CONFIDENCE INTERVALS, ACCEPTABILITY AND NET HEALTH BENEFITS IN COST EFFECTIVENESS ANALYSIS

Henry Glick
University of Pennsylvania
hlthsrvs@mail.med.upenn.edu
www.med.upenn.edu/dgimhsr

Economic Assessments in Clinical Trials: Design and Analysis

Drug Information Association
London, England
February 16, 2001

HISTORY OF UNCERTAINTY RESEARCH

- Prior to early 1990's, clinical economists did not have an answer to questions about stochastic uncertainty related to cost effectiveness ratios
  - Could express this uncertainty for the numerators and denominators of the ratios separately
  - For the ratio, however, we usually said that we evaluated uncertainty using sensitivity analysis
    * Sensitivity analysis was not completely cut-off from notions of stochastic uncertainty: in some cases, we used 95% confidence limits to define the upper and lower bounds of the variables being evaluated in sensitivity analysis

- In the early 1990's, investigators like Joe Heyse, Mike Drummond, Bernie O'Brien, and Andy Willan launched a revolution in the thinking about uncertainty in cost effectiveness ratios
  - In 7 short years it has led to dramatic advances in the theory of measuring this uncertainty

HISTORY (2)

- The initial research evaluating uncertainty focused on calculating confidence intervals for cost effectiveness ratios
  - “Let 100 flowers bloom [and 100 schools of thought contend]” phase, where we were evaluating every conceivable method for calculating confidence intervals

- Difficulties with some of the most important methods began to become apparent
  - Ordering problems -- and thus potentially undefined intervals -- when bootstrap replicates were found in all 4 quadrants of Black's cost effectiveness plane
  - Undefined intervals and exclusion (rather than inclusion) intervals when using Fieller's theorem

- Stinnett and Mullahy proposed the net health benefit transformation, and researchers began to evaluate similarities and differences between using cost effectiveness ratios and net health benefits for economic evaluation
Cost-effectiveness results are generally interpreted on the cost-effectiveness plane.

There have been some differences of opinion about whether change in effects goes on the X or Y axis, and there have been some differences of opinion in numbering the quadrants.

But there has been a growing agreement -- at least among some of the researchers that have attended these meetings -- we may want to use a plane like the one above.

CURRENT STATE OF THE ART

- May be my idiosyncratic view:
  - Develop hypotheses about cost effectiveness ratios
    * E.g., the incremental ratio of therapy X compared with therapy Y will be lower than £Z per QALY (where Z represents one’s estimate of the acceptable upper limit for the confidence interval, which has been referred to by Briggs as the ceiling ratio)
    * E.g., when you multiply effects times the ceiling ratio and subtract out costs, is the result statistically significantly greater in the treatment group than it is in the control group

HOW DID WE GET HERE FROM THERE?

CONFIDENCE INTERVALS FOR COST EFFECTIVENESS RATIOS

- A number of methods have been proposed for estimating confidence intervals
  - Resampling procedures
    * Bootstrap (parametric or nonparametric, bias corrected, etc.)
    * Jack Knife
  - Direct estimation using Fieller theorem intervals
  - Ratio of separate confidence intervals computed for costs and effects (the “Box” Method)
  - Others....
NONPARAMETRIC BOOTSTRAP METHOD

- Resample from the study sample and compute cost effectiveness ratios in each of the multiple samples

1. Draw a sample of size \( n \) with replacement from the empiric distribution and use it to compute a cost effectiveness ratio

2. Repeat this sampling and calculation of the ratio (by convention, at least 1000 times for confidence intervals)

3. The repeated estimates of the ratio are ordered from lowest (in some sense best) to highest (in some sense worst)

4. Identify a 95% confidence interval from this rank-ordered distribution
   - The percentile method is one of the simplest to compute. When using 1000 repeated estimates, the percentile method uses the 26th and 975th ranked cost effectiveness ratios to define the confidence interval
   - Carpenter and Bithell (2000) suggest using the BCa method
   - Do not use bootstrap methods that assume one can estimate the ratio’s standard error

- Ordering bootstrap replicates – and defining a confidence interval is relatively uncomplicated when the bootstrap replicates fall in 1, 2, or even 3 quadrants of the cost effectiveness plane
Although in all 3 examples the bootstrap replicates fall in 3 quadrants, we have no problems ordering the replicates or identifying which portions of the distribution are cutoff by the confidence intervals

- For the top 2 figures, we cannot be 95% confident about which is the better therapy
- For the bottom one, we can be if the ceiling ratio is less than $615,000

Ordering is much more controversial when the replicates fall in all four quadrants

- Disagreements have arisen how one defines 95% CI for this example
- Questions have also been raised about the interpretation of intervals that have negative values (i.e., intervals that fall in the northwest and/or southeast quadrants)

**FIELLER’S THEOREM METHOD**

- Parametric method based on the assumption that the differences in costs and the differences in effects follow a bivariate normal distribution
- However, real number solutions may not exist in some samples with inadequate sample size or large coefficients of variation
FIELLER’S THEOREM FORMULA (I)

- Lower limit
  \[
  L1 = \frac{(X - [X^2 - YZ])}{Y}
  \]

- Upper limit
  \[
  L2 = \frac{(X + [X^2 - YZ])}{Y}
  \]

- Where:
  - \( X = \Delta E \Delta C - f_{v,1-a} r s_{\Delta E} s_{\Delta C} \)
  - \( Y = \Delta E^2 - f_{v,1-a} r s_{\Delta E}^2 \)
  - \( Z = \Delta C^2 - f_{v,1-a} r s_{\Delta C}^2 \)
  - \( \Delta E \) and \( \Delta C \) denote mean difference in effect and cost; \( s_{\Delta E}^2 \) and \( s_{\Delta C}^2 \) denote estimated variances; \( r \) equals estimated Pearson correlation coefficient between costs and effects; \( f_{v,1-a} \) is the upper percentage point of the F-distribution with 1 and \( v \) degrees of freedom, \( v \) being the number of degrees of freedom upon which the estimated variance \( \Delta E - \Delta C \) is based.


FIELLER’S THEOREM FORMULA (II)

- Interpretation of the Fieller theorem intervals is uncomplicated if there is a statistically significant difference in effect
  - Then and only then will the denominators be positive, and the interval is \( L1 \) to \( L2 \).
- If there is no significant difference in effect, the denominators will be negative
  - In this case, the term that was thought to be the “upper limit” will be a larger negative number (i.e., smaller) than the lower limit
  - The interval consists of the union of the intervals \((-\infty, L2)\) and \((L1, \infty)\), (i.e., an exclusion interval)
- If there is no significant difference in both cost and effect, then the quantity \([X^2 - YZ]\) may be negative, and because the square root of this quantity is taken, \( L1 \) and \( L2 \) may not be defined (the method cannot exclude any ratio, and thus is an inaccurate 95% CI).
COMPARISON OF METHODS
Polsky, Glick, Willke, Schulman

- Used a Monte Carlo experiment to evaluate the coverage properties of the confidence intervals computed using a number of different methods in 20 simulated populations 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
  - Costs: Normal and Log Normal
  - Effects: 25% vs 15% morality
    55% vs 45% mortality
  - Correlations: -0.5, -0.25, 0.0, 0.25, 0.5
  - In all 20 populations, the true cost effectiveness ratio was $50,000 ($5000/0.10) per death averted
- The procedure with the best confidence interval was the one that came closest to the 5% target level of miscoverage

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Box</th>
<th>Jackknife</th>
<th>Bootstrap</th>
<th>Fieller</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.50</td>
<td>12.0</td>
<td>7.6</td>
<td>6.1</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>(11.0-12.4)</td>
<td>(7.0-8.2)</td>
<td>(5.4-6.8)</td>
<td>(5.0-6.0)</td>
</tr>
<tr>
<td>-0.25</td>
<td>8.8</td>
<td>6.5</td>
<td>5.7</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>(7.8-10.2)</td>
<td>(5.8-7.0)</td>
<td>(4.6-6.8)</td>
<td>(4.4-6.2)</td>
</tr>
<tr>
<td>0.00</td>
<td>5.8</td>
<td>5.5</td>
<td>5.4</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>(4.6-6.6)</td>
<td>(4.8-6.2)</td>
<td>(4.0-6.8)</td>
<td>(4.0-6.0)</td>
</tr>
<tr>
<td>0.25</td>
<td>3.1</td>
<td>4.3</td>
<td>5.1</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>(2.8-3.4)</td>
<td>(3.6-4.7)</td>
<td>(4.6-6.0)</td>
<td>(4.2-5.6)</td>
</tr>
<tr>
<td>0.50</td>
<td>1.3</td>
<td>4.1</td>
<td>4.9</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>(0.8-2.6)</td>
<td>(3.4-5.0)</td>
<td>(4.8-5.0)</td>
<td>(3.8-5.8)</td>
</tr>
</tbody>
</table>

COMMENT ON POLSKY ET AL.

- We varied distribution of costs, distribution of effects, and correlations, because we thought they might affect the coverage properties of the different methods.

- We realized after the fact that our experiment was located firmly in the northeast quadrant, and we didn’t have many hard cases to deal with.

- As the fairly large body of literature that followed our article demonstrated, we did not capture all of the variables that may affect the coverage properties. Also need to consider:
  - Magnitude of effect difference and its CI
  - Number of observations in the sample

RESPONSES TO PROBLEMS WITH CONFIDENCE INTERVALS FOR COST EFFECTIVENESS RATIOS

- In response to the problems that arise in certain situations when computing confidence intervals for cost effectiveness ratios, researchers began exploring a number of alternative approaches for evaluating stochastic uncertainty for cost effectiveness analysis.
  - Acceptability criterion (van Hout and colleagues)
  - Net health benefits (Stinnett and Mullahy)
  - Angular transformation (Heyse and Cook)

- In what follows, I discuss the first 2 approaches.

- Evaluation of the acceptability criterion is based on the probability that the estimated ratio falls below a specified ceiling ratio.

- The acceptability criterion is defined on the cost effectiveness plane as a line passing through origin with a slope defined by the ceiling ratio. Points from the cost/effect distribution falling in quadrants ENE, SE, and SSW satisfy the criterion.

- The probability that the criterion is met can be calculated by counting bootstrap replicates or, when they are defined, from the Fieller theorem intervals.
We previously indicated that for this example it is difficult to identify the 95% CI around the point estimate of $42,974.

However, it is straightforward to calculate that 56.6% of the bootstrap replicates satisfy an acceptability criterion of $50,000 (i.e., the cost effectiveness result is not significantly below the ceiling ratio).

Given that there is no agreed upon ceiling ratio for use in hypothesis testing of cost effectiveness ratios, we routinely estimate the probability of acceptability for a number of ceiling ratios (by varying the slope of the line defining the acceptability criterion).
• In this example, the ratio of $50,000 per QALY saved has the equivalent of a one-tailed p-value less than 0.05
  - i.e., 4% of the distribution falls above the $50,000 line)
• The ratio of $60,000 has the equivalent of a two-tailed p-value less than 0.05
  - i.e., less than 2.5% of the distribution falls above the $60,000 line)
The curve cuts the vertical axis (A) at the P-value (one sided) for the cost difference.

The 50% point (B) corresponds to the point estimate of cost-effectiveness.

The curve is tending to 1 minus the P-value (one-sided) for the effect difference.

D defines the (1-2X)% confidence interval.

NET HEALTH BENEFITS

- A composite measure (part cost-effectiveness, part cost benefit analysis), usually expressed in dollar terms, which is derived by rearranging the following decision rule:

\[
\text{CR} > \frac{(\text{Costs}_1 - \text{Costs}_2)}{(\text{Outcomes}_1 - \text{Outcomes}_2)}
\]

where CR = ceiling ratio (e.g., £40,000)

- More commonly expressed as what may be called net monetary benefits

\[
(CR \times [(\text{Outcomes}_1 - \text{Outcomes}_2)] - (\text{Costs}_1 - \text{Costs}_2)) > 0
\]

- Less commonly expressed as what is called net health benefits

\[
(\frac{(\text{Costs}_1 - \text{Costs}_2)}{\text{CR}}) - (\text{Outcomes}_1 - \text{Outcomes}_2) > 0
\]

- All else equal, one should adopt programs with net monetary (health) benefits that are greater than 0 (i.e., programs with incremental cost effectiveness ratios that are less than the ceiling ratio).

- Differs from cost effectiveness ratios in that -- except for the ceiling ratio -- a single ratio (made up of different costs and effects) does not represent a single net monetary (health) benefit.
NET MONETARY BENEFITS

- Motivation for estimating net monetary benefits:
  - Overcomes problems associated with parametric tests of the ratio
  - Can provide direct statistical tests with patient-level data

Plot of Cost Effectiveness Ratios in 4 Quadrants

NET MONETARY BENEFIT GRAPH

- Net monetary benefits are defined by a family of lines, all with the slope of the ceiling ratio

NET MONETARY BENEFIT GRAPH (II)

- The value of the net monetary benefits represented by a particular net monetary benefit line = -intercept
  - The net monetary benefit line running through the origin has net monetary benefits of 0, and is equivalent to the acceptability criterion defined using the same ceiling ratio
  - Lines below and to the right of the NMB line = 0 have positive net monetary benefits (i.e., acceptable cost effectiveness ratios)
WHERE IN THE ANALYTIC PROCESS SHOULD THE NET MONETARY BENEFIT TRANSFORMATION BE MADE?

- There are at least two places in the analytic process where the transformation from costs and effects to net monetary benefits can be made
  - Referred to here as aggregate and per-patient assessments
- Aggregate net monetary benefits
  - The transformation is made after the calculation of mean difference in costs and effects
    \[(\Delta \text{Effect} \times \text{CR}) - \Delta \text{Cost}\]
- Steps:
  1. Compute incremental effects (e.g., with regression analysis); multiply by the ceiling ratio
  2. Compute incremental costs (e.g., with regression analysis)
  3. Assess whether the difference between the incremental value of the outcome minus the incremental costs exceeds 0
  4. CI may be developed using bootstrap method

ADVANTAGES OF AGGREGATE ASSESSMENT OF NET MONETARY BENEFITS

- For net monetary benefits in general: The study result is a difference in means of net monetary benefits, not a ratio of means, and is always defined (i.e., no odd statistical properties like the ratio)
- For the aggregate approach: May be better able to estimate difference in outcomes and costs separately than to estimate a difference in the composite measure of net monetary benefits
  - Differences in costs and differences in effects analyzed separately may be more predictable than differences in net monetary benefits themselves
  - May need to adopt methods for analyzing censored data, in which case one estimates costs and effects, not net monetary benefits
PER-PATIENT ASSESSMENT OF NET MONETARY BENEFITS

- The aggregate approach to assessing net monetary benefits does not take full advantage of one's ability to calculate net monetary benefits using patient level data.

- In per-patient assessment, the transformation is made prior to analysis, and the analysis directly assesses differences in net monetary benefits.

- Derivation (given test that NMB>0):

\[
0 < (\sum (\text{Effect} \times CR) - \sum \text{Cost}) - \sum (\text{Cost} \times \text{Effect}) - (\sum \text{Cost} \times \text{Effect} - \sum \text{Cost} \times \text{Effect})
\]

where \( E_{Ai} \) is the effect for patient I among those who receive active intervention; \( N_A \) is the number of patients who receive active intervention; \( E_{ui} \) is the effect for patient I among those who receive usual care; \( N_U \) is the number of patients who receive usual care; \( C_A \) is the cost for patient I among those who receive active intervention; and \( C_U \) is the cost for patient I among those who receive usual care.

ADVANTAGES OF PER-PERSON ASSESSMENT OF NET MONETARY BENEFITS

- As with aggregate assessment, the study result is a difference in means of net monetary benefits, not a ratio of means, and is always defined (i.e., no odd statistical properties like the ratio).

- Each subject contributes a value to the average of net monetary benefits (i.e., in the same way that they contribute to the mean difference in costs or the mean difference in QALYs).

- Hypotheses about net monetary benefits can be directly evaluated statistically by the use of a test of differences in means of net monetary benefits (with CI).
  - e.g., t-tests or ordinary least squares regression.

- The distributional properties of net monetary benefits are much less complex than those of the ratio.

- There has been little if any research into the relative merits of the aggregate and per-person approaches.
  - May depend in part on the relative multivariable predictability of effects, costs and net monetary benefits.
The net monetary benefit line cuts the vertical axis (A) at $-\triangle C$

The upper limit of the CI of the net benefits curve cuts the horizontal axis (B) at the lower limit of the CI of the cost effectiveness ratio

The net benefits line cuts the horizontal axis (C) at the point estimate of the cost effectiveness ratio

The lower limit of the CI of the net benefits curve cuts the horizontal axis (D) at the upper limit of the CI of the cost effectiveness ratio
SUMMARY

- In the last 7 years, we as a community have made dramatic advances in how we address stochastic uncertainty in cost effectiveness analysis.

- In my view, the current state of the art is that we develop hypotheses about cost effectiveness ratios, but test these hypotheses by determining whether the net monetary benefits calculated using a ceiling ratio are significantly greater than 0.

- While in some sense, hypothesis tests about cost effectiveness ratios and net monetary benefits contain the same information, there are some benefits from the use of net monetary benefits that do not accrue to the use of cost effectiveness ratios. For example:
  - Net monetary benefits are always defined (i.e., no odd statistical properties like the ratio).
  - In some cases, hypotheses about net monetary benefits can be directly evaluated statistically by the use of a test of differences in means of net monetary benefits (with CI).

- Need to begin to evaluate the relative merits of the aggregate and per-person approaches to the calculation of net monetary benefits.