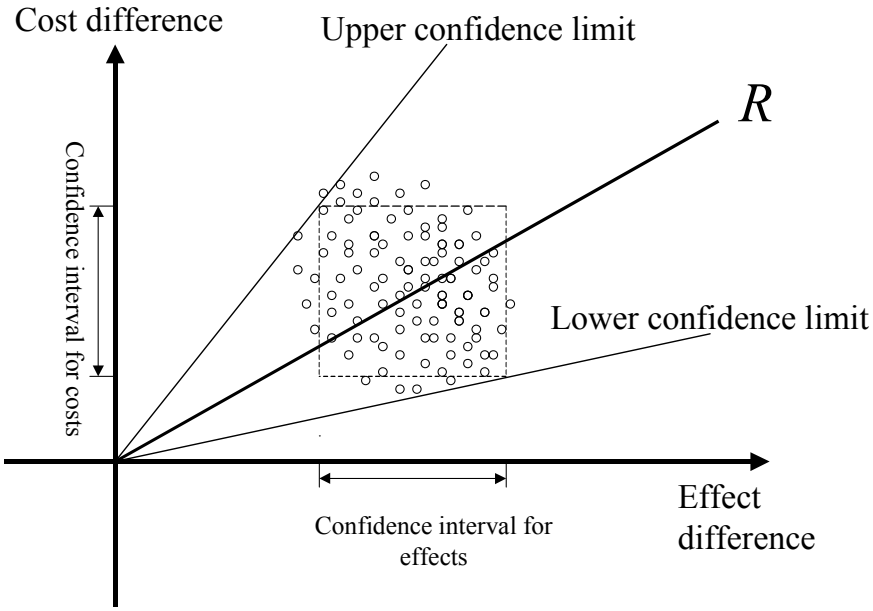
From the results in these tables there does not appear to be much to commend the bootstrap. Both bootstrap methods are always outperformed by the robust method. Although there are particular parameter combinations for which the bootstrap performs acceptably, in general, one would prefer the robust method for the parameter combinations here.

5. Methods (Cost-effectiveness data)

The aim here was to compare estimated confidence intervals for an incremental cost-effectiveness ratio estimated from a cluster RCT. This was of interest because costs and effects can be correlated at a cluster-level or at the individual-level or both. The degree of clustering and the relative strengths of these two levels of correlation may have implications for the coverage properties of confidence intervals based on those from costs and effects separately (the box method). The box method is often outperformed by methods such as the bootstrap. However the examination here is of interest because standard packages do not perform clustered bootstrapping correctly and the conditions in which such simple methods perform acceptably should be emphasised.

The figure below shows the 95% confidence intervals for costs and effects separately which are used to construct a box were in the cost-effectiveness plane. Such a method avoids the issues

concerning the statistical behaviour of ratios, but gives a confidence region that usually doesn't correspond to a genuine 95% confidence interval for the cost-effectiveness ratio since it does not take account of any correlation between costs and effects.



When the data are clustered there are two potential levels of correlation. Suppose, for example, a high ICC were to be accompanied by a zero correlation between the costs and effects at the cluster-level. This might be the case if, for example, costs of treatment were heavily influenced by regional factors rather than the treatment's effectiveness. Then, for example, a positive correlation between costs and effects within clusters might be offset with the result that the box method might give a confidence interval whose coverage was approximately right.

Data generation (cost-effectiveness ratios)

The structure of the hypothesised RCT is similar to that under the single outcome model. The main change is that there are two outcomes of interest for every individual, the cost and the effect of treatment. Effects (E) follow the same random effects model as that described above, but the distributions at the cluster and the individual-level are normal. Costs (C) follow a random intercept model (lognormal distributions) with additional restrictions:

$$\text{corr}\,(\alpha_{(E)hi}, \alpha_{(C)hi}) \approx \rho_i$$

$$\text{corr}\,(u_{hij}, \varepsilon_{hij}) \approx \rho_j$$

In the current model, it should be noted that there are two potential levels of correlation, that of costs and effects for an individual and that of costs and effects for a cluster. For instance, effective general practices could exhibit high costs on average, whilst for any given practice the cost of treating a patient may be unrelated to outcome. 5 possible correlations (-1,-0.5,0,0.5,1) at each of the two levels of variation gives 25 (5x5) combinations. Not all such combinations are sensible, however, and the results tables do not include entries for combinations such as a correlation of –1 at one level and 1 at the other. The mean and standard deviation of the cost and effect differences were set such that the coefficient of variation was fairly large (0.5 for costs and 0.25 for effects) as recommended by Briggs *et al*.[4]

Given time restraints, the need for additional parameter sets incorporating different correlations between costs and effects meant that economies had to be made with regard to distributional and variance combinations. Thus, for these simulation runs, no change in the ICC as a result of treatment was envisaged, nor was any change in the distribution as a result of treatment.

Data analysis (cost-effectiveness ratios)

The coverage of two methods have been investigated.

1. Under the first method separate confidence intervals for costs and effects have been estimated using the methods described in the cost section. A confidence box was used to estimate a confidence interval for the cost-effectiveness ratio.
2. Under the second method the double bootstrap was used to estimate a confidence interval for the cost-effectiveness ratio.

6. Results (Cost-effectiveness data)

Table 7 shows the mean confidence interval for the box method for various plausible combinations of correlation (at cluster and individual-level) for three values of the ICC, for a trial randomising six clusters of size 25 per treatment arm. Given time constraints, unlike the costs scenarios of sections 3 and 4, these are not averaged over various parameter sets involving changes in variance components and between-cluster distributions.

Thus each figure in the table represents the mean coverage for 5000 simulations where:

- The ICC was the same in both treatment arms (and the same for both costs and effects)
- The distribution of effects was normal at cluster and individual-level
- The distribution of costs was lognormal at cluster and individual-level

Note that the correlations in the tables are not strictly those between the normally distributed effect data and lognormally distributed cost data. Costs were initially generated from a normal distribution and the correlation between costs and effects was built in at this point. Exponentiating and transforming the cost data to make them lognormally distributed has the effect of shrinking the actual correlation towards zero. Thus the correlations of 1 and –1 are, in practice, illustrating correlations of around 0.7 and –0.7.

Table 7: Observed coverage (%) for box method of analysis

			Individual-level Correlation				
			-1	**-0.5**	**0**	**0.5**	**1**
ICC=0.01	**Cluster-level Correlation**	**-1**	92.2		95.0		
		-0.5		93.7	94.2		
		0	93.6	94.2	95.6	95.9	95.5
		0.5			96.0	96.3	
		1			95.6		96.3
ICC=0.1	**Cluster-level Correlation**	**-1**	92.4		93.8		
		-0.5		93.9	94.2		
		0	94.4	94.9	95.5	95.9	95.9
		0.5			96.0	96.3	
		1			95.7		96.2
ICC=0.25	**Cluster-level Correlation**	**-1**	92.3		93.4		
		-0.5		94.6	94.7		
		0	95.6	95.3	96.1	95.4	96.0
		0.5			96.1	96.0	
		1			96.0		96.3

Several points are apparent from these results:

- As expected, the smaller the ICC, the less effect the cluster correlation has upon coverage. In other words, for a small ICC and a given individual-level correlation, moving vertically through a table has less effect upon coverage compared to moving vertically when the ICC is large.

- As expected, the larger the ICC, the less effect the individual-level correlation has upon coverage. In other words, it is only when the ICC is small that coverage changes much when moving horizontally within a table .
- For a small ICC, 0.01, cluster-level correlation is only materially relevant (arbitrarily defined as ability to change coverage by 1 percentage point across the range of values for correlation) when the individual-level correlation is high anyway. Under such circumstances observed coverage moves even higher than 95%.
- For a large ICC, 0.25, individual-level correlation is only materially relevant when the cluster-level correlation is very small (very negative).
- For an intermediate ICC, 0.1, not unexpectedly, both correlations should not be too extreme (close to 1 or –1) if coverage is to be fairly close to 95% (in this case within 1.5% of the nominal level).

Table 8: Observed coverage (%) for BS2 method of analysis

			Individual-level Correlation				
			-1	**-0.5**	**0**	**0.5**	**1**
ICC=0.01	**Cluster-level Correlation**	**-1**	95.8		94.1		
		-0.5		95.5	94.1		
		0	96.0	96.1	95.3	94.1	91.1
		0.5			95.6	94.9	
		1			96.2		94.0
ICC=0.1	**Cluster-level Correlation**	**-1**	92.0		90.9		
		-0.5		93.6	91.1		
		0	94.3	93.6	92.6	91.5	87.6
		0.5			94.3	92.7	
		1			96.3		93.1
ICC=0.25	**Cluster-level Correlation**	**-1**	91.3		90.3		
		-0.5		91.7	90.3		
		0	93.6	92.5	92.5	90.4	90.1
		0.5			92.9	92.3	
		1			95.6		92.5

Table 8 shows similar results but for the double bootstrap method. As for the cost data, the downward bias in the second moments is clearly apparent in the results, particularly for larger ICCs,

leading to excessively narrow confidence intervals. The shaded entries indicate where the BS2 method was outperformed by the box method; in other words the box method achieved coverage that was closer to the nominal level of 95%.

Points to note include:

- Unlike the box method, increases in the individual-level correlation lead to lower rather than higher coverage. This is again probably due to the downward bias in the second moments: greater correlation leads to greater grouping of the cost-effectiveness estimates around the true cost-effectiveness ratio, which the bootstrap confidence interval takes account of. If the bootstrap confidence interval is too narrow anyway, Monte Carlo variation is then more likely to lead to non-inclusion of the true cost-effectiveness ratio. ICCs of closer to one would probably cause the results to exhibit a similar phenomenon for increases in the cluster-level correlation.
- For ICC=0.01 The coverage of the bootstrap and the box methods are similar. There is no consistent pattern in terms of relative coverage.
- For ICC=0.01 and 0.25 the situation is clearer and is reasonably consistent across the two ICCs. The bootstrap is nearly always outperformed (in terms of better matching of the 95% nominal level) by the box method. There is only a single combination for each ICC where the bootstrap gives a better coverage.
- Although the coverage of the bootstrap is clearly fairly good, at first glance it would appear that the investigator could do far worse than use the box method. However, the miscoverage rates (not tabulated), split according to whether the estimated confidence interval was too high or too low, suggest that the bootstrap is far superior in terms of taking account of the skewed data. In other words the mean rejection rate ‘above’ and ‘below’ the true cost-effectiveness ratio were similar, and usually much closer to the nominal 2.5% than those of the box method. Those of the box method were mostly on one side of the ratio.

7. Conclusions and further work

The results from the analysis of cost data seem to suggest that, for the parameter combinations considered here, the disadvantages of the cluster-adjusted bootstrap outweigh its advantages in dealing with skewed data. However, several points should be borne in mind:

1. Cost data from cluster RCTs are not yet widely available. If the conditions underlying the analysis here are insufficiently extreme, simple cluster-adjusted bootstrap methods such as those presented here may yet prove valuable.
2. More complex bootstrap methods have been developed and are being included in specialist packages for hierarchical data. These methods have the ability to deal with non-constant cluster size and are better able to match the higher moments of the data. The performance of such methods is therefore likely to be better than that of the simple methods described here.

The preliminary results from the cost-effectiveness simulations appear much more encouraging for the cluster-adjusted bootstrap. Even such 'theoretically incorrect' methods such as the double bootstrap method presented here perform quite well. However such a method still requires programming and until this, and more complex methods become accessible to non-specialists, the results of the work here can give recommendations as to when simple methods, such as the box method, are likely to give satisfactory results. In particular, the box method performs relatively well when the ICC is moderate to large and its coverage is never very poor. However, the inability of the box method to take account of heavily skewed data should act as a caution to researchers. Further work will ascertain whether these results here hold true when the ICC changes as a result of treatment, either through a change in the between-cluster variance or the within-cluster variance. Larger number of clusters will also be investigated to ascertain whether the performance of the double bootstrap can be further improved upon. These results suggest that the complex bootstrap methods that have begun to be utilised in hierarchical data structures are likely to prove useful.

Bibliography

1. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000;1-173.

2. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* 1996;**49**:435-439.

3. Davison AC, Hinkley DV. *Bootstrap methods and their applications*. Cambridge: Cambridge University Press, 1997.

4. Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat.Med.* 1999;**18**:3245-3262.

5. Polsky D, Glick HA, Willke RJ, Schulman K. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Economics* 1997;**6**:243-252.

6. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999;**3**.