International Journal of Technology Assessment in Health Care, **21:3** (2005), 403–409.

Cluster randomized trials: Another problem for cost-effectiveness ratios

Terry N. Flynn, Tim J. Peters
University of Bristol

Objectives: This work has investigated under what conditions cost-effectiveness data from a cluster randomized trial (CRT) are suitable for analysis using a cluster-adjusted nonparametric bootstrap. The bootstrap's main advantages are in dealing with skewed data and its ability to take correlations between costs and effects into account. However, there are known theoretical problems with a commonly used cluster bootstrap procedure, and the practical implications of these require investigation.
Methods: Simulations were used to estimate the coverage of confidence intervals around incremental cost-effectiveness ratios from CRTs using two bootstrap methods.
Results: The bootstrap gave excessively narrow confidence intervals, but there was evidence to suggest that, when the number of clusters per treatment arm exceeded 24, it might give acceptable results. The method that resampled individuals as well as clusters did not perform well when cost and effectiveness data were correlated.
Conclusions: If economic data from such trials are to be analyzed adequately, then there is a need for further investigations of more complex bootstrap procedures. Similarly, further research is required on methods such as the net benefit approach.

Keywords: Randomized controlled trials, Clustered data, Health services research, Economics

Cluster randomized trials (CRTs) are experimental investigations in which groups of individuals (clusters) are randomized rather than separate individuals (8). The relative complexity of CRTs has led to much methodological work concerning their design and analysis (19), but the analysis of cost-effectiveness data from these trials has received little attention. The conceptual issues arising in this context have been explored (12); briefly, there are four problems, the first two of which are already well-known in the context of individually randomized trials.

First, if the analysis is in terms of an incremental cost-effectiveness ratio (ICER), traditional statistical analysis is problematic. One difficulty is that the variance of a ratio of two stochastic variables cannot, generally, be calculated exactly (6). Asymptotically, the ratio of two standard normal variables follows a Cauchy distribution. Unfortunately, the theoretical mean of a Cauchy distribution does not exist, and the sample mean is unstable—the average ICER over any number of observations exhibits the same degree of variation as one single observation (20). In any case, in practice, the cost and effect differences in the numerator and denominator of the ratio are unlikely to be standard normal; therefore, the resulting sampling distribution is unknown (2). Hence, both determining sample size calculations and using parametric analytical methods have proved difficult for cost-effectiveness studies. The second well-known problem concerns negative cost-effectiveness ratios. The ICER can be negative from a negative numerator or denominator, but the implications of these two scenarios are very different. Inferences must be drawn carefully then, if an ICER confidence interval crosses zero.

Clustered data present two additional problems. First, the data requirements for a properly powered or analyzed cost-effectiveness study are even more onerous than an individually randomized trial. Previous work has illustrated the need for prior data on intracluster correlation coefficients

The work presented here formed part of T.N.F.'s PhD thesis, funded by the UK Medical Research Council (MRC Studentship Ref G78/5775). T.N.F. is funded currently by the MRC Health Services Research Collaboration, for which Bristol is the lead center. We acknowledge the assistance and helpful comments of Dr. Paul Mulheran, Dr. Elise Whitley, and Professor Allan Donner.

(ICCs) and on correlations between costs and effects at the cluster and individual level (12). Second, the use of the nonparametric bootstrap (3;9) introduces new problems in the analysis of clustered data. Specifically, to ensure that the bootstrap estimate (of, say, a mean) is valid for each replication, bootstrapped data must be independently and identically distributed. When stratification, cluster sampling, or probability weights are introduced, this assumption is violated. Work has been carried out in the 1980s and 1990s to generalize the bootstrap to survey sampling and regression analysis (7;17). For example the Stata statistical package bootstraps whole clusters (7;18), but then the second moments of the bootstrap estimates are biased downward (although consistent with respect to the number of clusters). The resulting confidence intervals are likely to be too narrow, particularly when the number of clusters takes a value typical of many CRTs (7).

The aim, therefore, was to compare the coverage of confidence intervals estimated from two cluster-adjusted bootstrap procedures for an ICER with that expected from theoretically correct confidence intervals (at a nominal 95 percent level). The simulated data used for these purposes encapsulated common features of CRTs with potential implications for the validity of the methods of analysis. Specifically, the following were incorporated: relatively few clusters; skewed cost distributions at both individual and cluster levels; different variances and, hence, ICCs between treatment groups; various correlations between costs and effects.

METHODS

Conceptualizing Cost-Effectiveness Data from a CRT

Nature of Effect and Cost Data. The interdependence of costs and effects necessitated a degree of joint consideration of these factors. However, because costs were conceptualized as being dependent upon effects, the latter were considered independently. Regarding these, effects at both the individual and cluster levels were assumed to be natural clinical units (such as blood pressure) that allowed appeals to the central limit theorem and so were generated from normal distributions. Whereas alternative distributional assumptions for effect data await future work, non-normality was introduced for the cost data. Specifically, three issues were considered. First, the expected distribution of individual patient costs among those normally eligible for treatment. Second, the extent to which cost distributions within clusters are representative of this population distribution has implications for the ICC and the distributional assumptions. Third, this might be influenced by the introduction of an intervention. For instance, as detailed previously, the existence of one unrepresentative cluster (such as a London teaching hospital) in one treatment group may affect the ICC, independently of treatment, or the treatment could directly change the ICC and the distribution (12).

Correlations between Costs and Effects. In a clustered setting, there are two potential levels of correlation—between costs and effects for an individual and between those for a cluster. For example, effective health-care organizations could exhibit high average costs on average, whereas within an organization the cost of treating an individual patient may be unrelated to their outcome. Although such ecological fallacies are possible, in practice, high positive correlations at one level would be unlikely to be observed with high negative correlations at the other level. The following scenarios, therefore, were considered to be most realistic: (i) zero cluster level correlation (where, for instance, costs are determined by purely geographical factors unrelated to clinical effects), but non-zero correlation at the individual level; (ii) vice versa (no correlation at an individual level but, for instance, where referral centers might have higher costs but lower success rates due to case mix); and (iii) the two correlations are approximately the same.

Data Generation Process

Data were constructed assuming n individuals in each of $2k$ clusters. Of these, nk individuals were randomized to an intervention and nk to a control/alternative intervention.

Effects. Effects (E) followed a random effects model utilized in previous research (13) with an additional grand mean that was arbitrarily set to 100. The effect of treatment was set such that the coefficient of variation was 0.25, to be realistic and to reduce the probability that the estimated ICER was distributed symmetrically (1).

Costs. Costs (C) followed a similar random effects model with additional restrictions: in particular, a cluster-level correlation factor and an individual-level correlation factor were introduced. The grand mean cost was arbitrarily set to 1,000, and the cost of treatment was set such that the coefficient of variation was 0.5 in value—again to be realistic and reduce the probability that the estimated ICER was distributed symmetrically (1).

Parameters Varied in the Simulation Model

As noted, normal distributions were utilized for all effect data. Individual level cost data were lognormally distributed. So as to ensure the generation of ICERs that were not too symmetrical (which would defeat the object of testing the methods' relative merits for skewed data), lognormal distributions for the between-cluster cost distributions in both treatment arms were used (11). The following six factors were then varied: the control group ICC, intervention ICC, number of clusters, cluster size, cost, and effect correlations at cluster and individual levels.

ICCs. With the total variance (between-cluster plus the within-cluster variance) arbitrarily fixed at 100, the "control" group ICCs used were 0.01 and 0.1. For each of these values, the intervention ICC was (i) the same; (ii) doubled, as a

result of an appropriate increase in the between-cluster variance; (iii) doubled, by decreasing the within-cluster variance; (iv) halved, by decreasing the between-cluster variance; or (v) halved, by increasing the within-cluster variance.

Sample Sizes. The number of clusters in each group (k) was six or twelve, reflecting the numbers of clusters recruited in many CRTs. The cluster size (n) was twenty-five or fifty; coupled with the cluster sizes, these values allowed alternative configurations to be investigated for a given total trial size.

Correlations between Costs and Effects. The correlations for each level of variation were set at −1, −0.5, 0, 0.5, and 1. Effects and costs were initially generated from normal distributions, and it was at this point that these correlations were incorporated. The cost data required transformation (exponentiation) to produce lognormal distributions, which had the effect of shrinking the actual correlations toward zero. Thus, the above five were ex ante (planned) correlations, whereas the ex post (resulting) correlations between lognormally distributed cost data and normally distributed effect data were closer to zero. Whereas five correlations at each of two levels gave twenty-five possible correlation combinations, given the conceptual points made earlier, only thirteen combinations were actually used in the simulations (Table 1).

Data Analysis

For the data set generated for each simulation, two methods of bootstrap confidence interval estimation were performed. Bootstrapping was performed independently within each of the two treatment groups, with the ICER being the statistic of interest. The sampling structure was maintained in a bootstrap replication by selecting k clusters with replacement from each treatment group. For each of the two procedures, 2,000 bootstrap estimates of the ICER were performed. A bias-corrected and accelerated (BC_a) confidence interval was then estimated at the same nominal percentage level (95 percent). Given the nature of the BC_a method, the resulting confidence interval need not be symmetric (10).

Bootstrap Method 1. Under the first procedure (a "single bootstrap" denoted here by BS1), only clusters were bootstrapped and each resampled cluster kept intact, as utilized by Stata (18) when the cluster () option is added to the bootstrap command. It can be shown that, for the bootstrap estimates, the expected variance and covariance of the resampled outcome data are slightly biased downward (7). 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 bias, therefore, is small, unless the number of clusters is low.

Table 1. Combinations of (ex ante) Correlations between Costs and Effects Utilized

		Individual level correlation				
		−1	−0.5	0	0.5	1
Cluster level correlation	−1	X		X		
	−0.5		X	X		
	0	X	X	X	X	X
	0.5			X	X	
	1			X		X

Bootstrap Method 2. An alternative method (the "double bootstrap" denoted here by BS2) involved resampling individuals as well as resampling whole clusters. This strategy used a first stage bootstrap applied to the estimated cluster means (sampling with replacement), and a second stage in which individuals were bootstrapped, involving resampling the deviations from the estimated cluster means. However, the estimated cluster means incorporate both within- and between-cluster variability and any analysis restricted to the cluster means would overestimate the variance in these means (7). Because incorporating the deviations from the estimated cluster means would effectively double-count the within-cluster variance, the cluster means were shrunk using Davison and Hinkley's shrinkage estimates.

Treatment of Negative ICERs and Confidence Limits

Problems accrued whenever an estimated ICER or confidence limit fell in either the top left or bottom right quadrant of the cost-effectiveness plane. When the intervention was, on average, more costly but less effective than the control, the latter was dominant—any bootstrap estimates in the top left-hand quadrant should not appear on the lower end of the ranking (that is, the lower centiles) of the bootstrap estimates (2).

Conversely, when the difference in mean effect was positive but the difference in mean costs was negative, intervention dominated control. Ranking was nonsensical—greater cost savings or greater incremental effectiveness is desired but each would move the ICER in different directions (the former making it more negative, the latter making it less negative). A conclusion of (unquantifiable) dominance was the only possible inference. Indeed, the coefficients of variation used for effects and, particularly, for costs meant that at least some negative differences were expected, leading to undefined lower confidence limits and perhaps estimates.

Performance Measures from the Simulations

Consequently, the comparisons were in terms of coverage of the various confidence intervals—that is, the percentage of simulations for which the estimated confidence interval contained the true value for the ICER. Distinguishing between observed noncoverage rates according to whether there was a spurious positive or negative treatment effect because skewed

distributions were expected to have different implications for them, the following percentages were obtained: (i) those for which the estimated ICER confidence interval did not contain the true ICER value *and* whose lower limit was greater than this value; (ii) those for which the estimated ICER confidence interval did not contain the true ICER value *and* whose upper limit was less than this value; (iii) those for which the estimated confidence interval contained the true ICER value (calculated simply as 100 minus the sum of the other two percentages. Ideally, then, (i) and (ii) should each be 2.5 percent, whereas (iii) should be 95 percent (the nominal level). Test runs found that 20,000 simulations were sufficient to reduce Monte Carlo variation to a level that enabled reasonable comparisons of the various methods for the different parameter sets.

RESULTS

Coverage Rates for Cost-Effectiveness Confidence Intervals

For the two control group ICCs investigated, Tables 2 and 3 show the coverage percentages for each of the methods of analysis for the various sample size combinations, averaging over the sixty-five parameter combinations (thirteen correlation combinations times five intervention ICC changes). From these results, neither of the two methods reproduced the nominal coverage of 95 percent but the BS2 method always achieved coverage closer to 95 percent than the BS1 method. When examining percentages along the diagonal in each table, the BS1 method performed much better for a large number of clusters and small cluster size compared with vice versa. This finding was probably due to the slight downward bias in the second moments; the degree of bias is an inverse function of the number of clusters. However, the double bootstrap did not exhibit this pattern. As the ICC in the control group increased (across tables) the performances of the bootstrap methods were noticeably poorer.

It is tempting to conclude at this point that the double bootstrap is the preferred method of analysis. However, it was hypothesized that correlations between costs and effects would have implications for the BS2 method. In particular, the rescaling procedure introduced by the double bootstrap is performed upon costs and effects separately and, therefore, takes no account of the correlation between them. It was unclear in advance how this would affect the BS2 coverage. The following sections, therefore, subdivide the coverage results first by correlation combination and then by variance changes for the BS2 method.

Table 2. Observed Coverage (%) for BS1 and BS2 Methods[a] (control ICC = 0.01)

		Cluster size	
		25	50
Number of clusters per arm	6	88.44 94.15	88.44 93.98
	12	91.42 93.98	

[a] Within each box, the first number represents the average coverage of the single bootstrap (BS1) method, whereasthe second represents the double bootstrap (BS2) method. ICC, intracluster correlation coefficients.

Table 3. Observed Coverage (%) for BS1 and BS2 Methods[a] (control ICC = 0.1)

		Cluster size	
		25	50
Number of clusters per arm	6	87.47 91.95	87.03 91.55
	12	90.09 92.05	

[a] Within each box, the first number represents the average coverage of the single bootstrap (BS1) method, whereas the second represents the double bootstrap (BS2) method.

Coverage Subdivided by Correlation Combination

Table 4 shows the coverage percentages for the BS2 method for each correlation combination when averaged over the five variance change combinations for a given control group ICC, using six clusters of size 25 per arm as an example.

The poor performance as the control ICC increased was again apparent. The BS2 method exhibited a 1 percent or greater decrease in coverage for ICC = 0.1 compared with ICC = 0.01 in eleven of the thirteen correlation combinations. Although there was some evidence of this for BS1 (not shown), a 1 percent decrease was apparent in only five of thirteen combinations. This finding was not surprising, given that the downward bias in the second moments is partly a function of the between cluster variance.

It seems that either or both of the rescaling mechanism and the second level of resampling utilized by the BS2 method affected confidence interval coverage. For zero cluster level correlation, as the individual level correlation increased from −1 to 1, coverage fell consistently from 95.0 percent to 90.7 percent for ICC = 0.01 and from 93.6 percent to 87.9 percent for ICC = 0.1. On the other hand, for zero individual level correlation, increases in the cluster level correlation from −1 to 1 led to increases in the coverage, from 93.4 percent to 95.7 percent and from 89.4 percent to 95.8 percent for the two ICCs, respectively. This difference in behavior may well be due to the rescaling mechanism or second level of resampling introduced by the BS2 method, especially because such trends were not apparent for BS1.

Repeating the above analyses for a cluster size of fifty rather than twenty-five led to the same patterns being observed (11). Moreover, for these other configurations, the BS2 method exhibited little if any improvement in matching the nominal 95 percent coverage probability. The

Table 4. Observed Coverage by Correlation Combination for Six Clusters of Size 25 per Arm: BS2 Method

			Individual level correlation				
			−1	−0.5	0	0.5	1
Control	Cluster level correlation	−1	94.45		93.40		
ICC = 0.01		−0.5		94.37	93.96		
		0	94.99	94.85	94.60	93.86	90.65
		0.5			95.21	94.54	
		1			95.65		93.38
Control	Cluster level correlation	−1	91.42		89.41		
ICC = 0.1		−0.5		91.69	90.71		
		0	93.60	93.09	91.93	90.89	87.93
		0.5			93.66	92.20	
		1			95.77		92.65

ICC, intracluster correlation coefficients.

performance of the BS1 method, however, improved for every correlation combination. More formally, the statistical consistency in the bootstrap moments was apparent; although twelve clusters was not sufficient to achieve 95 percent coverage, coverage was above 90 percent most of the time. Furthermore, for both values of the ICC, where the cluster level correlation was zero and the individual level correlation was 1, the BS1 method actually achieved better coverage than the BS2 (11).

Coverage Subdivided by Change in ICC

When coverage was subdivided by the change in the ICC, it was apparent that variance changes had comparatively little effect upon coverage. However, the relative performances of the two methods was much more dependent upon factors such as the bootstrap's bias in the second moments and the correlation combinations. Lastly, a larger control ICC tended on average to reduce the coverage of the bootstrap confidence intervals (11).

DISCUSSION

Summary of Findings

The conclusions were generally consistent with predictions based on *a priori* knowledge of the theoretical strengths and limitations of the methods. The single bootstrap performed badly on average, but it showed a marked improvement for a larger number of clusters per treatment arm. Furthermore, its performance was not strongly affected by the correlations between costs and effects or by changes in the ICC. The performance of the double bootstrap was somewhat disappointing. Although on average it more closely matched the nominal rates for coverage and rejection rates than the single bootstrap, its results demonstrated a high degree of variation. In particular, the two correlations between costs and effects had a greater impact than for the single bootstrap. Either the second level of resampling or, more probably, the rescaling method inherent in this method of analysis did not deal adequately with the nonstandard distribution of the ICER. A larger number of clusters per treatment arm did little to improve the performance of the double bootstrap, and there was some evidence to suggest that larger numbers of clusters per treatment arm would cause the single bootstrap to perform better than the double bootstrap on average.

Comparisons with the Literature

This work has drawn together the two distinct areas of "analysis of individually randomized economic data" and "analysis of clustered data". Comparing the simulation results for the bootstrap with those from the individually randomized literature suggests that the good confidence interval coverage of the bootstrap observed in several studies has not been replicated here (1;2;6;16). This finding is largely explained by the downward bias in the second moments introduced by simple bootstrapping of clustered data. However, attempts to improve the bootstrap's ability to take account of individual level factors by way of a second level of bootstrapping caused problems in analyzing cost-effectiveness data. The multilevel modeling software package MLWin version 1.10 includes bootstrap procedures (5). However, these involve parametric bootstrap procedures rather than nonparametric ones as utilized here, and they have yet to be evaluated in terms of their ability to take adequate account of correlations between costs and effects.

This work constitutes early stages in the further research that has been advocated to identify appropriate approaches to the analysis of economic outcomes from CRTs (4). The ICCs used in the present simulations were comparable to those estimated for costs in this previous study, but highly variable ICCs for costs at different levels and the magnitude of patient costs relative to total costs have both been emphasized as important issues (4). Whether any of the scenarios investigated here are relevant to future trials will depend,

in part, upon the issue of which cost component is most important.

Limitations of the Work Presented

Conceptualizing economic data from a CRT was difficult, given the almost total lack of empirical data on cost ICCs, the validity of distributional assumptions, and values for correlations between costs and effects. As a result, there were too many permutations of parameter values to be run in the time available and some simplifications of the models had to be made. Perhaps the most fundamental limitation of the models was that of a constant cluster size. Nonconstant cluster size is more realistic, but the bootstrap methods of Davison and Hinkley (7) were not intended to handle nonconstant cluster size and the more complex bootstrap methods of Rao and Wu (17) would be required.

Another limitation concerns the handling of effects, which were conceptualized as being a natural clinical unit such as blood pressure. The framework of analysis was then defined to be a cost-effectiveness analysis. By thus restricting the effect data, appeals to the central limit theorem were possible, thus reducing the generalizability of the results by ruling out alternative distributional assumptions that might be more reasonable if the outcome data were, for example, quality-adjusted life years.

The ICC (and the change in the ICC due to treatment) was constrained to be the same for both costs and effects. Further work on simulations using less-restrictive assumptions could be valuable, but again there is a lack of empirical data to inform the conceptualization and construction of such simulations.

Finally, a larger number of clusters per treatment arm, perhaps twenty-four, might have caused the single bootstrap to outperform the double bootstrap consistently. Unfortunately, given the complexity of the model and a large amount of time that the cost-effectiveness simulations took, such a large trial size was not feasible.

POLICY IMPLICATIONS

Despite these limitations, the results from the simulations present a coherent picture of the relative strengths of the two methods of analysis. Simple bootstrap methods such as those of Davison and Hinkley do not perform well when the number of clusters is small. The bias in the second moments is usually fairly large when the number of clusters per treatment arm takes values common to many CRTs.

This work should prompt researchers to consider the economic aspects of future CRTs so that such trials are not subject to poor design and analysis. Given the possibly extensive data requirements that will result from future economic evaluations and the difficulties encountered with the use of the bootstrap, this work may have provided a strong motivation for rethinking how costs and effects may be combined throughout clinical trials. For individually randomized trials, there has been work undertaken to combine costs and effects in a framework that rejects ratio statistics and moves back to a unidimensional outcome that can be analyzed using traditional methods that are utilized for clinical outcomes (the net benefit approach; 15). Such an approach may permit a simpler conceptualization of outcomes in CRTs. Furthermore, by considering issues such as the value and costs to society of obtaining information on treatments' effectiveness, the net benefit approach is consistent with the philosophy behind work on the design of CRTs (14). At the very least, it has the potential to promote a more unified approach to the design and analysis of economic data from such studies.

CONTACT INFORMATION

Terry N. Flynn, PhD, Research Fellow (terry.flynn@bristol.ac.uk), MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK
Tim J. Peters, PhD, Professor of Primary Care Health Services Research (tim.peters@bristol.ac.uk), Academic Unit of Primary Health Care, Department of Community Based Medicine, University of Bristol, The Grange 1, Woodland Road, Bristol BS8 1AU, UK

REFERENCES

1. 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.
2. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health Econ.* 1997;6:327-340.
3. Campbell M, Torgerson D. Confidence intervals for cost-effectiveness ratios: The use of 'bootstrapping.' *J Health Serv Res Policy.* 1997;2:253-255.
4. Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: Estimation of intracluster correlation coefficients and sample size. *Stat Med.* 2001;20:391-399.
5. Centre for Multilevel Modelling. *MLwiN software package*. (1.10). London: Centre for Multilevel Modelling; 2000.
6. Chaudhary MA, Stearns SC. Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Stat Med.* 1996;15:1447-1458.
7. Davison AC, Hinkley DV. *Bootstrap methods and their applications*. Cambridge: Cambridge University Press; 1997.
8. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold; 2000.
9. Efron B. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Rev.* 1979;21:460-480.
10. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci.* 1986;1:54-77.
11. Flynn TN. *Design and analysis of randomised controlled trials: Economic aspects of cluster randomisation*. PhD thesis. University of Bristol; 2002.

12. Flynn TN, Peters TJ. Conceptual issues in the analysis of economic data from cluster randomised trials. *J Health Serv Res Policy.* 2005;10:97-102.
13. Flynn TN, Peters TJ. Use of the bootstrap in analysing cost data from cluster randomised trials: Some simulation results. *BMC Health Serv Res.* 2004;4:33.
14. Flynn TN, Whitley E, Peters TJ. Recruitment strategies in a cluster randomized trial—cost implications. *Stat Med.* 2002;21:397-405.
15. Hoch JS, Briggs AH, Willan AR. Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Econ.* 2002;11:415-430.
16. Polsky D, Glick HA, Willke RJ, Schulman K. Confidence intervals for cost-effectiveness ratios: A comparison of four methods. *Health Econ.* 1997;6:243-252.
17. Rao JNK, Wu CFJ. Resampling inference with complex survey data. *J Am Stat Assoc.* 1988;83:231-241.
18. StataCorp. *Stata statistical software*: Release 8.0. College Station, TX: Stata Corporation; 2003.
19. 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 Technol Assess.* 1999;3:iii-92.
20. Wakker P, Klaasen MP. Confidence intervals for cost-effectiveness ratios. *Health Econ.* 1995;4:373-381.