Sampling Uncertainty for Cost-Effectiveness Analysis in Clinical Trials

Henry Glick
Institutt for helseledelse og helseøkonomi
November 25, 2008

Good Value for the Cost

• A common goal of an economic analysis is to identify when we can be confident that one therapy is good value compared to another
• A threat to such confidence arises because the economic result observed in an experiment may not truly reflect the result in the population
  – Single sample drawn from a population
• This form of uncertainty is referred to as sampling (or stochastic) uncertainty
  – A commonly used approach for addressing this threat is to use the data from the experiment to identify when we can be confident about the value for the cost

Outline

• Describe methods for identifying when we can and cannot be confident about a therapy’s value for the cost
  – Point estimates
  – Confidence intervals
  – Decision threshold
• Goal is to demonstrate the quantification and interpretation of sampling certainty by use of CI for CER, CI for NMB, and acceptability curve
Sampling Uncertainty and Clinical Outcomes

- We can be confident that a therapy is clinically effective when its confidence interval excludes our decision threshold; we can’t be confident when its interval includes our decision threshold.
- For odds ratios / relative risks. Decision threshold = ???
  - OR = 0.30; 95% CI, 0.15 to 0.63
  - OR = 0.30; 95% CI, 0.09 to 1.02
- For risk differences or changes in blood pressure or cholesterol. Decision threshold = ???
  - Risk difference = 30%; 95% CI, 18% to 42%
  - Risk difference = 30%, 95% CI, -4% to 64%

Sampling Uncertainty and Clinical Outcomes (2)

1) Not confident A differs from B (A - B)

2) Confident A greater than B (A - B)

3) Confident A less than B (A - B)

Implications

- If the confidence interval includes the decision threshold, we CANNOT be confident that the alternatives differ from one another.
- If the confidence interval excludes the decision threshold, we CAN be confident that the alternatives differ from one another.
- It doesn’t matter what else is included or excluded from the interval.
Sampling Uncertainty and Economic Outcomes

- Confidence statements about economic outcomes are also based on whether or not the confidence interval for the economic outcome includes the decision threshold
- Methods for assessing confidence
  - Confidence intervals for cost-effectiveness ratios
  - Confidence intervals for net monetary benefits
  - Acceptability curve
- Decision threshold
  - FOR CER: Maximum willingness to pay (W) for a unit of health outcome or maximum acceptable cost-effectiveness ratio (e.g., 30,000 GBP or 50-100,000 USD)

Sampling Uncertainty and CI for CER

- Suppose we calculate a point estimate and 95% CI for the incremental cost-effectiveness ratio
- We determine whether we can be 95% confident a therapy is good value by comparing the confidence interval to our decision threshold (maximum WTP)
  - If maximum willingness to pay is included within the confidence interval, we CANNOT be confident that the two therapies differ in their cost-effectiveness
  - If it is excluded from / outside the interval, we CAN be 95% confident that one of the therapies is cost-effective compared to the other

Sampling Uncertainty and NMB

- Suppose we calculate a point estimate for NMB and its 95% CI
  - We determine whether we can be 95% confident that therapy A is good value by comparing the confidence interval to the NMB decision threshold (0)
    - Includes 0, cannot be confident of a difference
    - Excludes 0, can be confident the therapies differ
  - Primary difference between interpretation of CI for CER and CI for NMB is that we compare CER to WTP whereas we build WTP into NMB
Sampling Uncertainty and the Acceptability Curve

- Suppose we calculate a point estimate for fraction of the distribution that is acceptable
  - We determine whether we can be 95% confident that therapy A is good value by comparing the fraction acceptable to the decision threshold (horizontal lines drawn at 0.025 and 0.975
  - If $0.025 < \text{fraction} < 0.975$, we cannot be 95% confident that the two therapies differ in their cost-effectiveness
  - If $\text{fraction} < 0.025 \text{ OR fraction} > 0.975$, we can be 95% confident that one of the therapies is cost-effective compared to the other

Differences in Clinical and Economic SU (1)

- While there are similarities between how we interpret sampling uncertainty for clinical and economic outcomes, there are also at least 2 differences
  - For differences in height, weight, or risk, there is little debate that the decision threshold is 0; for odds ratios and relative risks, there is little debate that the decision threshold is 1.0
  - For economic outcomes, jurisdictions exist in which there is much less agreement about the appropriate willingness to pay for a unit of health outcome
    - It is expected that the maximum willingness to pay can differ among decision makers, particularly in different decision making jurisdictions

Differences in Clinical and Economic SU (2)

- While it sometimes is stated that 95% confidence is arbitrary for judging clinical outcomes, there is near universal agreement among regulatory bodies and medical journals that this level of confidence is required for making a claim that two therapies differ clinically
  - Such agreement is less clear in the economics community, in which there is discussion about whether we need to have the same degree of confidence about health returns on our investments as we do about clinical outcomes
  - Our conclusions are independent of the level of confidence we are seeking (e.g., 95% confidence or even 1% confidence)
Conclusions (1)

• For any given willingness to pay, an experiment **ALWAYS** allows us to draw one of three conclusions:
  – We can be confident that one therapy is good value compared to the alternative
  – We can be confident that the alternative therapy is good value compared to the first
  – We cannot be confident that the two therapies differ in their economic value

Conclusions (2)

• If our goal is to identify which of these 3 statements holds for a given willingness to pay, confidence intervals for cost-effectiveness ratios, confidence intervals for NMB, and acceptability curves **ALWAYS** provide the same answer:
  – e.g., if our WTP is included within the CI for the CER, then:
    • The CI for the NMB that is calculated by use of our WTP will include 0, and
    • The fraction of the distribution that is acceptable at our WTP will fall between the horizontal lines that define the decision threshold (e.g., between 0.025 and 0.975)
Conclusions (3)

• Confidence intervals for cost-effectiveness ratios provide decision makers with concise information (i.e., 0, 1, or 2 numbers) that allows them to determine — based on their own WTP — if they can be confident about a therapy’s value.

• Acceptability curves provide the added advantage of allowing decision makers to assess alternate levels of confidence if such alternate levels are of interest.

Experiment 1

• Suppose you conducted an economic evaluation of two therapies and found that:
  – Therapy A on average cost 1000 more than therapy B, SE = 325, p=0.002.
  – Therapy A on average yielded 0.01 QALYs more than therapy B, SE = 0.001925, p<0.0000.
  – The correlation between the difference in cost and effect was -0.71; and there were 250 participants per group in the trial.

• Point estimate CER:
  \[
  \frac{1000}{0.01} = 100,000 / \text{QALY saved}
  \]

CI for CER

• Lines through the origin that each exclude \(\alpha/2\) of the distribution of the difference in costs and effects.
The interval stretches from the lower (clockwise) limit to the upper (counter-clockwise) limit.

\[ \Delta C = 1000; SE_C = 325; \Delta Q = 0.01; SE_Q = 0.001925; \rho = -0.71; DOF = 498 \]

Interpreting the CI for the CER

1) Not confident A differs from B

2) Confident A cost-effective compared to B

3) Confident B cost-effective compared to A
NMB Recap

\[ \text{NMB} = (W \Delta Q) - \Delta C \]

- For a WTP of 50,000, NMB for experiment 1:
  \[ (50,000 \times 0.01) - 1000 = -500 \]
- The study result is a difference in means of net benefits, not a ratio of means, and is always defined (i.e., no odd statistical properties like the ratio) and continuous
- Unlike the cost-effectiveness ratio, the standard error of net benefits is always defined
- Given that not all decision making bodies have an agreed upon maximum willingness to pay, we routinely estimate net benefit over the range of policy relevant values of willingness to pay

NMB on the Cost-Effectiveness Plane

- Net benefit is defined on the cost-effectiveness plane by a family of lines, all with a slope equal to \( W \)
- Each line represents a single value of net benefit, which for NMB equals -intercept
- For the line passing through the origin, NMB = 0
  - Lines below and to the right of the net benefit=0 line have positive net benefits (i.e., acceptable cost-effectiveness ratios)
  - Lines above and to the left have negative net benefits
- Lines increase in value as one travels southeasterly

Constructing CI for NMB for Experiment 1

\[ \text{WTP: 50,000; NMB: -500; 95\% CI, -1284 to 284} \]
Interpreting CI for NMB

1) \( \lambda = 50,000 \): Not confident A differs from B (A - B)

<table>
<thead>
<tr>
<th>Less</th>
<th>0</th>
<th>Greater</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1283</td>
<td>0</td>
<td>283</td>
</tr>
</tbody>
</table>

2) \( \lambda = 250,000 \): Confident A net beneficial compared to B

<table>
<thead>
<tr>
<th>Less</th>
<th>0</th>
<th>Greater</th>
</tr>
</thead>
<tbody>
<tr>
<td>-30</td>
<td>0</td>
<td>2970</td>
</tr>
</tbody>
</table>

3) \( \lambda = 10,000 \): Confident B net beneficial compared to A

<table>
<thead>
<tr>
<th>Less</th>
<th>0</th>
<th>Greater</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1283</td>
<td>0</td>
<td>236</td>
</tr>
</tbody>
</table>

Additional Information in the NMB Graph

- Slope = Change in effect
Additional Information in the NMB Graph (II)

(A) NMB curve cuts the vertical axis at -C, because at that point W equals 0, \(Qx0=0\), and the NMB formula reduces to -C
   - NMB confidence limit curves cut the vertical axis at -1 times the confidence interval for C for same reason
(B) NMB curve cuts the horizontal axis at the point estimate for the cost-effectiveness ratio because when \(W = C/Q\), the NMB equation reduces to C – C
   The slope of the NMB curve equals Q because the NMB equation defines a line with slope Q and intercept –C
(C & D) The NMB confidence limit curves intersect the horizontal axis at the lower and upper limits of the confidence interval for the cost-effectiveness ratio

Acceptability Curve

- We calculate the probability a therapy is acceptable by calculating the probability that it falls below a specified value of WTP (e.g., the maximum WTP)
- The acceptability criterion is defined on the cost-effectiveness plane as a line passing through the origin with a slope equal to WTP
- The proportion of the distribution of the difference in cost and effect that falls below and to the right of this line is "acceptable" (i.e., has positive NMB); the proportion that is above and to the left of this line is "unacceptable"
Constructing the Acceptability Curve

The Acceptability Curve

Additional Information in Acceptability Curve
**Additional Information in Acceptability Curve (II)**

(A) The curve cuts the vertical axis at the p value (one sided) for $\Delta C$ because when $W=0$, $\Delta Q$ is ignored

- Fraction of the distribution of $\Delta C$ that falls above and below the X axis (0) defines a one-sided test of the difference in cost

(B) The 50% point on the curve corresponds to the point estimate of the cost-effectiveness ratio

(C) As $W$ approaches $\infty$, the curve asymptotically approaches 1 minus the p value (one-sided) for $\Delta Q$

(D) Cutting $\alpha\%$ from either end of the vertical axis and mapping the points on the curve on to the horizontal WTP defines the $(1-2\alpha\%)$ confidence interval for the cost-effectiveness ratio

---

**Review of Results for Experiment 1**

Confidence interval for CER

Lower (clockwise) limit = 28,200; upper (counter-clockwise) = 245,200

Confidence frontier for NMB

CI intersect decision threshold (0): Lower limit intersects at 28,200; upper limit at 245,200

Acceptability curve

Intersection with 0.025 and 0.975: Intersects 0.025 at 28,200 and 0.975 at 245,200

---

**Bottom Line**

- If the goal is to provide information for decision makers about whether or not we can be confident that a therapy is cost-effective, all three methods provide the same information
  - i.e., for any given WPT, all 3 yield the same recommendation

- The difference between the methods is in the way the information is expressed

- The choice of method should be based on the most effective way to express the point estimate and the uncertainty to decision makers rather than basing the choice on the statistical properties of each estimator
"Pattern 1" Findings

- We refer to findings like those in experiment 1 as pattern 1 findings.
- They occur when the difference in effect is significant.
- We know we are observing a pattern 1 finding when:
  - The confidence interval for the cost-effectiveness ratio excludes the Y axis (i.e., LL < PE < UL).
  - Both NMB confidence limits curves intersect the decision threshold (0) once.
  - The acceptability curve intersects horizontal lines drawn at both 0.025 and 0.975 on the Y axis.

Pattern 1 Findings (2)

PATTERN #1

<table>
<thead>
<tr>
<th>One can be confident the</th>
<th>One cannot be confident the two therapies differ from one another</th>
<th>One can be confident the more effective therapy is good value</th>
</tr>
</thead>
<tbody>
<tr>
<td>more effective therapy is not good value</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* In cases where some of the boundaries between the regions occur at negative willingnesses to pay, we may not always observe all 3 regions on an acceptability curve or NMB plot.

Experiment 3

- Suppose you conducted an economic evaluation of two therapies and found that:
  - Therapy A on average cost 400 more than therapy B, SE = 325, p=0.22
  - Therapy A on average yielded 0.02 QALYs more than therapy B, SE = 0.02, p<0.32
  - The correlation between the difference in cost and effect was 0.25; and there were 250 participants per group in the trial
- Point estimate CER:
  \[
  \frac{400}{0.02} = \frac{20,000}{\text{QALY saved}}
  \]
\[ \Delta C = 400, \ SE_C = 325 \ (p = 0.22); \ \Delta Q = 0.02, \ SE_Q = 0.02 \ (p = 0.32); \ p = .25; \ DOF = 498 \]

Last definable limit, Experiment 3

Last definable limit excludes 7.76%

\[ LL = UL = -11,500 \]

CI for NMB and Acceptability Curve
Review of Results for Experiment 3

- Confidence interval for CER
  CER CI: Undefined

- Confidence frontier for NMB
  CI intersect decision threshold (0):
  Does not intersect

- Acceptability curve
  Intersection with 0.025 and 0.975:
  Does not intersect

Pattern 3 Findings

- We refer to findings like those in experiment 3 as pattern 3 findings.
- They occur only when the difference in effect is not significant.
- We know we are observing a pattern 3 finding when:
  - The confidence interval for the CER is undefined.
  - Neither NMB confidence limit curve intersects the decision threshold (0).
  - The acceptability curve never intersects horizontal lines drawn at either 0.025 and 0.975 on the Y axis.

Pattern 3 Findings (2)

PATTERN #3

One cannot be confident the two therapies differ from one another.

-∞ ← Willingness to Pay → ∞
Estimate the CI for Our Maximum WTP

Dependable Accuracy?

- If the confidence interval is undefined (i.e., includes every possible ratio), how can 95% of experiments “cover” the true ratio and 5% fail to cover?

- Heitjan has shown that even in the case where Fieller’s theorem CI are undefined, a coverage experiment will demonstrate that intervals are dependably accurate
  - In the coverage experiment, \((\alpha \times N)\) of the repeated experiments will have defined Fieller theorem intervals that will fail to cover truth; a number of additional repeated experiments will have defined Fieller theorem intervals that will cover truth; and the remaining experiments will be undefined (i.e., cover truth by definition)

Experiment 2

- Suppose you conducted a set of experiments all with the same means and SDs for cost and QALYs, with the only difference being in their sample sizes
  - Therapy A on average cost 35 more than therapy B, SD for cost = 8692.7143 per group
  - Therapy A on average yielded 0.04 QALYs more than therapy B, SD for QALYs = 0.25043961 per group
  - The correlation between the difference in cost and effect was 0.706
- Point estimate CER:
  \[
  \frac{35}{0.04} = \frac{875}{QALY\ saved}
  \]
Experiment 2, N = 1,000 / Group

- SE cost = 388.75; SE QALYs = 0.0112

Same Experiment, But N = 500 / Group

- SE cost = 549.78; SE QALYs = 0.015839

Same Experiment, But N = 250 / Group

- SE cost = 777.5; SE QALYs = 0.0224
What Just Happened?

<table>
<thead>
<tr>
<th>N / Group</th>
<th>Point Est</th>
<th>P, QALYs</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>875</td>
<td>0.0004</td>
<td>-34,100 to 15,300</td>
</tr>
<tr>
<td>500</td>
<td>875</td>
<td>0.01</td>
<td>-91,00 to 20,300</td>
</tr>
<tr>
<td>250</td>
<td>875</td>
<td>0.07</td>
<td>LL, 245,200 to ∞</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UL, - ∞ to 28,200</td>
</tr>
</tbody>
</table>

- As the sample size shrinks, the limits grow wider
- When the limits grow wide enough to include the Y axis, (-∞/∞) they develop what some consider "odd" properties
  - Occurs when p-value for effectiveness > 0.05

"Odd" Properties

- Lower limit represents a larger number than the upper limit
- Point estimate either a larger number than both the lower and upper limits OR a smaller number than both limits
- WTP excluded from the confidence interval include the ratios with values between the upper and lower limits
  - In the example where N = 250 / group, between 28,200 and -245,200

How is it Possible for the Lower Limit to be a Larger Number than the Upper Limit?
How is it Possible for the Point Estimate to be Larger or Smaller than Both Limits?

Why Does the Confidence Interval Have a Gap Between 28,200 and 245,200

Interpreting CI for CER for Experiment 2 (N=250/g)

- What confidence statements can we make about this experiment?
  - So long as our WTP is between 28,200 and 245,200, we can be confident that the therapy is good value
Same Confidence statements from NMB and Acceptability Curves

Review of Results for Experiment 2

Confidence interval for CER
Lower (clockwise) limit = 245,200; upper (counter-clockwise) = 28,200

Confidence frontier for NMB
CI intersect decision threshold (0):
Lower limit intersects twice at 28,200 and 245,200

Acceptability curve
Intersection with 0.025 and 0.975:
Intersects 0.975 twice at 28,200 and 245,200

Pattern 2 Findings

• We refer to findings like those in experiment 2 as pattern 2 findings
• They occur only when the difference in effect is not significant
• We know are observing a pattern 2 finding when:
  – The confidence interval for the CER includes the Y axis (i.e., LL > UL > PE OR PE > LL > UL)
  – One NMB confidence limit curve intersects the decision threshold (0) twice; the other limit never intersects the decision threshold
  – The acceptability curve intersects a horizontal line drawn at either 0.025 and 0.975 on the Y axis twice and never intersects the other line
Pattern 2 Findings (2)

PATTERN #2

One cannot be confident the two therapies differ from one another

One can be confident that one of the therapies is good value

Willingness to Pay

Two Lines Through Origin, 3 Different Experiments

Experiment 1
CER CI: 28,200 to 245,200

Experiment 2(a)
CER CI: 245,200 to 28,200
$\Delta C = 35; \Delta Q = 0.04$

Experiment 2(b)
CER CI: 245,200 to 28,200
$\Delta C = 1985; \Delta Q = 0.0001$

Two Lines, 3 Different Experiments (II)

- 2 of the CI must include the Y axis, 1 must not
- In 2 experiments $\Delta Q$ is not significant, whereas in 1 it is
- 1 CI includes ratios between 28,200 and 245,200 and excludes all else; 2 CI exclude ratios between 28,200 and 245,200 and include all else
- In at least one of the experiments $\Delta C$ must be significant; in at least one one must not be
- W must fall within the interval defined for at least 1 of the experiments and fall outside the interval defined for at least 1 of the experiments
- Which of the experiments is which?
Overview of Patterns of Results

PATTERN 1
One cannot be confident the more effective therapy is not good value
One cannot be confident the two therapies differ from one another
One can be confident the more effective therapy is good value

PATTERN 2
One cannot be confident the two therapies differ from one another
One can be confident that one of the therapies is good value
One cannot be confident the two therapies differ from one another

PATTERN 3
One cannot be confident the two therapies differ from one another

Overview of Patterns of Results (II)

- The observed pattern is a function of ΔC and ΔE, their SEs and correlation, the DOF, and the confidence level
  - All experiments will display all 3 patterns depending on the confidence level
    - Experiment 1: pattern 1, $t < 5.184$; pattern 2, $5.184 \leq t \leq 7.571$; pattern 3, $t > 7.571$
  - The observed pattern is independent of method used to express sampling uncertainty
  - When calculating NMB and acceptability curves, one may not recognize these patterns
    - Patterns are defined over $W$ values that range from $-\infty$ to $\infty$, but NMB and acceptability curves usually calculated for positive values of $W$ only
Adoption Recommendations: No Consideration of Sampling Uncertainty

- Required to calculate CER/NMB for one set of outcomes:
  - \( C_a \) significantly greater than \( C_b \) and \( Q_a \) significantly greater than \( Q_b \)
- Not required to compare cost and effect if:
  - \( C_a \) significantly smaller than \( C_b \) and \( Q_a \) significantly greater than \( Q_b \) (dominance)
  - \( Q_a \) not significantly different from \( Q_b \)
  - \( C_a \) not significantly different from \( C_b \)
  - Therapies fail to differ significantly in both their cost and effect

Adoption Recommendations: Considering Sampling Uncertainty

- Required to calculate CER/NMB for one set of outcomes:
  - \( C_a \) significantly greater than \( C_b \) and \( Q_a \) significantly greater than \( Q_b \)
  - \( Q_a \) not significantly different from \( Q_b \)
  - \( C_a \) not significantly different from \( C_b \)
  - Therapies fail to differ significantly in both their cost and effect
- Not required to compare cost and effect if:
  - \( C_a \) significantly smaller than \( C_b \) and \( Q_a \) significantly greater than \( Q_b \) (dominance)

**C_a Significantly Different from C_b and Q_a Not Significantly Different from Q_b**

\[ \Delta C = -1250; \ SD = 500; \ p = 0.01 \]
\[ \Delta Q = 0; \ SD = 0.025; \ p = 1.0 \]
\[ SD \ for \ effectiveness = 0.280 \]
\[ SE = 0.5; \ DOF = 498 \]
Ca Not Significantly Different from Cb and Qa
Significantly Different from Qb

Conclusions

• For any given willingness to pay, an experiment ALWAYS allows us to draw one of three conclusions:
  – We can be confident that one therapy is good value compared to the alternative
  – We can be confident that the alternative therapy is good value compared to the first
  – We cannot be confident that the two therapies differ in their economic value

• If our goal is to identify which of these 3 statements holds for a given willingness to pay, confidence intervals for cost-effectiveness ratios, confidence intervals for NMB, and acceptability curves ALWAYS provide the same answer

Conclusions (2)

• Confidence intervals for cost-effectiveness ratios provide decision makers with concise information (i.e., 0, 1, or 2 numbers) that allows them to determine – based on their own WTP – if they can be confident about a therapy’s value

• Acceptability curves provide the added advantage of allowing decision makers to assess alternate levels of confidence if such alternative levels are of interest