Q J Med 1999; **92**:177–182

Commentary

QJM

Bootstrapping: estimating confidence intervals for cost-effectiveness ratios

M.K. CAMPBELL and D.J. TORGERSON[1]

From the Health Services Research Unit University of Aberdeen, Aberdeen, and [1]Centre for Health Economics, University of York, York, UK

Summary

Economic evaluations are increasingly being conducted alongside clinical trials of health interventions, with resource consequences being estimated from stochastic data. It is, therefore, important that economic evaluation results, like the clinical results, reflect the underlying variance within the sample data. A statistical methodology, known as bootstrapping, has recently been put forward as a potential method for calculating confidence intervals for cost-effectiveness ratios, yet it is still unusual to see economic evaluations reporting confidence intervals. In this paper we demonstrate the practical application of bootstrapping using real data from clinical trials, and conclude that bootstrapping is easily transferable from theory to practice for the estimation of confidence intervals for cost-effectiveness ratios. We encourage further investigation into its applicability and use.

Introduction

Traditionally, in many economic evaluations, the cost profile of a treatment has been informed by clinical judgement about what resources a typical patient might use for a given treatment. Over recent years, however, an increasing number of economic evaluations are being conducted alongside clinical trials, with resource consequences now being estimated from observations of a sample of patients.

Confidence intervals have been used for many years in the reporting of clinical data to reflect the stochastic nature of data collected from a sample of patients. The transfer of this methodology to economic reporting has not been straightforward, however, as methods to calculate exact confidence intervals for the more commonly used economic measures, such as cost-effectiveness ratios, do not exist.

Several authors have explored methods for the approximation of confidence intervals in this situation, and the use of a statistical methodology known as bootstrapping has been put forward as a potential solution.[1–5] Bootstrapping is a computationally intensive technique which allows the distribution of the cost-effectiveness ratio to be constructed empirically. Despite the proposal of these techniques as feasible alternatives for the calculation of confidence intervals, there have been few cost-effectiveness ratios reported in the literature to date when both costs and effects are variable.[6]

The aim of this paper is to review the principles of the bootstrap methodology for estimating confidence intervals for cost-effectiveness ratios when both cost and effectiveness are variable, and to highlight its practical application use through two examples using empirical data from clinical trials.

Economic measures

There are two commonly occurring objectives in economic evaluations. First, within a clinical trial

Address correspondence to Ms M.K. Campbell, Health Services Research Unit, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD situation, is the desire to describe the most cost-effective treatment alternative between at least two comparators. Second, there is the need for a wider comparison of efficiency between a large range of different competing health-care interventions. The two objectives require different economic approaches.

Comparison of treatments within a trial

Within a trial of two interventions, the incremental cost-effectiveness ratio (ICER) is the measure primarily used to compare the cost-effectiveness of the experimental treatment relative to the control treatment.[7] The ICER can be described as the ratio of the difference in costs to the difference in effects between the two treatments, or:

$$\frac{\bar{C}_e - \bar{C}_c}{\bar{E}_e - \bar{E}_c}$$

where $\bar{C}_e$ and $\bar{C}_c$ are the mean costs, and $\bar{E}_e$ and $\bar{E}_c$ are the mean effects for the experimental and control treatments, respectively.

Comparison of treatments outside a trial

To compare the cost-effectiveness of a particular treatment against other treatments outside the context of a trial, for example by comparing against published data, requires the use of a different economic measure. A commonly used measure is that of the marginal cost-utility ratio.[7] In this case, the effect of the treatment must be expressed in terms of a standardized measure to ensure comparability across treatments. The most common standardized measure of effect is quality-adjusted life years (QALY). In this case, the marginal cost-utility ratio would be described as ratio of the cost of the treatment to the number of QALYs gained, or

$$\frac{\bar{C}_t}{\bar{E}_t}$$

where $\bar{C}_t$ and $\bar{E}_t$ are the average cost of and the average QALY gain for the treatment.

Both these methods, when they are using stochastic data, require a statistical technique which will appropriately describe the underlying variance.

Bootstrap methods

Bootstrapping is a non-parametric technique which involves large numbers of repetitive computations to estimate the shape of a statistic's sampling distribution empirically.[8–10] The basic concept behind bootstrapping is to treat the study sample as if it were the population, the premise being that it is better to draw inferences from the sample in hand rather than make potentially unrealistic assumptions about the underlying population.

Using the bootstrap approach, repeated random samples of the same size as the original sample are drawn *with replacement* from the data. As such, the fact that an observation has been selected for inclusion in a resample does not preclude it from being selected again for the same resample. The statistic of interest is calculated from each resample, and these bootstrap estimates of the original statistic are then used to build up an empirical distribution for the statistic. The number of bootstrap resamples, *B*, required depends on the application, but typically *B* should be at least 1000 when the distribution is to be used to construct confidence intervals.[5,8] When constructing confidence intervals, this large number of resamples is required to ensure that the tails of the empirical distribution are filled. This process is pictorially represented in Figure 1.

For example, to generate a bootstrap distribution for an ICER using trial data, the following steps would be required (we assume that there were n_e patients in the experimental treatment group and n_c in the control treatment group):

1. Generate a sample of n_e cost and effect pairs from the experimental group data with replacement. The resampling procedure must reflect that by which the original data were obtained,[9] hence cost and effect pairs need be resampled together as they are inter-dependent.
2. Similarly, generate a sample of n_c cost and effect pairs from the control group data with replacement.
3. Calculate the ICER for this bootstrap resample.
4. Repeat this procedure 1000 times, to get 1000 bootstrap estimates of the ICER. These estimates then define the empirical sampling distribution of the ICER.

Bootstrap confidence intervals

A range of procedures have been developed for the construction of bootstrap confidence intervals, which include the normal approximation method, the percentile method, the percentile-t method, the bias-corrected percentile and the accelerated bias-corrected method. The optimal choice of method is, however, application-specific. A number of authors give a full description of each technique together with a summary of the main advantages and disadvantages of each method.[5,8] A full discussion of all these techniques is beyond the scope of this paper; we would refer readers to these other texts for a detailed comparison. We will, rather, illustrate the methodology through the use of the simple bias-corrected percentile method. We have chosen to use

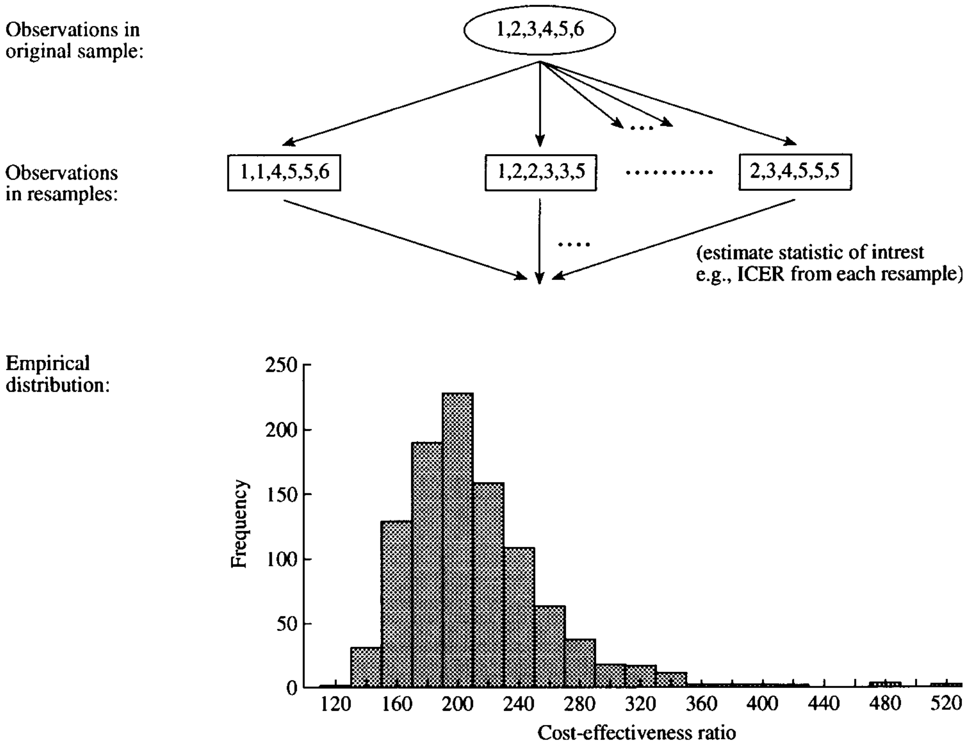

Figure 1. Bootstrap process.

a bias-corrected method to illustrate the technique as it has been shown that an ICER calculated from sample data is a biased estimate of the true population ICER.[11] We have chosen the simple bias-corrected approach to demonstrate the technique; however, the accelerated bias-corrected approach (which is a refinement of the simple approach) has been shown to perform better under a wider variety of assumptions.[5]

The bias-corrected percentile method adjusts for any bias in the bootstrap estimate, and, as the name implies, percentile-based methods use the percentiles of the generated bootstrap distribution to determine the limits of the confidence interval. To adjust for potential bias in the bootstrap estimates, two steps must be followed:

1. Calculate the bias-correcting constant, z_0, which is the standard normal deviate corresponding to the proportion of bootstrap estimates which are less than or equal to the estimate from the original sample. The estimate from the original sample ought to fall at the 50th percentile. If it does not, the bias-correcting constant makes a correction to adjust the confidence intervals in the appropriate direction. If the estimate from the original sample does fall at the 50th percentile, the resulting bootstrap confidence interval will be symmetric around the original estimate; if it does not, the bias-correcting constant allows for the confidence interval to be asymmetrical around its expected value.
2. Use this bias-correcting constant to modify the percentiles used to calculate the limits of the desired confidence interval, such that the lower limit of the bias-corrected confidence interval is the value of the bootstrapped estimate at the $\Phi[z_{\alpha/2}+2z_0]\times 100$ percentile and the upper limit is the value at the $\Phi[z_{1-\alpha/2}+2z_0]\times 100$ percentile; α is the desired level of significance eg 0.05; $z_{\alpha/2}$ is the standard normal deviate associated with the value $\alpha/2$; z_0 is the bias-correcting constant; and Φ represents the cumulative distribution of the standard normal function.

Example 1: Cost-effectiveness comparison within a trial

The Aberdeen Birthright randomized trial of alternative policies for managing mild cervical dyskaryosis

The cost-effectiveness of immediate colposcopy versus cytological surveillance for the management of mild cervical dyskaryosis was examined within the context of the Aberdeen Birthright randomized trial conducted in the North East of Scotland.[12] Women in the immediate colposcopy group had fixed treatment costs but variable effects, but women randomized to surveillance had variable costs due to differences in subsequent management: completion of surveillance with no recurrent dyskaryosis;

default from surveillance; or recurrent dyskaryosis leading to colposcopy.

One hundred and forty-five women were randomized to immediate colposcopy and 158 were randomized to the surveillance group. The average cost per woman for immediate colposcopy was £82.02 and for surveillance was £54.42, with 66 (46%) cases of disease detected in the immediate colposcopy group and 43 (27%) in the surveillance group. This leads to an ICER of:

$$\frac{£82.02-£54.42}{0.46-0.27}=\frac{£27.60}{0.19}=£145.26$$

Following the steps outlined above, 145 effect and cost pairs from the immediate colposcopy group were resampled with replacement, and 158 effect and cost pairs from the surveillance group. An ICER using this data was calculated. This process was repeated 1000 times. The 1000 bootstrap estimates of the ICER then provided the empirical sampling distribution from which the limits of the confidence interval would be taken.

Four hundred and fifty-eight of the 1000 bootstrap ICER estimates had values which were less than or equal to £145.26 (the estimate obtained from the trial data). Thus the bias correcting constant, z_0, is calculated to be:

$$z_0=\Phi^{-1}(0.458)=-0.105$$

Assuming a 95% confidence interval is desired, i.e. $\alpha=0.05$, then $z_{\alpha/2}=-1.96$ and $z_{1-\alpha/2}=1.96$. From this the appropriate confidence interval endpoints become: lower CI endpoint, the estimated ICER at the $\Phi[-1.96-0.21]\times 100=0.015\times 100=1.5$th percentile of the bootstrap distribution (i.e. the 15th largest bootstrap ICER estimate); upper CI endpoint, the estimated ICER at the $\Phi[1.96-0.21]\times 100=0.960\times 100=960$th percentile of the bootstrap distribution (i.e. the 960th largest bootstrap ICER estimate). (Software packages such as Microsoft Excel or MINITAB[13] contain a standard normal cumulative distribution function and can be used to return the values of z and Φ). This results in a 95% bootstrap bias-corrected confidence interval for the ICER of £94.01 to £309.33.

Example 2: Cost-utility comparisons outside a trial

This example uses data looking at the cost and health improvement associated with orthopaedic management of patients having orthopaedic care for a variety of musculo-skeletal conditions.

The cost-effectiveness of the routine service provided by orthopaedic surgeons for the management of non-surgical musculoskeletal conditions was examined for 233 patients at the Princess Margaret Rose Hospital, Edinburgh, Scotland.

Data relating to costs and benefits were collected for all patients, benefit being measured as absolute increase in quality of life score (based on the EuroQol quality of life measure[14]). As before, this resulted in both variable costs and benefits for patients.

The average cost of treatment for these patients was £335.15, with a corresponding average increase in EuroQol score of 0.102 points per patient. This leads to a marginal cost-utility ratio, or cost per unit increase in EuroQol score, of:

$$\frac{\bar{C}_t}{\bar{E}_t}=\frac{£335.15}{0.102}=£3285.78$$

As before, 233 effect and cost pairs from the data were resampled with replacement. The marginal cost-utility ratio for this data was calculated. Again, this process was repeated 1000 times.

On this occasion, 488 of the 1000 bootstrap estimates had values which were less than or equal to the original marginal cost-utility ratio. Thus the bias-correcting constant, z_0, for this dataset is calculated to be:

$$z_0=\Phi^{-1}(0.488)=-0.0301$$

Assuming a 95% confidence interval as before, recall that $z_{\alpha/2}=-1.96$ and $z_{1-\alpha/2}=1.96$. From this the appropriate confidence interval endpoints become: lower CI endpoint, the estimated cost-utility ratio at the $\Phi[-1.96-0.0602]\times 100=0.022\times 100=$ 2.2nd percentile of the bootstrap distribution (i.e. the 22nd largest bootstrap estimate); upper CI endpoint, the estimated cost-utility ratio at the $\Phi[1.96-0.0602]\times 100=0.971\times 100=97.1$th percentile of the bootstrap distribution (i.e. the 971th largest bootstrap estimate). This results in a 95% bootstrap bias-corrected confidence interval for the marginal cost-utility ratio of £2170.51 to £5369.18.

This cost-utility result can now be compared with other common health-care interventions to assess its relative worth. For example, comparing this result with other published cost-utility ratios,[15] we can show that the point estimate of cost utility for orthopaedic surgery renders it less cost-effective than routine treatment for hypertension (Table 1). However, the 95% confidence interval extends much lower, suggesting there is unlikely to be a real difference in cost utility between the two procedures. On the other hand the upper confidence limit places orthopaedic treatment lower than breast screening.

Discussion

Recently, randomized trials have started to include contemporaneous economic evaluations, and indeed

Table 1 Cost per QALY for range of interventions

Intervention	Cost/QALY (£)
Advice by GP to stop smoking	330
HRT for menopausal symptoms	550
Coronary artery bypass grafting for severe angina	1925
Treatment of hypertension	3135
Routine orthopaedic treatment for musculo-skeletal disorders	*3291* *(95% CI: 2171 to 5369)*
Breast cancer screening	6105
Heart transplantation	14735

Published cost per QALY data[15] adjusted to 1997 costs.

there is obvious intuitive appeal in measuring both cost and effect data on the same patients. With increasing emphasis on the use of confidence intervals when reporting the results of clinical trials, simple point estimates of cost-effectiveness ratios based on data which are variable will rapidly become unacceptable.

In recent years, the problem of confidence interval generation for economic analysis has been highlighted, and bootstrap techniques raised as a potential solution.[1–5] The primary benefit of bootstrap techniques is that they require no assumptions as to the shape of the sampling distribution of the statistic of interest. In this paper we have shown the practical application of the technique to stochastic cost and effect data, and have demonstrated that the technique is straightforward to apply with real-life data.

To date, however, there have been few cost-effectiveness ratios reported in the literature when both costs and effects are variable.[6] Computational difficulties with the technique have historically restricted the use of resampling techniques such as bootstrapping, but with the advances of modern computing power, these difficulties should no longer exist. Despite this, as the routine adoption of resampling techniques is a fairly recent trend, the majority of software programs currently available to undertake bootstrapping have been custom-built. Statistical packages such as STATA[16] and RATS[17] do have bootstrap procedures in-built, however, and the macro and/or syntax facilities within other statistical packages can be adapted to run the procedure.

Bootstrapping does have limitations, however. For example, Briggs *et al.* raise the concern that a theoretical assumption of the bootstrap, that the second moment exists, may be questionable if there is a distinct possibility of obtaining a zero or near-zero value on the denominator of the ICER.[5] Other concerns have been raised by a number of authors[3,5,8] into the validity of other assumptions for particular applications of the bootstrap, such as the applicability of bootstrapping when the initial sample is small. Further research is currently being carried out to address these issues.

Mathematical techniques, such as the parametric method based on Fieller's theorem, have also been put forward as potential methods for calculating confidence intervals for cost-effectiveness ratios.[4,18] Fieller's method does provide analytic solutions to the confidence limits, and may be seen as a more powerful approach than bootstrapping. There are, however, limitations to this technique, one of the most important being that implausible values may be returned for the confidence limits (e.g. returning a negative value when only positive values are possible in practice).[2] There is also a concern over the validity of parametric assumptions, when the sampling distribution of statistics such as the ICER are unknown.[5]

Economists have also traditionally used sensitivity analysis rather than confidence intervals to express uncertainty with regard to estimates of costs and/or benefits. It is, however, possible to combine sensitivity analysis with confidence intervals.[3] For example, if the cost of a procedure is subject to external variation e.g. regional variation, the cost of the procedure can be varied through sensitivity analysis with different average estimates and confidence intervals generated. In the Aberdeen study, for example, the cost of routine cervical smears was £7.01.[12] In other centres, however, other costs have been quoted. The NHS cervical screening programme, for example, estimated the costs of routine cervical smears at £17.[19] Leaving all other parameters unchanged, but varying the cost of routine smears to £17, a new bias-corrected bootstrap confidence interval for the ICER can be calculated, leading to a revised ICER from the sample data of £45.85 with a 95% bootstrap confidence interval ranging from £19.55 to £104.88.

In conclusion, we have shown that non-parametric bootstrapping for the calculation of confidence intervals for cost-effectiveness ratios is straightforward to apply within the practical context of a randomized trial, and we encourage further investigation into its application and use.

Acknowledgements

We thank Doug Altman for comments on an earlier draft of this paper, and the referee for helpful comments. Marion Campbell, through the Health Services Research Unit, is supported by the Chief Scientist Office of the Scottish Office Department of Health. David Torgerson is a Research Fellow in the Centre for Health Economics at York. The views expressed are not necessarily those of the funding bodies.

References

1. O'Brien BJ, Drummond MF, Labelle RJ, *et al.* In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med Care* 1994; **32**:150–63.
2. Chaudhary MA, Stearns SC. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Stat Med* 1996; **15**:1447–58.
3. Campbell MK, Torgerson DJ. Confidence intervals for cost-effectiveness ratios: the use of bootstrapping. *JHSRP* 1997; **2**:253–5.
4. Polsky D, Glick HA, Willke R, *et al.* Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Economics* 1997; **6**:243–52.
5. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics* 1997; **6**:327–40.
6. Wakker P, Klassen MP. Confidence intervals for cost/effectiveness ratios. *Health Economics* 1995; **4**:373–81.
7. Drummond MF. *Principles of economics appraisal in health care*. Oxford, Oxford University Press, 1980.
8. Mooney CZ, Duval RD. *Bootstrapping: a non-parametric approach to statistical inference*. Newbury Park CA, Sage, 1993.
9. Efron B, Tibshirani RJ. *An introduction to the bootstrap.* New York, Chapman & Hall, 1993.
10. Stine R. An introduction to bootstrap methods. In: Fox J, Long JS, eds. *Modern Methods of Data Analysis*. Newbury Park CA, Sage, 1990.
11. Stinnett A. Adjusting for bias in C/E ratio estimates. *Health Economics* 1996; **5**:470–2.
12. Flannelly GM, Campbell MK, Meldrum P, *et al.* Immediate colposcopy or cytological surveillance for women with mild dyskaryosis: a comparative cost analysis. *J Pub Health* 1997; **19**:419–23.
13. Ryan BF, Joiner BL, Ryan TA. *Minitab handbook*, 2nd edn. Boston, Duxbury Press, 1985.
14. The EuroQol Group. A new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16**:199–208.
15. Daly E, Vessey MP, Barlow D, Gray A, McPherson K, Roche M. In Berg G, Hammar M, eds. *Hormone replacement therapy in a risk-benefit perspective: the modern management of the menopause*. London, 1994.
16. Computing Resources Center. *STATA reference manual: release 3*. Santa Monica, 1992.
17. Doan TA. *RATS User 's Manual: Version 4*. Evanston, Estima, 1992.
18. Willan AR, O'Brien BJ. Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Economics* 1996; **5**:297–305.
19. Havelock C. The cost of the cervical screening programme—an activity based approach. In: Muir Gray JA, National Coordinating Network, eds. *NHS Cervical Screening Programme*. Oxford, 1994.