CONFIDENCE INTERVALS FOR COST-EFFECTIVENESS RATIOS: A COMPARISON OF FOUR METHODS

DANIEL POLSKY1, HENRY A. GLICK1,*, RICHARD WILLKE2 AND KEVIN SCHULMAN3
1Division of General Internal Medicine and Leonard Davis Institute of Health Economics, University of Pennsylvania, USA
2Pharmacia and Upjohn, Inc., Kalamazoo, USA
3Clinical Economics Research Unit, Division of General Internal Medicine, Georgetown University Medical Center, Washington, DC, USA

SUMMARY
We evaluated four methods for computing confidence intervals for cost–effectiveness ratios developed from randomized controlled trials: the box method, the Taylor series method, the nonparametric bootstrap method and the Fieller theorem method. We performed a Monte Carlo experiment to compare these methods. We investigated the relative performance of each method and assessed whether or not it was affected by differing distributions of costs (normal and log normal) and effects (10% absolute difference in mortality resulting from mortality rates of 25% versus 15% in the two groups as well as from mortality rates of 55% versus 45%) or by differing levels of correlation between the costs and effects (correlations of -0.50, -0.25, 0.0, 0.25 and 0.50). The principal criterion used to evaluate the performance of the methods was the probability of miscoverage. Symmetrical miscoverage of the intervals was used as a secondary criterion for evaluating the four methods.

Overall probabilities of miscoverage for the nonparametric bootstrap method and the Fieller theorem method were more accurate than those for the other the methods. The Taylor series method had confidence intervals that asymmetrically underestimated the upper limit of the interval. Confidence intervals for cost–effectiveness ratios resulting from the nonparametric bootstrap method and the Fieller theorem method were more dependably accurate than those estimated using the Taylor series or box methods. Routine reporting of these intervals will allow individuals using cost–effectiveness ratios to make clinical and policy judgments to better identify when an intervention is a good value for its cost. © 1997 by John Wiley & Sons, Ltd.

INTRODUCTION
As costs have become more important in health care decision making, the number of economic evaluations of medical treatments has grown.1–4 This trend has been accompanied by an increase in the number of economic evaluations that are conducted as part of clinical trials.5–7 Many of these evaluations have used confidence intervals to assess the precision of separate estimates of costs and effects. However, for comparisons of costs and effects (for example, using cost-effectiveness ratios), direct measures of statistical precision are not frequently used. Instead, the convention has been to perform sensitivity analyses to evaluate the robustness of the comparisons in these analyses.
One reason for this difference is that the methods for computing separate confidence intervals for costs and effects are well developed, whereas the methods for assessing confidence intervals for cost-effectiveness ratios are not. Several techniques have recently been proposed for calculating these intervals, but few researchers have evaluated the relative performance of these techniques. In this paper, we use Monte Carlo simulation to assess the performance of four methods that have been proposed for computing confidence intervals for cost-effectiveness ratios.

**METHODS**

We compared four methods for estimating confidence intervals for cost-effectiveness ratios, a method based on independent confidence intervals for costs and effects (referred to as the 'box' method), the Taylor series method, the non-parametric bootstrap method and the Fieller theorem method. We investigated the relative performance of each method and assessed whether it was affected by differing distributions of costs and effects or by differing levels of correlation between the costs and effects.

**Economic analysis in randomized clinical trials**

Randomized clinical trials are frequently used to estimate the effect of an active intervention relative to the effect of another active intervention or placebo. When data on both costs and effects are collected prospectively, changes in costs associated with a therapy can also be estimated.

Differences in costs and effects between two therapies can fall into one of the four quadrants defined in Fig. 1. One therapy may be more costly and less effective (quadrant II), in which case it is said to be dominated by the less costly, more effective alternative therapy. One therapy may be less costly and more effective (quadrant IV), in which case it is said to dominate the more costly, less effective alternative therapy. Finally, a therapy may be more costly and more effective (quadrant I) or it may be less costly and less effective (quadrant III). In these last two cases, the tradeoff between costs and effects is summarized by an incremental (or marginal) cost-effectiveness ratio (referred to as the cost-effectiveness ratio).

The cost-effectiveness ratio is computed by dividing the difference in the mean costs of the two therapies (referred to as costs) by the difference in the mean effects of the therapies (referred to as effects). In Fig. 1, the ratio for any point in quadrants I and III is defined as the slope of the ray connecting the origin and the point. All points on the ray have the same cost-effectiveness ratio. The steeper the slope of this ray, the greater is the ratio.

**Measures of precision**

Because they are based on a sample from the population, the costs, effects and cost-effectiveness ratio reported from a randomized clinical trial are estimates of the true population values. The degree of precision of these estimates is related to the sample size and the amount of random variation in the population.

A 95% confidence interval is a standard statistical measure of precision for estimates with sample variation. This interval defines a range within which one can be 95% confident the true value lies. That is, if a randomized clinical trial was repeated an infinite number of times, the confidence intervals computed in 95% of them would include (or 'cover') the true value from the population, whereas in 5% of them, the true value would fall outside the interval (i.e. fail to cover). Intervals that miscover — either by over- or under-covering — are poorly specified. If a confidence interval over-covers (i.e. the confidence intervals computed in more than 95% of the repeated experiments contain the true value for the population), it expresses too little confidence in the estimate; in this case, a narrower confidence interval that provides 95% coverage should be identified. If a confidence interval under-covers (i.e. if the confidence intervals in fewer than 95% of repeated experiments contain the true value), it expresses too much confidence in the estimate; in this case, a wider confidence interval that provides 95% coverage should be identified.

For a two-tailed 95% confidence interval, miscoverage should be symmetric. In other words, the lower limit of the confidence interval should exceed the true value for the population 2.5% of the time and the true value should exceed the...
upper limit of the confidence interval 2.5% of the time.

Procedures for computing confidence intervals

Computing a 95% confidence interval for the estimate of either costs or effects depends on the distribution of the variable, its mean and variance and the sample size. Formulae for computing these intervals are readily available in texts, the intervals themselves are reported by many commonly used statistical packages. These intervals are reliable because unbiased and efficient estimates are available for the standard errors of the costs and effects and because the distributions are approximately normal when the sample size is sufficiently large.

Direct methods for computing confidence intervals for cost–effectiveness ratios, on the other hand, are not appropriate because the distribution of the ratio may not be known and may not be well behaved and because there is no known unbiased and efficient estimator of the ratio’s standard error.

In an attempt to overcome these difficulties, several procedures have recently been proposed for calculating confidence intervals for cost–effectiveness ratios. We evaluated four of these proposed procedures: the box method, the Taylor series method, the nonparametric bootstrap method and the Fieller theorem method.

Box method. The box method uses confidence limits computed separately for costs and effects to calculate a confidence interval for the cost–effectiveness ratio. The lower limit of the confidence interval for the ratio is defined by the lower confidence limit for costs divided by the upper confidence limit for effects. The upper limit of the confidence interval for the ratio is defined by the upper limit for costs divided by the lower limit for effects. Several authors have commented that using confidence limits from the 95% confidence intervals for costs and effects for the box method yields an inappropriately wide confidence interval for the cost–effectiveness ratio. This problem is due to the fact that the box method fails to account for the elliptical shape of the bivariate distribution. For example, when costs are below their lower confidence limit 2.5% of the time, the estimate of effects can fall anywhere within its distribution (rather than falling above the upper limit of the effect confidence limit). To avoid this problem, we defined...
the confidence interval for the ratio using narrower independent confidence intervals for costs and effects such that they jointly yielded a 95% confidence interval for the ratio. For costs to fall on one side of their confidence limit 2.5% of the time (e.g. below the lower limit of costs), while effects fall on the other side of their limit 2.5% of the time (e.g. above the upper limit of effects), the upper and lower confidence limits on cost and effects must each exclude 15.8% of the distribution (0.0251/2). We thus used the confidence limits from the 68.4% [1 - (2 × 0.158)] confidence intervals for costs and effects. The z-score for these intervals is approximately 1.0 and the confidence intervals for cost and effects are defined as their mean ± (1.0 × standard error) [rather than mean ± (1.96 × standard error)]. Wakker and Klaassen11 have proposed a modified box method to be used when the difference in the effects of two treatments (i.e. the denominator of the cost–effectiveness ratio) significantly differs from zero. The box method implicitly assumes that the correlation between costs and effects is zero but does not make potentially restrictive assumptions about the normality or symmetry of the distribution of the cost–effectiveness ratio.

Taylor series method. This method, proposed by O’Brien et al.9 involves estimating the standard error of the cost–effectiveness ratio itself by a Taylor series approximation (see their paper for computational details9,21). The confidence interval is defined by two points, each of which is 1.96 estimated standard errors from the sample ratio. This method incorporates the correlations between costs and effects into its standard error calculation, but assumes the cost–effectiveness ratio estimate is normally distributed (which, among other properties, requires that the confidence interval be symmetric about the ratio).

Nonparametric bootstrap method. The third method involves resampling from the study sample and computing cost–effectiveness ratios in each of the multiple samples.9,15 A n underlying principle of the nonparametric bootstrap technique is that a random sample of size n (for example, the patients in a randomized, controlled trial) is an empirical distribution function that estimates the probability distribution of the population.22 Given this principle, the bootstrap algorithm for estimating a confidence interval is performed by first drawing a sample of size n with replacement from the empiric distribution and using it to compute a cost–effectiveness ratio. Second, this sampling and calculation of the cost–effectiveness ratio are repeatedly replicated (by convention, at least 1000 times23). Third, the repeated estimates of the ratio are ordered from lowest (best) to highest (worst). Many methods are available for identifying a 95% confidence interval from this rank-ordered distribution.22 The percentile method is one of the simplest to compute, but it may not be as dependably accurate as other, more complex, methods.24 When 1000 repeated estimates have been made, the percentile method uses the 26th and 975th ranked cost–effectiveness ratios to define the confidence interval. The nonparametric bootstrap method considers the correlation between the costs and effects observed in the sample in its confidence interval estimate. Additionally, it does not make parametric assumptions, such as normality or symmetry, on the distribution of the ratio.

Fieller theorem method. This method is an application of Fieller’s theorem,25 a parametric method for computing confidence intervals of a ratio based on the assumption that the numerator and denominator of the ratio follow a bivariate normal distribution. Willan and O’Brien13 and Chaudhary and Stearns15 have proposed applying this method to estimate confidence intervals for cost–effectiveness ratios. They also provide formulae for computing the confidence intervals. The estimated correlation between costs and effects is included in the computation and the Fieller theorem method does not restrict the distribution of the ratio to be normal or symmetric.

The Monte Carlo experiment

We used a Monte Carlo experiment to evaluate the performance of the four procedures.26 We started with a simulated population with known means and variances of costs and effects, known distributions of costs and effects, a known correlation between costs and effects and a known cost–effectiveness ratio. We then drew 500 samples of 250 for each treatment group from this population, which was equivalent to repeating a randomized trial 500 times. For each sample, we used each of the four methods to compute confidence intervals for the cost–effectiveness ratio observed...
in the sample. Finally, we recorded whether the confidence intervals covered the true ratio in the population from which the sample was drawn. When it did not cover the true ratio, we identified whether the lower limit of the interval exceeded the true ratio or whether the upper limit of the interval was below the true ratio.

We performed this experiment in 20 different populations, defined by two distributions of costs, two distributions of effects and five levels of correlation between costs and effects (2 × 2 × 5). In half the populations, costs in each treatment group were drawn from normal distributions; in the other half, costs were drawn from log normal distributions. The difference in cost between the two treatment groups in each population equalled $5000 and was the result of two distributions with fixed means of $20 000 and $25 000 and fixed standard deviations of $20 000.

For effects, we set a 10% absolute difference in mortality between the two groups (one treatment group had higher costs and lower mortality and the other had lower costs and higher mortality). To vary the distribution, in half of the populations we defined mortality rates of 25% versus 15% in the two groups; in the other half, we defined mortality rates of 55% versus 45%.

Finally, we varied the correlations of costs and effects between −0.50 and 0.50 using five points separated by four intervals (the size of each interval was 0.25). In distributions with negative correlations, persons who died cost more than those who survived (possibly because they were sicker to begin with or because they received heroic measures before they died). In distributions with positive correlations, persons who survived cost more than those who died (possibly because those who survived were in the hospital longer and thus incurred greater costs). In all 20 populations, the true cost–effectiveness ratio was $50 000 ($5000/0.10) per death averted.

Criteria for assessing the adequacy of a confidence interval

The principal criterion used to evaluate the performance of the four methods for computing confidence intervals was the overall probability of miscoverage (i.e. the percentage of samples in a Monte Carlo simulation where the true ratio fell outside the estimated confidence interval). The procedure with the best confidence interval was the one that came closest to the target miscoverage level of 5%. Symmetric miscoverage of the intervals was used as a secondary criterion for evaluating the four methods. Symmetric miscoverage occurred when the upper limit of the confidence interval fell below the true value with the same frequency with which the lower limit fell above the true value.

Analysis

We evaluated the effect of the different distributions of costs and effects and their correlations on the resulting distribution of cost-effectiveness ratios, as well as how these distributions may have affected the relative performance of the confidence intervals derived using the four methods.

Effects of population assumptions on the distribution of the cost–effectiveness ratios

We identified the effects on the distribution of the cost–effectiveness ratios of different distributions of costs and effects and the correlations between costs and effects by calculating the standard deviation, skewness and kurtosis of the observed distributions of the ratios resulting from the different costs and effects. These distributions of ratios are the result of the 500 repeated trials from each of the 20 populations; they are independent of the different methods used to compute confidence intervals.

The standard deviation is a measure of the spread of the distribution (larger standard deviations indicate wider distributions). Skewness is a measure of the symmetry of the distribution (i.e. whether or not one tail of the distribution is longer than the other). Skewness greater than zero indicates that the distribution of the ratio has a longer right tail; skewness less than zero indicates that the distribution has a longer left tail.

Kurtosis is a measure of peakedness of the distribution, which often is judged relative to the peakedness of a normal distribution. Kurtosis greater than three represents a distribution that is flatter than a normal distribution; kurtosis less than three represents a distribution that is more peaked than a normal distribution.

Coverage properties of the four methods

To compare the coverage properties of the four methods.
methods we evaluated their probabilities of miscoverage in the 20 populations (principal criterion).

To evaluate our secondary criterion, the degree of symmetry of miscoverage of the four methods, we calculated a fraction defined as the number of times the true ratio fell below the lower limit of the ratio's confidence interval (i.e. confidence intervals that overestimated the magnitude of the true ratio) divided by the total number of mistakes in coverage. A fraction equal to 0.5 represents symmetric miscoverage; fractions >0.5 represent confidence intervals that overestimated the magnitude of the true ratio; and fractions <0.5 represent confidence intervals that underestimated the magnitude of the true ratio.

RESULTS

Effects of population assumptions on the distribution of the cost-effectiveness ratios

Table 1 shows the average effects of the different distributions of costs and effects and their correlations on the standard deviation, skewness and kurtosis of the cost-effectiveness ratios resulting from the 500 repeated trials within each of the 20 populations. Differences in the distribution of effects substantially altered the distribution of the cost-effectiveness ratio. When the 10% absolute difference in mortality was the result of mortality rates of 55% versus 45% rather than of 25% versus 15%, the standard deviation increased, the distribution changed from one with more symmetric tails (skewness, 0.9) to one with a longer right tail (skewness, 4.2) and the distribution changed from one that was somewhat flatter than a normal distribution (kurtosis, 6.1) to one that was substantially flatter (kurtosis, 44.5). The reason for this finding is that due to the increased variance in the case of 55% versus 45% mortality rates, there are a greater number of estimates of effects that approach zero and thus more samples with extremely high cost-effectiveness ratios.

As the correlation between costs and effects changed from highly negative to highly positive, the standard deviation of the distribution of the cost-effectiveness ratio became smaller and the distribution changed from being right-tailed and substantially flatter than a normal distribution (skewness, 4.2; kurtosis, 42.1) to being left-tailed with a peak similar to that of a normal distribution (skewness, 0.4; kurtosis, 4.7).

Finally, both the skewness and kurtosis were modestly decreased when the cost distribution was changed from normal to log normal. However, the effect of this change on the distribution of the cost-effectiveness ratios was much smaller than the effect of differing distributions of effects and differing levels of correlation.

Figure 2 shows the two extreme distributions of the ratios that resulted from the different assumptions about costs and effects. The more peaked dashed line represents the distribution of the

Table 1. Effects of the population assumptions on distribution of cost-effectiveness ratios

<table>
<thead>
<tr>
<th>Population assumption</th>
<th>Standard deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of costs:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>27 111</td>
<td>2.7</td>
<td>29.9</td>
</tr>
<tr>
<td>Log normal</td>
<td>26 863</td>
<td>2.3</td>
<td>20.6</td>
</tr>
<tr>
<td>Distribution of effects:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-15%</td>
<td>22 049</td>
<td>0.9</td>
<td>6.1</td>
</tr>
<tr>
<td>55-45%</td>
<td>31 925</td>
<td>4.2</td>
<td>44.5</td>
</tr>
<tr>
<td>Correlation of costs and effects:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.50</td>
<td>36 547</td>
<td>4.2</td>
<td>42.1</td>
</tr>
<tr>
<td>-0.25</td>
<td>33 278</td>
<td>4.1</td>
<td>44.7</td>
</tr>
<tr>
<td>0.00</td>
<td>26 587</td>
<td>2.6</td>
<td>23.6</td>
</tr>
<tr>
<td>0.25</td>
<td>21 538</td>
<td>1.5</td>
<td>11.2</td>
</tr>
<tr>
<td>0.50</td>
<td>16 985</td>
<td>0.4</td>
<td>4.7</td>
</tr>
</tbody>
</table>

*Mean cost-effectiveness ratio equals $50,000 for each of the 20 populations.*
ratios for the population with log normally distributed costs, mortality rates of 25% versus 15% and a correlation between costs and effects of 0.50. The more right-tailed and flatter solid line represents the distribution for the population with normally distributed costs, mortality rates of 55% versus 45% and a correlation between costs and effects of -0.50. Although the distributions differ markedly in their spread, degree of symmetry and the height of their peak, both are peaked at $50,000 because that is the true ratio in the two populations from which they were drawn.

Coverage properties of the four methods

Table 2 reports the coverage probabilities for the four methods in the different populations. Because we found that the distributions of costs and effects had little impact on the probability of miscoverage while correlations between costs and effects had a greater impact, we report average miscoverage among the four populations for each of the five evaluated correlations. We also report the extreme miscoverage probabilities (i.e. highest and lowest) among the four populations included in the average to ensure that these averages do not hide large variations in miscoverage.

Inspection of Table 2 indicates that the results from the bootstrap and Fieller theorem methods were relatively unaffected by the level of correlation between costs and effects (range of miscoverage in the 20 populations, 4.0-6.8% for the bootstrap method and 3.8-6.2% for the Fieller method). For the box and Taylor methods, on the other hand, negative correlations led to undercoverage and positive correlations led to overcoverage. For example, when the correlation was -0.5, the box method did not cover an average of 12% of the time; when the correlation was 0.5, it did not cover 1.3% of the time. In addition, when no correlation was seen between costs and effects, all four methods had similar overall coverage properties.

When costs and effects were positively correlated, the bootstrap method tended to perform slightly better than the Fieller theorem method (i.e. the range of probabilities of miscoverage for the four populations was closer to 0.05). When the correlation of costs and effects was negative, the

Figure 2. Distribution of cost–effectiveness ratios (in thousands of dollars) among the 500 samples from two of the populations (defined by the distributions of costs and effects and the correlation between costs and effects). The solid line represents the distribution of cost–effectiveness ratios from the population with normally distributed costs, mortality rates of 35% compared with 45% and a -0.5 correlation between costs and effects. The standard deviation of the distribution of the ratios is 46,000; the skewness is 6.5, and the kurtosis is 72. The dashed line represents the distribution of the ratios from the population with normally distributed costs, mortality rates of 25% compared with 15% and a 0.5 correlation between costs and effects. The standard deviation of the distribution of the ratios is 15,000, the skewness is 0.1 and the kurtosis is 3.2.

Table 2. Miscoverage (%) of the four methods by correlation assumption between costs and effects^a

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Method</th>
<th>Box (95%)</th>
<th>Taylor (95%)</th>
<th>Bootstrap (95%)</th>
<th>Fieller (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.50</td>
<td>12.0</td>
<td>6.2</td>
<td>6.1</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11.0–12.4)</td>
<td>(5.2–7.0)</td>
<td>(5.4–6.8)</td>
<td>(5.0–6.0)</td>
<td></td>
</tr>
<tr>
<td>-0.25</td>
<td>8.8</td>
<td>5.7</td>
<td>5.7</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.8–10.2)</td>
<td>(3.6–7.0)</td>
<td>(4.6–6.8)</td>
<td>(4.4–6.2)</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>5.8</td>
<td>4.5</td>
<td>5.4</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.6–6.6)</td>
<td>(3.4–5.4)</td>
<td>(4.0–6.8)</td>
<td>(4.0–6.0)</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>3.1</td>
<td>3.4</td>
<td>5.1</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.8–3.4)</td>
<td>(3.0–4.0)</td>
<td>(4.6–6.0)</td>
<td>(4.2–5.6)</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>1.3</td>
<td>2.7</td>
<td>4.9</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.8–2.6)</td>
<td>(2.0–3.2)</td>
<td>(4.8–5.0)</td>
<td>(3.8–5.8)</td>
<td></td>
</tr>
</tbody>
</table>

^aTable entries represent the probability of miscoverage averaged across the four distributions of costs and effects; values in parentheses represent the largest and smallest of the miscoverage probabilities for the four distributions of costs and effects for each level of correlation. For example, for the bootstrap method given a correlation of 0.50, the probability of miscoverage in the four populations ranged from 4.8 to 5% and had a mean of 4.9%. The target level of miscoverage is 5%.

Our evaluation of the degree of symmetry of the miscoverage of the four methods indicated that for the box, bootstrap and Fieller theorem methods, 52% of the errors in coverage occurred because the lower limit of the interval exceeded the true ratio in the population; 48% of the errors occurred because the upper limit of the interval fell below the true ratio (i.e., errors were relatively symmetric). For the Taylor series method, only 2% of the errors in coverage occurred because the lower limit of the confidence interval exceeded the true ratio, whereas 98% of the errors in coverage occurred because the upper limit of the confidence interval fell below the true ratio. This finding implies that the upper limits of the confidence intervals derived using the Taylor series method were often too low.

**DISCUSSION**

Economic evaluations of new medical therapies are increasingly being performed in randomized clinical trials. Several methods have been proposed for calculating confidence intervals for the resulting cost-effectiveness ratios. In this study, we found that confidence intervals derived using the nonparametric bootstrap method and the Fieller theorem method were more dependably accurate than those constructed with either the box or Taylor series methods. In particular, the results of the bootstrap and Fieller theorem methods were more robust than the results of the box and Taylor series methods to variations in the distributions of costs and effects and to the correlation between costs and effects.

Our analysis suggests that there may be small differences in the accuracy of the results of the bootstrap and Fieller methods when the distributions of costs and effects and the correlations between costs and effects vary. For example, the bootstrap method may perform better when costs and effects are positively correlated. Further research is needed to characterize the differences in results of these methods under a broader set of conditions (e.g., more distributions, different sample sizes).

We also found that the Taylor series method had confidence intervals that asymmetrically underestimated the upper limit of the interval. This result is consistent with the conclusions in two recent comparisons of methods using an example cost-effectiveness ratio. Use of this method to evaluate which of two therapies should be adopted may yield results that are biased toward the adoption of the more expensive, more effective therapy, because its confidence intervals underestimate the probability that the true cost-effectiveness ratio will be substantially higher than that calculated from the study.

The box and Taylor series methods generally provide weak approximations of the 95% confidence intervals for the cost-effectiveness ratios. Nevertheless, when costs and effects in the underlying population were not correlated, we found that all four methods had reasonably similar overall coverage properties (although the Taylor series method still had problems with symmetry). In these cases, one might use the box method because of the ease of computation it provides. To allow readers to judge the likely reliability of confidence intervals computed with this method, authors should report the correlation between costs and effects in the sample. Our simulation indicated that this correlation is probably ade...
quate for identifying when the correlation in the population is zero.

Finally, the distributions of costs and effects did not seem to affect the relative performance of the four methods. The reason for this finding is that all four methods could appropriately incorporate the variance from normal distributions in their calculation of confidence intervals. Although the distribution of costs and effects among individual persons in the sample may not be normally distributed (for example, they may be log normal or binomial), the distribution of the means of the costs and effects—which are used to compute the cost-effectiveness ratio—are asymptotically normal.

One traditional method for computing sample sizes for trials that have both clinical and economic outcomes has been to compute separate sample size estimates for the clinical and cost outcomes and to use the larger of the two estimates. The resulting sample size is the same whether the costs and effects are positively or negatively correlated. Our finding that the correlation between costs and effects has a substantial effect on the distribution of the cost-effectiveness ratio (and thus the confidence interval for the ratio) indicates that this approach may need to be modified if the cost-effectiveness ratio itself is the endpoint of interest. When costs and effects are positively correlated, the resulting sample size may be sufficient. However, if costs and effects are negatively correlated, relatively larger samples are required to overcome the fact that the resulting distribution of the ratio will be wider and have a longer right tail. In this case, it may be possible to use the formula for the Fieller theorem method to estimate sample size.

Our study had several limitations. First, the relative performance of the methods may be affected by other variables that we did not evaluate. For example, we varied the distribution of costs but always assumed that the distribution was the same in the two groups. One effect of a new therapy may be to shift the distribution of costs as well as its mean and variance. A nother effect of a new therapy may be to alter the correlation between costs and effects. Nevertheless, given that the bootstrap and Fieller theorem methods appear to make fewer restrictive assumptions than do the other two methods, we see no reason they would provide worse estimates in these situations.

Second, we did not evaluate every proposed method for developing confidence intervals. Other methods that have been proposed include the bias-corrected bootstrap and jack-knife procedures. The former may have advantages over the percentile method bootstrap in cases when the distribution of the ratio is skewed. The latter, on the other hand, may have particular advantages if informed removal of observations can be exploited in the estimation of the confidence intervals.

Finally, we focused on measures of precision for outcomes for which there is an unambiguous tradeoff between costs and effects (i.e. the intervals all fall within quadrant I or quadrant III). The evaluation of precision may need to be redefined when separate confidence intervals for costs and effects indicate that the confidence intervals for the comparison of costs and effects fall in more than one quadrant. For example, if one therapy appears to dominate another, one might use the results of the bootstrap method to identify the probability that the therapy truly was dominant. Alternatively, one might use these results to identify the proportion of the time that one therapy dominated or had a ratio less than a prespecified target such as $50,000 per year of life saved compared with another therapy.

CONCLUSION

We found that confidence intervals for cost-effectiveness ratios resulting from the nonparametric bootstrap method and Fieller theorem method were more dependably accurate than those estimated using the Taylor series or box methods. Routine reporting of these intervals will allow individuals using cost-effectiveness ratios to make clinical and policy judgments to better identify when an intervention is a good value for its cost.

ACKNOWLEDGMENTS

The authors wish to thank Dr Sankey Williams, Dr Joseph Heyse and Dr John Cook for their helpful suggestions and reviews of earlier drafts of the manuscript. Grant support was provided in part by a grant from Pharmacia and Upjohn Inc., Kalamazoo, M.l. K.S. is also a recipient of a career development award in Pharmacoeconomics from the PHRMA foundation.
REFERENCES


