UNCERTAINTY OF INCREMENTAL COST-EFFECTIVENESS RATIOS

A Comparison of Fieller and Bootstrap Confidence Intervals

Johan L. Severens
Theo M. De Boo
Emmy M. Konst
University of Nijmegen

Abstract

Objective: To compare different methods to estimate the confidence interval of the incremental cost-effectiveness ratio (ICER).
Methods: The adequacy of Fieller intervals and three methods for calculating bootstrap intervals are compared based on a simulation of 10,000 trials, using data from one trial.
Results: Both Fieller and bootstrap methods lead to unsatisfactory results when the difference in effectiveness is approximately zero. Where this difference is significant, the four methods for calculating confidence intervals for ICER do not give very different results, but Fieller's interval performs best.
Conclusions: Since Fieller's confidence limits are relatively easy to compute compared with bootstrap simulations, we recommend using this method.

Keywords: Cost-effectiveness, Confidence Intervals, Ratios

Increasingly, cost-effectiveness analyses are conducted as part of clinical trials (5;8). Gathering data on both effectiveness and cost for each patient taking part in a trial makes it possible to determine the mean and standard deviation of costs, *C,* and effects, *E,* of the medical interventions. The incremental cost-effectiveness ratio (ICER) comparing a treatment or diagnostic facility to some alternative (1 and 0, respectively) is considered to be the main result of cost-effectiveness analysis (1;9;11). This incremental cost-effectiveness ratio is defined as:

$$\hat{R} = \frac{\overline{C}_1 - \overline{C}_0}{\overline{E}_1 - \overline{E}_0}$$

However, $\hat{R}$ is a point estimate that does not give any real insight into uncertainty of $\hat{R}$ itself. Several methods have been explored to estimate a confidence interval for R (3;4;10;12;13;16;18;19;20). There is some agreement in the literature that methods that fully rely on a normal distribution of the ICER should be avoided in determining confidence intervals of R. Different methods of calculating nonparametric bootstrap intervals and the Fieller method fulfill this criterion. In this paper we report our experiences comparing Fieller intervals and three methods for calculating bootstrap intervals: the percentile method and two bias-corrected and accelerated methods.

METHODS

Trial Data and Simulation

The data used for comparing the methods in estimating confidence intervals were gathered in the framework of the Dutch PSOT study. The PSOT study was designed as a two-group randomized multicenter clinical trial concerning orthodontic treatment of children born with unilateral cleft lip and palate. More detailed information about the PSOT study is reported elsewhere (14).

From the trial data, mean and variance of cost and effectiveness for each group and the correlation between them were computed. These statistics were used to simulate the subsequent 10,000 trials by repeatedly drawing 2 times 10 cases from normal distributions with these statistics as parameters. For each simulation, $\hat{R}_s^*$ was calculated and the mean ICER $\overline{R}_s$ was used as an estimate for the true population ICER, R.

Confidence Intervals of the ICER

The Fieller theorem method is a parametric method for calculating a confidence interval of a ratio of means. The assumption on which this method is based is bivariate normality of numerator and denominator, here the difference between the means of the cost and the means of effect (10). On the basis of each $\hat{R}_s^*$ of the 10,000 simulations, a 90% Fieller confidence interval was calculated (see Appendix 1 for the complete calculations).

The principle of bootstrapping is that a random sample of size n with replacement from the data is taken a large number of times (7). As a result the bootstrap ratio $\hat{R}_b^*$ can be calculated from each bootstrap series. For each trial simulation we performed 25,000 bootstrap replicates, and three methods were used to calculate bootstrap confidence intervals for R: a) the percentile method; b) the bias corrected and accelerated (BCA) percentile 1 method (BCA1); and c) the BCA2 method (Appendix 2).

For each of the four methods of calculating 90% confidence intervals (Fieller theorem, bootstrap percentile method, bootstrap BCA1, and bootstrap BCA2), 10,000 confidence intervals were determined. The adequacy of the four methods was investigated by comparing the percentages of confidence intervals containing $\overline{R}_s$. Approximately 90% of the intervals should contain this estimate, because a level of miscoverage of 10% (type 1 error) was prespecified (Figure 1).

Effect of $E_1 - E_0 \approx 0$ on the Confidence Intervals

Our original trial data showed a significant difference in effectiveness between the two groups of 1.34 on the scale of 1 to 10 (a professional judgment of overall speech and language performance). To investigate the impact of the relative importance of the magnitude of the difference in effectiveness between the groups, we simulated four other situations: the difference in effectiveness was assumed to be 0.5 and 1.0 less, and 1.0 and 5.0 more than in the original trial, which reflect the situations where $E_1 - E_0 \approx 0$ and where $E_1 - E_0 >> 0$, respectively. Hence, in total five approaches were used to compare the results of Fieller intervals and three bootstrap intervals.

RESULTS

Trial Data

The trial concerned 20 patients (10 receiving orthodontic treatment for unilateral cleft lip and palate, 10 with no treatment). For the treatment group the mean

https://doi.org/10.1017/S0266462399153157 Published online by Cambridge University Press

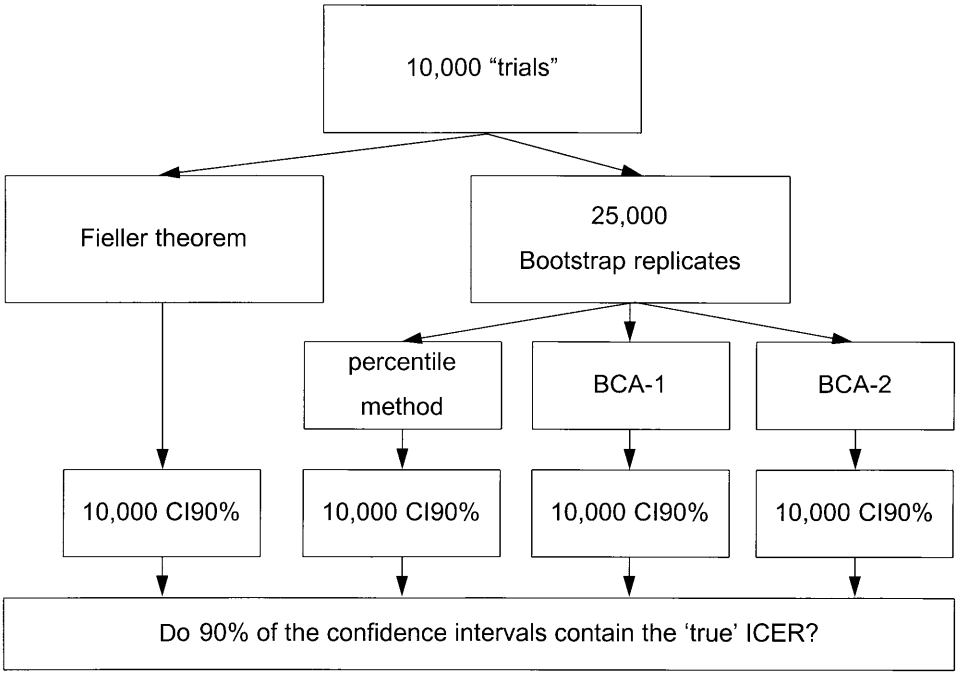

Figure 1. Overview of the simulation data and methods for calculating confidence intervals (Fieller theorem, bootstrap percentile method, bootstrap bias corrected and accelerated 1 [BCA1], bootstrap bias corrected and accelerated 2 [BCA2]), and calculation of the adequacy of these methods. Based on the 10,000 simulated trials, for each simulation 90% confidence intervals were computed using the four methods. The adequacy of these methods was investigated by calculating the percentage of confidence intervals containing the estimate of the true incremental cost-effectiveness ratio.

medical cost was Dutch guilders (Dfl) 2,544, standard deviation Dfl 646, and the mean effectiveness score (for speech and language development) and standard deviation were 3.52 and 1.75, respectively. For this group, correlation between costs and effectiveness was 0.35. Mean costs and standard deviation for the no-treatment group were Dfl 881 and Dfl 151. Mean effectiveness and standard deviation were 2.18 and 0.62, respectively. Correlation between costs and effectiveness for no treatment was −0.06.

Both the difference between the means of the costs (Dfl 1,663) and the means of effectiveness (1.34) were significant. The correlation between the difference in cost and the difference in effects was +0.30. The ICER based on the trial data for the speech development score is Dfl 1,241/point score improvement.

Simulations and Confidence Intervals for $\hat{R}$

Table 1 summarizes the result of the simulations for the chosen levels of difference in effectiveness. The number of simulated trials on which $\overline{R}_s$ is based is not always equal to 10,000, due to the fact that whenever a simulated trial leads to $E^*_{s1} - E^*_{s0} = 0$, $\hat{R}^*_s$ cannot be calculated for this simulation and therefore is neglected when calculating $\overline{R}_s$.

In Table 2, the percentages of the confidence intervals that contain $\overline{R}_s$ are mentioned for the four different methods, separately for the different levels of difference in effectiveness. In the case of $E_1 - E_0 \approx 0$ (baseline difference minus 1.0),

https://doi.org/10.1017/S0266462399153157 Published online by Cambridge University Press

Table 1. Results of the Simulations for Five Levels of Difference in Effectiveness

Level of difference in effectiveness	n	$\overline{R}_s$ (mean ICER)
Baseline difference minus 1.0	9,935	2,388
Baseline difference minus 0.5	9.982	2,692
Baseline difference (1.34)	9,997	2,053
Baseline difference plus 1.0	10,000	941
Baseline difference plus 5.0	10,000	633

both bootstrap-based BCA1 and BCA2 calculations lead to confidence intervals that contain $\overline{R}_s$ in less than 70% of the cases. In contrast, the bootstrap percentile method contains $\overline{R}_s$ 96% of the time, indicating (much) too wide intervals. This is not surprising in view of the fact that the denominator of R_b^* is often practically zero. The Fieller method seems to have intermediate results; however, the results do not appear to be stable when comparing the situation regarding a baseline difference minus 0.5 (reflecting $E_1 - E_1 \approx 0$ to a lesser extent) and the baseline effectiveness difference (reflecting $E_1 - E_0 \neq 0$). When $E_1 - E_0 > 0$ (baseline difference plus 1.0 and 5.0, respectively), the target percentage of 90% coverage of $\overline{R}_s$ by the 10,000 confidence intervals is best obtained by the Fieller method. All bootstrap-based calculating methods give somewhat too narrow intervals, thus less than 90% contain $\overline{R}_s$.

DISCUSSION

Recently, an alternative approach for performing cost-effectiveness analysis was suggested by Tambour et al. (17) that gives a solution to the difficulty in calculating confidence intervals for ratios. In this approach the effectiveness units as used in a study are multiplied by the price per effectiveness unit, thus resulting in expression of both costs and effectiveness in monetary terms. The authors describe that in this way the net benefits of the medical interventions that are compared can be determined, and that standard statistical techniques can be used to calculate confidence intervals for the net benefits. However, we think that this approach in general is difficult to apply. First, the method makes it necessary to determine a price per effectiveness unit, which seems to be rather arbitrary. Although sensitivity analysis might be used to explore the impact of varying the unit price on the studies' conclusions (2), it can be argued that specific measures of effectiveness are difficult to express in monetary terms. Second, the example that is used by the authors expresses effectiveness in quality-adjusted life-years (QALYs), which seem possible

Table 2. Percentage of the 10,000 90% Confidence Intervals that Contain the Mean Incremental Cost-effectiveness Ratio $\overline{R}_s$ for Four Methods of Calculating the Confidence Intervals

		Bootstrap		
Level of difference in effectiveness	Fieller	Percentile	BCA1	BCA2
Baseline difference minus 1.0	75.9	96.5	69.2	60.6
Baseline difference minus 0.5	89.0	90.9	77.0	72.8
Baseline difference (1.34)	84.1	82.0	79.6	77.7
Baseline difference plus 1.0	88.2	85.7	85.5	85.6
Baseline difference plus 5.0	89.2	85.5	85.5	85.5

to translate into monetary terms quite easily. However, the question arises of how to translate other effectiveness measures, because QALYs are only one of many ways in which to express effectiveness (15).

Before answering the question of which method should be used to determine a confidence interval for an ICER, a distinction must be made between the situation where $E_1 - E_0 \approx 0$ and where $E_1 - E_0 >> 0$. Actually, in case a trial does not show a significant difference in effectiveness, it does not make sense to calculate confidence intervals for R, making the discussion about which confidence interval method to use academic (4).

In conclusion, whenever there is a strictly significant difference in effectiveness between groups, calculating a confidence interval for the ICER is useful. Fieller's theorem, which is based on a single formula, leads to satisfactory results, and the necessary calculations are done relatively easily compared with bootstrap simulations, which require a rather powerful computer. We therefore recommend using the Fieller confidence limits in such cases.

REFERENCES

1. Briggs, A., & Fenn, P. Trying to do better than average: A commentary on 'statistical inference for cost-effectiveness ratios.' *Health Economics*, 1997, 6, 491–95.
2. Briggs, A., Sculpher, M., & Buxton, M. Uncertainty in the economic evaluation of health care technologies: The role of sensitivity analysis. *Health Economics*, 1994, 3, 95–104.
3. Briggs, A. H., Wonderling, D. E., & Mooney, C.Z. Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health Economics*, 1997, 6, 327–40.
4. Chaudhary, M. A., & Stearns, S. C. Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Statistics in Medicine*, 1996, 15, 1447–58.
5. Drummond, M. F., & Davies, L. Economic analysis alongside clinical trials. *International Journal of Technology Assessment in Health Care*, 1991, 7, 561–73.
6. Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 1987, 82, 171–88.
7. Efron, B., & Tibshirani, R. J. An introduction to the bootstrap. In D. R. Cox, D. V. Hinkley, N. Reid, D. B. Rubin, & B. W. Silverman, (eds.), *Monographs on statistics and applied probability*, 57. New York: Chapman and Hall, 1993.
8. Glick, H. G. Strategies for economic assessment during the development of new drugs. *Drug Information Journal*, 1995, 29, 1391–403.
9. Johannesson, M. On the estimation of cost-effectiveness ratios. *Health Policy*, 1995, 31, 225–29.
10. Laska, E., Meisner, M., & Siegel, C. Statistical inference for cost-effectiveness ratios. *Health Economics*, 1997, 6, 229–42.
11. Laska, E. M., Meisner, M., & Siegel, C. The usefulness of average cost-effectiveness ratios. *Health Economics*, 1997, 6, 497–504.
12. O'Brien, B., Drummond, M. F., Labelle, R. J., & Willan, A. In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care*, 1994, 32, 150–63.
13. Polsky, D., Glick, H. A., Willke, R., & Schulman, K. Confidence intervals for cost-effectiveness ratios: A comparison of four methods. *Health Economics*, 1997, 6, 243–52.
14. Severens, J. L., Prahl, C., Kuijpers-Jagtman, A. M., & Prahl, B. Short term cost-effectiveness of presurgical orthopaedic treatment in children with complete unilateral cleft lip and palate. *Cleft Palate-Craniofacial Journal*, 1998, 35, 222–26.
15. Severens, J. L., & van der Wilt, G. J. Economic evaluation of diagnostic tests: A review of published studies. *International Journal of Technology Assessment in Health Care*, 1999, 15, 480–96.

16. Tambour, M., & Zethraeus, N. Bootstrap confidence intervals for cost-effectiveness ratios: Some simulation results. *Health Economics*, 1998, 7, 143–47.
17. Tambour, M., Zethraeus, N., & Johannesson, M. A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care*, 1998, 14, 467–71.
18. Van Hout, B. A., Al, M. J., Gordon, G. S., & Rutten, F. F. H. Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics*, 1994, 3, 309–19.
19. Wakker, P., & Klaassen, M. P. Confidence intervals for cost-effectiveness ratios. *Health Economics*, 1995, 4, 373–81.
20. Willan, A.R., & O'Brien, B. Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. *Health Economics*, 1996, 5, 297–305.

Appendix 1

Fieller Theorem

Let ΔE and ΔC denote the mean difference in effect and cost respectively, $S^2_{\Delta E}$ and $S^2_{\Delta C}$ their estimated variances and r the estimated (Pearson) correlation coefficient between them. Let $f_{\upsilon,1-\alpha}$ denote the upper percentage point of the F-distribution, with 1 and υ degrees of freedom, υ being the number of degrees of freedom upon which the estimated variance of $\Delta E - \Delta C$ is based (18 in our case). Compute (10):

$$L_1 = \frac{(\Delta E\Delta C - f_{\upsilon,1-\alpha} r s_{\Delta E} s_{\Delta c}) - [(\Delta E\Delta C - f_{\upsilon,1-\alpha} r s_{\Delta E} s_{\Delta c})^2 - (\Delta E^2 - f_{\upsilon,1-\alpha} s^2_{\Delta E})(\Delta C^2 - f_{\upsilon,1-\alpha} s^2_{\Delta C})]^{1/2}}{\Delta E^2 - f_{\upsilon,1-\alpha} s^2_{\Delta E}}$$

and

$$L_2 = \frac{(\Delta E\Delta C - f_{\upsilon,1-\alpha} r s_{\Delta E} s_{\Delta c}) + [(\Delta E\Delta C - f_{\upsilon,1-\alpha} r s_{\Delta E} s_{\Delta c})^2 - (\Delta E^2 - f_{\upsilon,1-\alpha} s^2_{\Delta E})(\Delta C^2 - f_{\upsilon,1-\alpha} s^2_{\Delta C})]^{1/2}}{\Delta E^2 - f_{\upsilon,1-\alpha} s^2_{\Delta E}}$$

Now, if there is a statistically significant difference in effect, then (and only then) the denominators of L_1 and L_2 are positive. The Fieller $(1 - \alpha)$ interval is then the interval (L_1,L_2). If there is no significant difference in effect, the denominators are negative and Fieller's interval consists of the union of the intervals $(-\infty,L_2$ and $(L_1,+\infty)$. For further details see Laska et al. (10). (Note that in that paper there is a misprint on page 235: the second plus-sign in the numerator of the formula for the upper confidence limit estimator should be a minus-sign, as in the lower confidence limit.)

Appendix 2

Bootstrap Confidence Intervals

The percentile method is based on the principle of sorting the $\hat{R}^*_\beta$. When 25,000 replicates have been made, the percentile method uses the 1,250th and 23,750th ranked $\hat{R}^*_\beta$ as the 90% confidence interval limits. The basic principle of the BCA methods is a modification of the percentile method making a correction for bias and the skewness of the estimator of the sampling distribution (3). For an extensive description of this method we suggest Efron and Thibshirani (7) and Efron (6). The BCA method uses an acceleration constant which is used to adjust for the skewness of the sampling distribution of $\hat{R}^*_\beta$. This acceleration constant is calculated using a jack-knife estimate. Since the jack-knife method is not straightforward

described in case of comparison of two groups, we used two options: BCA1 and BCA2. For BCA1 an estimator is used that leaves out only one measurement at a time, irrespective of the origin (treatment or no treatment). BCA2 uses an estimator based on simultaneously leaving out one measurement from each of the two samples, treatment and no treatment.

https://doi.org/10.1017/S0266462399153157 Published online by Cambridge University Press