Sampling Uncertainty: What We Are Doing Right and Where We Are Going Wrong

Henry A. Glick, Ph.D.
ISPOR Western Pharmacoeconomics Conference 2009
March 20, 2009
www.uphs.upenn.edu/dgimhsr

CATEGORY
Economists’ Answers to Epidemiologists’ Questions

“No, we perform sensitivity analysis”

Jeopardy Question
Don’t you calculate confidence intervals or p-values for cost-effectiveness ratios?
"No, We Perform Sensitivity Analysis"

- "No, we perform sensitivity analysis" was the only answer economists had for the epidemiologists' question until 1994
- In that year, O'Brien et al. published “In search of power and significance: issues in the design and analysis of stochastic cost effectiveness studies in health care” (Med Care. 1994;32:150-63)
- The last 15 years have seen dramatic advances in the theory of measuring this uncertainty
  - But assessment of sampling uncertainty still has not reached a mature stage where everyone understands it and where it routinely is correctly reported in published cost-effectiveness analyses

Outline

- Describe similarities and differences between confidence statements for clinical outcomes and confidence statements for cost-effectiveness analysis
- Illustrate construction and interpretation of confidence intervals for cost-effectiveness ratios, confidence intervals for net monetary benefit and acceptability curves
- Illustrate hard cases
- Discuss some of the ways we are going wrong in addressing sampling uncertainty for cost-effectiveness analysis

Section 1

Describe Similarities And Differences Between Confidence Statements For Clinical Outcomes And Confidence Statements For Cost-effectiveness Analysis

- Review confidence statements for clinical outcomes
- Describe the similarities between confidence statements for clinical and economic outcomes
- Discuss 2 potential differences between confidence statements for clinical and economic outcomes
- Summarize general properties of confidence statements for economic outcomes
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 results from experiments are derived from single samples and thus may not truly reflect the result in the population.
- This form of uncertainty is referred to as sampling (or stochastic) uncertainty.
- Methods for estimating sampling uncertainty for economic results have much in common with methods for estimating sampling uncertainty for clinical results, but there are also differences.

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.
- Which of these odds ratios allow us to be 95% confident about clinical effectiveness? Why?
  - OR = 0.30; 95% CI, 0.09 to 1.02
  - OR = 0.30; 95% CI, 0.15 to 0.63
- Which of these risk differences allow us to be 95% confident about clinical effectiveness? Why?
  - Risk difference = 30%; 95% CI, 18% to 42%
  - Risk difference = 30%, 95% CI, -4% to 64%

Confidence Statements for Clinical Outcomes

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 express sampling uncertainty by use of confidence intervals for cost-effectiveness ratios
• We estimate the point estimate and 95% CI for the ratio and determine whether we can be 95% confident that a therapy is good value by comparing the confidence interval to our maximum willingness to pay
  – 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
    • If the interval includes 0, cannot be confident of a difference
    • If the interval 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 the definition of 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 < fraction < 0.975, we cannot be 95% confident that the two therapies differ in their cost-effectiveness
    • If fraction < 0.025 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 odds ratios and relative risks, there is little debate that the decision threshold is 1; for differences in height, weight, or risk, there is little debate that the decision threshold is 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

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.

While my conclusions depend on the idea of making confidence statements, they are independent of the level of confidence we are seeking (i.e., they refer equally to 1% or 95% confidence statements).

Confidence Statements (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.

Confidence Statements (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).
Confidence Statements (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.

Section 2
Illustrate Construction And Interpretation Of Confidence Intervals For Cost-effectiveness Ratios, Confidence Intervals For Net Monetary Benefit, And Acceptability Curves

- Introduce a well-behaved economic experiment
- Use this experiment to describe the 3 methods for quantifying sampling uncertainty for economic outcomes
- Demonstrate that all 3 methods yield the same confidence statement for the experiment

Cost-Effectiveness Plane

- Upper Left: Treatment Dominated
- Upper Right: Cost Effectiveness Ratio
- Lower Left: Cost Effectiveness Ratio
- Lower Right: Treatment Dominates
- (-) Difference in Cost (+)
- (+) Difference in Effect (+)
Experiment 1

• Suppose we 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:
  \[ \frac{1000}{0.01} = 100,000 / \text{QALY saved} \]

• 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
ΔC = 1000; SEC = 325; ΔQ = 0.01; SEQ = 0.001925; ρ = -.71; DoF = 498

Confidence Statements for 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

Concerns With CI for CER

• If every experiment was like this one, we probably wouldn’t have seen the development of net monetary benefit and acceptability curves
• But experiments can occur in which the CI for CER have “odd properties” that most people at least initially find counter-intuitive
  – CI can be undefined, or
  – On the real number line, either PE > LL > UL or LL > UL > PE
• One response to this problem was the development of net monetary benefits (or NMB)
Net Monetary Benefit

- Composite measure (part cost-effectiveness, part cost benefit analysis), usually expressed in dollar terms, that is derived by rearranging the cost-effectiveness decision rule:

\[ W^* > \frac{\Delta C}{\Delta Q} \]

where \( W^* \) = maximum acceptable cost-effectiveness ratio (e.g., 50,000 per QALY)

- NMB expressed on the cost scale (NHB expressed on the health scale)

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

- For a WTP of 50,000, NMB for experiment 1:

\[(50,000 \times .01) - 1000 = -500\]

NMB Recap

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

- 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 monetary benefit, which equals -intercept (because at the intercept, \( \Delta Q = 0 \), thus \( W^* \Delta Q = 0 \), and we are left with -C)

- For the line passing through the origin, \( \text{NMB} = 0 \)
  - Lines below and to the right of the net monetary benefit=0 line have positive net monetary benefits (i.e., acceptable cost-effectiveness ratios)
  - Lines above and to the left have negative net monetary benefits

- Lines increase in value as we travel southeasterly down the plane
Constructing CI for NMB for Experiment 1, WTP=50K

Constructing CI for NMB for Experiment 1, WTP=100K

Constructing CI for NMB for Experiment 1, WTP=200K

WTP: 50,000; NMB: -500; 95% CI: -1284 to 284

WTP: 100,000; NMB: 0; 95% CI: -945 to 945

WTP: 200,000; NMB: 1000; 95% CI: -290 to 2291
Confidences Statements for CI for NMB

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

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

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

Concerns With CI for CER (2)

- A second response to the “odd properties” of CI for CER was the development of the acceptability curve
- For this curve, we calculate the proportion of the distribution of the difference in cost and effect that is acceptable (i.e., has positive NMB) by calculating the fraction of the distribution that falls below and to the right of 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 fraction that is above and to the left of this line is "unacceptable"
Acceptability Criterion

Constructing the Acceptability Curve

The Acceptability Curve
Review of Results for Experiment 1

Confidence interval for CER
CER CI: (28,200 to 245,200)

Confidence frontier for NMB
CI intersect decision threshold (0) at 28,200 to 245,200

Acceptability curve
Acceptability curve intersects 0.025 and 0.975 at 28,200 and 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)**

We can be confident the more effective therapy is not good value. We cannot be confident the two therapies differ. We can be confident the more effective therapy is good value.

\[ \begin{array}{c}
\text{Willingness to Pay} \\
-\infty & \infty
\end{array} \]

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

**Section 3**

*Illustrate Hard Cases*

- Introduce 2 less well-behaved experiments, and show that even though some of the results of the methods may seem counter-intuitive, the resulting confidence statements remain consistent for the 3 methods of measuring sampling uncertainty.
- (Re)Summarize findings for confidence statements for cost-effectiveness analysis.

**Odd Properties**

- If all experiments were like experiment 1, there wouldn’t be so many issues surrounding confidence intervals for cost-effectiveness ratios.
- But for some experiments the 95% CI for CER cannot be defined, and for some they can have
  \[ PE > LL > UL \lor LL > UL > PE \]
- Odd properties or not, in these latter experiments, confidence intervals for cost-effectiveness ratios, confidence intervals for NMB, and acceptability curves always provide the same confidence statements.
  - i.e., we can be confident therapy A is good or bad value compared to therapy B or we can’t be confident they differ.
Experiment 3

- Suppose we 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:
  \[ \frac{400}{0.02} = 20,000 \text{ QALY saved} \]

\[ \Delta C = 400, \text{ SEC} = 325 (p=0.22); \Delta Q = 0.02, \text{ SEQ} = 0.02 (p = 0.32); \rho = .25; \text{ DOF} = 498 \]
Pattern 3 Findings

- We refer to findings like those in experiment 3 as pattern 3 findings.
- They are 1 of 2 patterns that 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)

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

\[ \text{Willingness to Pay} \]

\[ -\infty \leq \text{Willingness to Pay} \leq \infty \]
Experiment 2

• Suppose we 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: 
  \[
  \frac{35}{0.04} = 875 \text{ QALY saved}
  \]

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

- What confidence statements can we make about this experiment?
  - So long as our WTP is between 28,200 and 245,200, we can be 95% 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 are one of two patterns that occur only when the difference in effect is not significant.
- We know we 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)

- We cannot be confident the two therapies differ from one another.
- We can be confident that one of the therapies is good value.
- We cannot be confident the two therapies differ from one another.

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

(Repeat) Confidence Statements (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.
Confidence Statements (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)

Confidence Statements (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.

Progress to Date

- Have come a long way since 1994
  - Developed 3 dependably accurate methods for evaluating sampling uncertainty for cost-effectiveness ratios
  - Understand how to interpret the results of these methods
  - Introduced probabilistic sensitivity analysis, which is a method for quantifying sampling uncertainty, into decision analysis, which is were the epidemiologists’ question generally originated
- But not everything has been so successful
Section 4
What are (Some) of the Ways We Are Going Wrong In Addressing Sampling Uncertainty

• Failing to report sampling uncertainty
• "Point-estimate" decision making, by which I mean reporting sampling uncertainty and then ignoring it when making adoption recommendations
• Mistakenly overstating the confidence in their results
• Not knowing how to calculate the CI

1) Failing to Report Sampling Uncertainty

• Although we know how to report sampling uncertainty, many investigators are still not doing so
• Doshi and colleagues* have reported that of the economic assessments of trials published in 2003 that reported both costs and effects, only 57% (24 of 42) reported sampling uncertainty for cost-effectiveness


Failing to Report Sampling Uncertainty (2)

• An additional 38 studies that reported only cost differences may have done so because they found no significant difference in effects (i.e., cost-minimization)
  – This historical recommendation is no longer appropriate, because it doesn’t distinguish underpowered trials from adequately powered trials that demonstrate no difference
2) “Point Estimate” Decision Making

- Many investigators calculate and report sampling uncertainty, but then ignore it when they make recommendations about value.
- For example, in one paper the authors found that the therapy under investigation significantly increased direct costs (\(\epsilon_{449}, p=0.01\)), had no significant effect on total direct and indirect costs (\(\epsilon_{332}, p\sim 0.2\)) and had no significant effect on life years gained (0.04, \(p\sim 0.4\)).
- The resulting cost-effectiveness ratios were \(\epsilon_{10,993}\) (DC) and \(\epsilon_{8127}\) (TC).
“The predicted cost-effectiveness ratios were well below the threshold values generally considered cost-effective...[The therapy] appeared to be cost-effective....”

Point Estimate Decision Making (2)

- Depending on whether we consider direct or total cost, the therapy significantly increased cost or had no effect on cost
- The therapy had no significant effect on effect
- Under the old decision rules, we’d have said either the therapies don’t differ or we’d reject the new therapy
- The acceptability curve approaches 80%; thus, we can be at most 60% confident the therapy is good value even if we are willing to pay as much as ε144,000 per life year gained
- The authors ignored all of these findings, and instead made their overly optimistic recommendation based on the magnitude of the point estimate

3) Mistakenly Overstate the Confidence in Their Results

- When the reported confidence limits don’t allow us to be confident of good value, investigators commonly use language that appears like a confidence statement, but which overstates normal confidence levels
Pressure Mattresses for Prevention of Pressure Ulcers

- The authors of a study that compared alternating pressure mattresses vs alternating pressure overlays for the prevention of pressure ulcers found:
  - "The differences in health benefits and total costs...were not statistically significant."
- They did not report the fact, but the 95% CI for the CER was undefined
- They added the following: "however a cost-effectiveness acceptability curve indicated ... [that pressure mattresses] were associated with an 80% probability of being cost saving."
- They concluded: "Alternating pressure mattresses...are more likely to be cost effective..."

Statistical Restatement

- It probably would have been more obvious that this was simply an underpowered trial that did not allow us to differentiate between the therapies if they'd used common statistical language to report their result:
  - "The differences in health benefits and total costs...were not statistically significant," and p=0.4 that pressure mattresses saved money.

The Authors Claim

- The authors of a recent study reported this acceptability curve and concluded: "There was a 90% probability that the cost-effectiveness of early treatment would be less than $40,000"
CER and NMB Confidence Statements

- Had the authors calculated 90% CI for the cost-effectiveness ratio, it would have included $40,000 (i.e., we wouldn't have been 90% confident)
- Had they calculated 90% CI for NMB, it would have included $0 (i.e., we wouldn't have been 90% confident)
- The appropriate formal (2-tailed) confidence statement associated with this acceptability curve is that we can be 80% confident that the therapy has a cost-effectiveness ratio that is below $40,000

More Pessimistic Conclusion

- Can’t be 95% confident that the therapy represents good value

4) Not Knowing How to Calculate the CI

- Many investigators don’t know how to calculate confidence intervals for the hard cases (and probably wouldn’t know how to interpret the results if they did)
- A recent study reported the following results

<table>
<thead>
<tr>
<th>Therapy</th>
<th>Cost (SD)</th>
<th>Effect (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy 1 (n=205)</td>
<td>1415 (968)</td>
<td>6.84 (10.9)</td>
</tr>
<tr>
<td>Therapy 2 (n=200)</td>
<td>1729 (1291)</td>
<td>5.27 (9.96)</td>
</tr>
<tr>
<td>Difference</td>
<td>-314</td>
<td>1.57</td>
</tr>
</tbody>
</table>
Expectation

- Given the reported standard deviations and sample sizes, it is possible to calculate approximate standard errors for the differences *
  - Difference in cost: -314, SE 113.6, p = 0.006
  - Difference in effect: 1.57, SE 1.03, p = 0.13
  (Reported Mann Whitney p = 0.168)
- Because the difference in mean effect is not significant, we should expect a pattern 2 result (LL > UL, and an interval that includes $-\infty$/$+\infty$)

\[
\text{SE} = \sqrt{\frac{\text{SD}_1^2}{N_1} + \frac{\text{SD}_2^2}{N_2}}
\]

Reported Intervals

- The authors tried to use the bootstrap percentile method and reported a pattern 1 interval that ranged from -2069 to 1809
  - Interpretation, We can be confident of value for WTP greater than 1809
- Using Fieller's theorem, the interval is pattern 2 and -- assuming the correlation of the difference is 0 -- includes the region between $-\infty$ and -48 and the region between 650 and $+\infty$
  - Interpretation: We can be 95% confident of value for WTP between 0 and 650; we can’t be confident the therapies differ for WTP greater than 650

The Authors’ Mistake

- When calculating a bootstrap percentile method CI, we generate a set of bootstrap replicates and order them
- The authors used the naïve ordering of most negative to most positive
The Authors’ Mistake (2)

• Had they plotted the data on the cost-effectiveness plane, they would have seen that they had constructed an incorrect interval.

Correct Ordering

4) Not Knowing How to Calculate the CI (2)

• “The ellipse defines the 95% confidence interval for the true incremental cost-effectiveness…”
Other Concerns

• Calculation of confidence intervals for decision models may give a false sense of precision because sampling uncertainty is probably not the most important source of uncertainty in these models
• Multi-way acceptability curves, which report the probability that a therapy is best when compared with multiple different therapies, throw away information and are inappropriate for decision making
• The rumor attributed to Karl Claxton that confidence statements aren’t important for decision making

Summary, Section 1

• Confidence statements for clinical and economic outcomes are similar in that when the decision threshold (for cost-effectiveness ratios, WTP) is excluded from the confidence interval we can be confident the therapies differ, and when it is included in the interval we can’t be confident they differ
• They differ in part because we agree on the decision threshold for odds ratios (1), relative risks (1), and absolute differences (0), but we expect differences among various decision makers’ WTP

Summary, Section 2

• For a given experiment, confidence intervals for cost-effectiveness ratios, confidence intervals for NMB, and acceptability curves ALWAYS provide the same confidence statement
  – The therapy is good value, bad value, or we can’t be confident that the therapies’ value differs
• Confidence intervals for cost-effectiveness ratios provide decision makers with concise information (i.e., 0, 1, or 2 numbers) that allows them to determine if they can be confident about a therapy’s value
Summary, Section 3

• Even when an experiment’s confidence measures have “odd properties,” we can always draw 1 of the 3 possible confidence statements AND
• Confidence intervals for cost-effectiveness ratios, confidence intervals for NMB, and acceptability curves ALWAYS yield the same confidence statement (you just have to know how to interpret it)

Summary, Section 4

• Our training is inadequate because we are still making too many mistakes related to sampling uncertainty for cost-effectiveness analysis
  – Investigators don’t always report sampling uncertainty
    • The recommendation to cost minimize when a therapy costs less and has no difference in effect is no longer appropriate
    – Some report it, but then ignore it when making adoption recommendations
    – Some investigators mistakenly overstate the confidence in their results
    – They don’t always know how to calculate the CI
• IMPLICATION: We still need more education

Conclusion

• The science related to sampling uncertainty for cost-effectiveness analysis is only 15 years old
• We’ve made a lot of progress in that time
• There is still ongoing work being done
  – There was just a new (incorrect) method for calculating confidence intervals for cost-effectiveness ratios published in Statistic in Medicine
  – There is yet to be an appreciation of a number of issues related to sample size and power for cost-effectiveness analysis
• There are still many mistakes being made
• But I’m optimistic that our methods are continuing to improve