GENEALIZING THE RESULTS OF TRIALS *

Two of the most important issues confronting the interpretation of the results of multinational trials are whether they apply to countries that:

- Participated in the trial
- Did not participate in the trial

* Much of the following material was developed in conjunction with John Cook, Mike Drummond, and Joseph Heyse. A manuscript currently is under review.
Multicenter and multinational trials are the norm for the evaluation of new medical therapies, in part because they:

- Speed the development process

- Broaden the representativeness of the patients who receive the therapy

- Familiarize clinical decision makers with the therapy prior to its approval by regulatory agencies such as the FDA

There has been a growing trend to incorporate economic evaluations within these trials.

The information developed from these evaluations is intended to inform decision makers about the value for the cost of new drugs and technologies.

It will shed the most light on the question of value for the cost if the trial and the evaluation are appropriately designed, if appropriate data are collected and are appropriately analyzed, and if the many sources of uncertainty surrounding these evaluations are adequately addressed.

The issues being described apply equally to generalization to:

- Individual countries in multinational trials

- Individual centers in multicenter trials conducted in a single country and

  * These issues may be embedded within one another when one wants to make decisions for individual centers in individual countries that participated in a multinational trial

To simplify explication (and because the data used in the example are drawn from a multinational trial), in what follows, we substitute the phrase individual countries for the phrase individual centers and countries and substitute the phrase multinational trials for the phrase multicenter and multinational trials.

- In most, if not all, cases, however, one can substitute the terms center for country and multicenter for multinational without changing the meaning of the statements.
SOURCES OF UNCERTAINTY

- Parameters measured without variation (e.g., unit cost estimates, discount rates, etc.)
  - Usually addressed with sensitivity analysis
- Projection beyond the trial
  - Usually addressed with decision analysis
- Sampling error
  - Usually addressed by estimating CI for the economic result
- Generalizability
  - Topic of this session’s discussion

DIFFERING MORBIDITY/MORTALITY PATTERNS

- Concerns about generalizability may arise due to variations in underlying morbidity/mortality patterns in different countries
  - e.g., cost effectiveness of cholesterol-lowering agents may vary if underlying risk for CHD differs
- Potential sources
  - Naturally occurring differences
  - Differences between populations who meet official entry criteria (e.g., cream-skimming or reverse cream-skimming)
  - Differences in disease identification

DIFFERING PRACTICE PATTERNS

- Concerns about generalizability may arise due to variations in practice patterns in different countries
  - NOTE: While a commonly expressed concern is that protocols dramatically affect and homogenize therapy, in a number of trials in which I have been involved, practice variation has been observed across countries
  - Effectiveness of intervention may be related to the other supportive care received by the patient
    - e.g., therapies can be complementary with intensive care (i.e., more effective when combined with intensive care), and may be less cost effective in settings that use less intensive therapy
    - Alternatively, therapies can be substitutes for intensive care (i.e., more effective when not combined with intensive care), and may be more cost effective in settings that use less intensive therapy

GENERALIZABILITY

- There has been growing concern that the pooled (i.e., average) clinical and economic results from multinational trials may not be reflective of the results that would be observed in individual countries that participated in the trial
  - Thus it may be difficult for decision makers in specific countries to draw conclusions about the value for the cost of the therapies that were evaluated in the trials
- This difficulty falls under the more general issue of generalizability, i.e., the applicability of the results of a given clinical trial to other populations or subpopulations
- In this presentation, we propose statistical models that evaluate the homogeneity of the economic results among different countries
DIFFERING PRACTICE PATTERNS (II)

- Practice pattern may affect and be affected by relative cost
  - There is (limited) evidence that practice styles vary in the face of differing relative prices
  - If the intervention either substitutes for or is complimentary with intensive care, one would expect its effectiveness to vary when relative prices of hospital units vary by country

DIFFERING UNIT COSTS

- Probably the most frequently cited reason for lack of generalizability, possibly with the smallest impact on the results of the study

Tirilazad Mesylate for subarachnoid hemorrhage, cost per death averted (subanalysis using data from 5 countries) *

<table>
<thead>
<tr>
<th>Country</th>
<th>Country-Specific Unit Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46,818</td>
</tr>
<tr>
<td>2</td>
<td>57,636</td>
</tr>
<tr>
<td>3</td>
<td>53,891</td>
</tr>
<tr>
<td>4</td>
<td>69,145</td>
</tr>
<tr>
<td>5</td>
<td>65,800</td>
</tr>
<tr>
<td>Overall</td>
<td>45,892</td>
</tr>
</tbody>
</table>


ARE WE LOOKING AT THE RIGHT DATA?

- These concerns often arise/fail to arise when we compare data from administrative data bases in the countries with data from the trial
  - e.g., we may look at rates of events or costs in the trial and compare them to rates and costs in a country and become uncomfortable about the applicability of the trial results when they fail to match and become comfortable when they do

- Are we sure that the two sets of data are comparable?
  - Mean rates and costs available from administrative data bases potentially not from same population as in trial
  - Even if they are, it is possible that the therapy isn’t affecting the average events

- **Alternative strategy**: Use data from the trial to determine whether heterogeneity in the economic result was observed within the countries/centers/ practice patterns observed within the trial
  - The observation of heterogeneity or the lack thereof within the trial is the strongest direct evidence we have that the results will apply to the individual countries that participated in the trial
TRADITIONAL APPROACHES TO GENERALIZABILITY

Traditional approaches to generalizing the economic results from multinational trials to individual countries include:

- Using trial-wide clinical results with costs based on trial wide utilization but using unit prices of the country in question

- Using trial-wide clinical results and costs based on patients treated in the country

These approaches ignore the fact that clinical and economic outcomes may influence one another (differences in costs may affect practice patterns, which in turn may affect outcomes; differing practice patterns may affect outcomes, which in turn may affect costs)

An alternative approach has been to use decision analytic models (see Drummond et al. IJTAHC. 1992;8:671-82.)

IMPACT OF UNIT COSTS VS OTHER VARIATION

Tirilazad Mesylate for subarachnoid hemorrhage, cost per death averted (subanalysis using data from 5 countries) *

<table>
<thead>
<tr>
<th>Country</th>
<th>Trial-Wide Effects</th>
<th>Country-Specific Costs and Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit Costs</td>
<td>Costs †</td>
</tr>
<tr>
<td>1</td>
<td>46,818</td>
<td>5,921</td>
</tr>
<tr>
<td>2</td>
<td>57,636</td>
<td>91,906</td>
</tr>
<tr>
<td>3</td>
<td>53,891</td>
<td>90,487</td>
</tr>
<tr>
<td>4</td>
<td>69,145</td>
<td>93,326</td>
</tr>
<tr>
<td>5</td>
<td>65,800</td>
<td>**</td>
</tr>
<tr>
<td>Overall</td>
<td>45,892</td>
<td>45,892</td>
</tr>
</tbody>
</table>

† Country-specific resource use × Country-specific unit costs
** New therapy dominates
THE PROBLEM

! A finding of homogeneity of a therapy’s independent impacts on costs and outcomes need not guarantee the homogeneity of the resulting cost effectiveness ratios or net health benefits associated with the therapy.

! Statistical tests of the clinical endpoints of trials often are based on relative measures such as odds ratios, hazard ratios, or relative risks.

! Economic outcomes, on the other hand, are the result of absolute differences:
  - Cost-effectiveness ratios are computed as the ratio of absolute differences in cost and outcome.
  - Net health benefits are computed by multiplying the maximum acceptable ratio (i.e., the ceiling ratio) times the absolute difference in outcome and then subtracting costs.

! Heterogeneity in absolute treatment effects (measured as a difference) can occur when there are large country-to-country differences in underlying rates of events coupled with a constant multiplicative treatment effect (i.e., homogeneity in relative treatment effects).

EXAMPLE

! Suppose that the probability of death in the placebo group is 20% in country 1 and 10% in country 2.

! Suppose in one case the odds ratio for death associated with active intervention is 0.5 and has been found to be homogeneous in countries 1 and 2.

! Suppose in another case, the odds ratios for death associated with active intervention are 0.70 and 0.5 in the two countries, and have been found to be heterogeneous.

! In which case will the country-specific absolute differences in outcome be similar to one another and in which case will they differ?

EXAMPLE (II) *

<table>
<thead>
<tr>
<th>Country</th>
<th>$p_{Vehicle}$</th>
<th>OR</th>
<th>$p_{Active}$†</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneous odds ratios</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.111</td>
<td>0.099</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.053</td>
<td>0.047</td>
</tr>
<tr>
<td>Heterogeneous odds ratios</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.72</td>
<td>0.153</td>
<td>0.047</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.053</td>
<td>0.047</td>
</tr>
</tbody>
</table>

* $p_{active}$ = probability of death among those receiving active intervention; $p_{Vehicle}$ = probability of death among those receiving vehicle; OR = odds ratio; Difference = $p_{Vehicle}$ - $p_{active}$

† $p_{active} = (p_{Vehicle} \times OR) / ((p_{Vehicle} \times OR) + (1-p_{vehicle}))$
AN ALTERNATIVE APPROACH

The complexities related to assessing the homogeneity of the country-specific cost effectiveness ratios (and net health benefits) by evaluating the homogeneity of the country-specific treatment effects on cost and effects suggest an alternative approach:

- Estimate country-specific cost effectiveness ratios (net health benefits) and evaluate their homogeneity directly

One would use the more precise pooled (average) ratio (net health benefits) for the overall study to represent these countries’ ratios only if:

1) It appears that there is no country-by-ratio (country-by-net health benefits) interaction and

2) The minimum detectable difference was small enough to be economically important

In what follows, I am going to focus on homogeneity of cost effectiveness ratios; however, similar techniques -- without the complexity of statistical tests of ratios are available for testing the homogeneity of net health benefits

THE STRATEGY

The methods currently available for the development of confidence intervals for cost effectiveness ratios can also be used to evaluate the homogeneity of cost effectiveness ratios

Tests of homogeneity can be used to determine:

- Whether the cost effectiveness ratios are inconsistent in both direction and magnitude (i.e., some ratios are acceptable and some are not)
  
  or

- Whether they are consistent in direction but not in magnitude (i.e., the ratios are all acceptable or they are all unacceptable, but they differ in their magnitude)

TESTS OF HOMOGENEITY

Clinical investigators have long recognized the need to evaluate the homogeneity of the outcomes observed in multinational trials

Heterogeneity of treatment effects among different subsets of patients (e.g., those in different countries) is called an interaction between treatment and the variable which was used to create the subsets (i.e., a treatment-by-country interaction)

A number of methods are available for assessing whether the treatment effect is homogeneous

- e.g., F or \( \chi^2 \) tests to determine whether a set of country-by-outcome interaction terms are statistically significant
THE GAIL AND SIMON TESTS OF INTERACTION

Gail and Simon have proposed two tests of interaction to determine whether results are inconsistent in both direction and magnitude or whether they are consistent in direction but not in magnitude.

- They defined a qualitative or crossover interaction as one where the treatment effect is positive (i.e., has acceptable ratios) in some countries, and negative (i.e., has unacceptably high ratios) in other countries (i.e., inconsistent in both direction and magnitude).

- They defined a non-crossover interaction as one where there is variation in the magnitude of the effect (i.e., variation in the ratio), but not in its direction (e.g., when the treatment effect has acceptable ratios in all countries). Peto has termed the latter type of interaction a quantitative interaction.

INTERPRETATION OF QUALITATIVE INTERACTION

A finding of qualitative interaction suggests:

- For the clinical outcome, if the treatment is effective in some countries and ineffective in others.

- For the cost outcome, if the treatment saves money in some countries and adds costs in others.

- **INTERPRETATION**: For the cost effectiveness ratio, if the treatment has acceptable ratios in some countries and has unacceptably high ratios in others.

Which is the same as:

- For net health benefits, if net health benefits are greater than 0 in some countries and less than 0 in others.

INTERPRETATION OF QUANTITATIVE INTERACTION

A finding of qualitative interaction suggests:

- For the clinical outcome, if the treatment is effective (ineffective) in all countries, but differs in the degree of its effectiveness (ineffectiveness).

- For the cost outcome, if the treatment saves (costs) money in all countries, but differs in the degree of savings (costs).

- **INTERPRETATION**: For the cost effectiveness ratio, if the ratio is acceptable (unacceptable) in all countries, but differs in its acceptability (unacceptability).

Which is the same as:

- For net health benefits, if treatment is net health beneficial (non-net health beneficial) in all countries, but differs in its magnitude.

QUALITATIVE INTERACTION

The formal test for qualitative treatment-by-country interaction of effects uses estimates of the treatment effect and its variance for each of the countries being evaluated (see appendix).

The statistical test is based on a likelihood ratio, with critical values of the test given in Gail and Simon (see table in Appendix).

The power of the test has been described by Pan GH, Wolfe DA. Test for qualitative interaction of clinical significance. Stats in Med. 1997;16:1645-52.
QUANTITATIVE INTERACTION

- The test for quantitative interaction is based on the sum of squared errors of the country-specific treatment effects and the variance of these effects (see Appendix)

- For cost effectiveness analysis, country-specific cost effectiveness ratios and their variance estimates are used to compute the test statistic

- A weighted mean is used in estimating the errors rather than the arithmetic mean

- The test statistic is compared to critical values of the $\chi^2$ distribution with one less degree of freedom than there are countries being evaluated

ESTIMATION ISSUES

- The estimate of treatment effect (e.g., cost effectiveness ratio) and variance for each country should be derived separately and independently

- The treatment effects can be the difference in treatment means, odds ratios, relative risks, cost effectiveness ratios, etc.

- If covariates are available, then they can be the estimated parameters from a regression model (e.g., linear, logistic, poisson), or from a survival model (e.g., Cox Proportional Hazards Model)

- Net health benefits can be evaluated directly

- For the cost effectiveness ratio, its ill-behaved distributional properties lead to use of an angular transformation of the ratio calculated by using the arctan function

- For example, for the evaluation of qualitative interaction, the effects that are being compared are obtained by subtracting the angle corresponding to the prespecified ceiling ratio from the observed country-specific cost-effectiveness angles
FOLLOW-UP/TIME HORIZON

- The trial enrolled patients for 20 months, and follow-up of all available patients was discontinued on a given calendar date (maximum potential follow-up, 41 months)
- For the principal analysis, we adopted a time horizon for analysis of 35 months (because of the sparseness of the data after the 35th month)
- For the current country-specific analyses, for which we calculate results for subgroups as small as 130 individuals, the data are too sparse for making estimates even for 35 months
  - For this subanalysis we have adopted a 27 month time horizon (we report a revised pooled result for this shorter time period as well as country-specific results)

EXAMPLE

- Data drawn from a randomized, double blinded, placebo controlled trial evaluating a drug for severe heart failure
- A total of 1663 patients enrolled in 16 countries were used in our analysis
  - Five countries enrolled more than 100 patients (N’s = 130, 372, 382, 236, and 254)
  - The remaining 10 countries enrolled 289 patients (80 from the developing world and 209 from the developed world)
COSTING

Costs were estimated for hospitalization, active drug therapy, and ambulatory care.

Unit cost estimates for hospitalizations were obtained from 4 of the 5 countries that enrolled more than 100 patients, and from one that enrolled fewer.

- Tried to obtain estimates for 49 of the most frequent reasons for admission to the hospital (representing 92.6% (2486/2684) of hospitalizations).

- For the remaining 137 reasons recorded in the trial (representing 7.4% (198/2684) of hospitalizations), we imputed costs using the available unit cost data as well as data on relative weight and geometric mean length of stay from the U.S. Medicare prospective payment system.

We valued resource use in the 3 developing countries in which unit cost data were not collected by use of unit costs from the 1 developing country in which they were.

We valued resource use in the 7 developed countries in which unit cost data were not collected by use of a simple average of the unit costs from 4 developed countries in which they were.

ANALYSES

All results are based on nonparametric bootstrap analyses for the pooled data and for the country-specific data.

- We report a pooled result, results for four countries that enrolled more than 100 patients and for which unit cost data are available, and results for a category made up of the remaining 11 countries (N = 525).

- For the country-specific results we use resource use and unit cost data from the specific country under evaluation (except as noted above for countries where unit cost data were not collected).

Incremental Costs and QALYS: We report pooled and country-specific estimates of incremental costs and QALYS (mean, 95% CI). We also report tests for country-by-treatment interactions for costs and QALYS.
ANALYSIS OF NET HEALTH BENEFITS

Net Health Benefits

- For some of the country-specific analyses, results of the bootstrap indicate that the estimate may fall into all four of the quadrants of the cost-effectiveness plane (i.e., therapy may dominate placebo, it may be dominated by placebo, it may cost more and be more effective, or it may cost less and be less effective)

- Given the difficulty that this pattern of results poses to ordering the bootstrap replicates and identifying a confidence interval, we have chosen instead to report the net health benefits for the therapy

- Net health benefits are calculated by multiplying incremental QALYS times a cost-effectiveness ratio that represents the acceptable upper limit of cost-effectiveness (hereafter referred to as the ceiling ratio) and then subtracting incremental costs from this product. If net health benefits exceed 0, then the therapy is acceptable

REPORTED RESULTS

- For the results reported here, we report in tabular form the net health benefits calculated using three ceiling ratios, $20,000, $50,000, and $80,000

- The 95% CI for net health benefits -- for which there is no ambiguity in ordering -- are derived from the nonparametric bootstrap

- We also plot the net health benefits for ceiling ratios ranging from $0 to $100,000 and report the probability that net health benefits are greater than $0 for ceiling ratios ranging from $0 to $100,000.

- Finally, we report statistical tests of the homogeneity of the net health benefits between the four individual countries for which the analyses were performed and the remaining pool of countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Incremental Costs Mean</th>
<th>95% CI</th>
<th>Incremental QALYS Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>-847</td>
<td>-2015 to 316</td>
<td>0.08</td>
<td>0.04 to 0.12</td>
</tr>
<tr>
<td>C1</td>
<td>-802</td>
<td>-4895 to 2872</td>
<td>0.09</td>
<td>-0.02 to 0.22</td>
</tr>
<tr>
<td>C2</td>
<td>-932</td>
<td>-2647 to 654</td>
<td>0.08</td>
<td>0.01 to 0.16</td>
</tr>
<tr>
<td>C3</td>
<td>-2457</td>
<td>-6056 to 945</td>
<td>0.06</td>
<td>-0.03 to 0.14</td>
</tr>
<tr>
<td>C4</td>
<td>-452</td>
<td>-3195 to 2606</td>
<td>0.15</td>
<td>0.05 to 0.25</td>
</tr>
<tr>
<td>Other</td>
<td>283</td>
<td>-1243 to 1869</td>
<td>0.04</td>
<td>-0.01 to 0.10</td>
</tr>
</tbody>
</table>

Statistical tests of homogeneity

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Qualitative p</th>
<th>Quantitative p</th>
</tr>
</thead>
<tbody>
<tr>
<td>p &gt; 0.50</td>
<td>p = 1.00</td>
<td></td>
</tr>
<tr>
<td>p = 0.65</td>
<td>p = 0.50</td>
<td></td>
</tr>
</tbody>
</table>
Pooled and Country-Specific Estimates of Net Health Benefits for the First 27 Months of the Trial for Selected Ceiling Ratios

<table>
<thead>
<tr>
<th>Ceiling Ratio</th>
<th>Pooled</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% CI</td>
<td>Mean</td>
<td>95% CI</td>
<td>Mean</td>
<td>95% CI</td>
</tr>
<tr>
<td>$20,000</td>
<td>2429</td>
<td>1157 to 3793</td>
<td>2570</td>
<td>-1959 to 7432</td>
<td>2509</td>
<td>400 to 4743</td>
</tr>
<tr>
<td></td>
<td>1157 to 3793</td>
<td>2741 to 6953</td>
<td>1521 to 13,073</td>
<td>768 to 8697</td>
<td>400 to 4743</td>
<td>67 to 10,849</td>
</tr>
<tr>
<td>$50,000</td>
<td>4801</td>
<td>2741 to 6953</td>
<td>5224</td>
<td>-1521 to 13,073</td>
<td>4876</td>
<td>768 to 8697</td>
</tr>
<tr>
<td></td>
<td>2741 to 6953</td>
<td>1521 to 13,073</td>
<td>13,073 to 13,073</td>
<td>8697 to 8697</td>
<td>67 to 10,849</td>
<td>67 to 10,849</td>
</tr>
<tr>
<td>$80,000</td>
<td>7174</td>
<td>4171 to 10,389</td>
<td>7877</td>
<td>-1876 to 19,292</td>
<td>7242</td>
<td>1118 to 13,094</td>
</tr>
<tr>
<td></td>
<td>4171 to 10,389</td>
<td>-1876 to 19,292</td>
<td>19,292 to 19,292</td>
<td>13,094 to 13,094</td>
<td>1118 to 13,094</td>
<td>-557 to 14,595</td>
</tr>
</tbody>
</table>

Statistical tests of homogeneity

<table>
<thead>
<tr>
<th>Interaction</th>
<th>P = 1.00</th>
<th>p = 1.00</th>
<th>p = 1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative</td>
<td>p = 0.51</td>
<td>p = 0.40</td>
<td>p = 0.43</td>
</tr>
<tr>
<td>Quantitative</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Net health benefits calculated for ceiling ratios ranging from $0 to $100,000 for C1, C2, C3, C4, the (pooled) 11 miscellaneous countries for which unit cost data were not available (Oth), and the trial as a whole (P.A.)
Figure 2. The probability that net health benefits exceed $0 for ceiling ratios ranging from $0 to $100,000 for C1, C2, C3, C4, the (pooled) 11 miscellaneous countries for which unit cost data were not available (Oth), and the trial as a whole (P.A.)

EXAMPLE DISCUSSION (I)

This example addressed the question, “Are the pooled results of a trial representative of the results that would be expected in the individual countries that participated in the trial?” by evaluating the homogeneity of incremental costs, incremental QALYS, and net health benefits observed in four individual countries and the (pooled) 11 miscellaneous countries for which unit cost data were not available.

- We evaluated homogeneity of net health benefits rather than of the cost-effectiveness ratios because the pattern of results on the cost-effectiveness plane led to problems with ordering of observations for the estimation of confidence intervals for the country-specific cost-effectiveness ratios.

EXAMPLE DISCUSSION (II)

In the example, we found little evidence of difference in the incremental costs and QALYS associated with aldactone therapy for severe heart failure among the different countries that were evaluated.

- We also found little evidence of heterogeneity in the country-specific net health benefits.

In this case, one may want to use the more precise trial-wide (i.e., pooled) estimate for each of the countries in the trial.

EXAMPLE DISCUSSION (III)

The pooled point estimate suggested that usual care plus active therapy dominated usual care alone, and the upper limit of the 95% confidence interval for the cost per QALY ratio equalled $3865 (for the pooled result, there was no problem with ordering).

- This finding translates into a finding that for any ceiling ratio greater than $3865, the lower limit of the 95% confidence interval for net health benefits will exceed 0.

EXAMPLE DISCUSSION (IV)

Could there be a difference in the country-specific results yet we not be able to detect it?

- A qualified yes

The power of the test of homogeneity is low.

- However, except for the result for 11 miscellaneous countries for which unit cost data were not available, all of the country-specific evidence suggests that aldactone therapy saved money and increased QALYS.
DISCUSSION

In this presentation, we outlined a method for evaluating the homogeneity of the country-specific cost effectiveness ratios calculated in multinational trials to determine whether reporting a pooled result is appropriate for all countries or whether it is more appropriate to report separate ratios for each country.

In addition to a general test of interaction, we proposed determining if the treatment effects are inconsistent in both direction and magnitude or if they are consistent in direction but not in magnitude.

This set of homogeneity tests is a reformulation of Gail and Simon’s tests for the evaluation of the homogeneity of clinical outcomes to make them applicable to the evaluation of the homogeneity of cost effectiveness ratios.

HOW TO INTERPRET THE RESULTS

A finding of homogeneity is a necessary but not sufficient condition for attributing the pooled results of the trial to all of the countries that participated in the trial.

- Given the potential limitations in the power of the test of homogeneity, failure to detect heterogeneity does not mean the results are homogeneous.

- Ex-post power calculations can provide additional information about the degree of homogeneity that may exist, when heterogeneity is not detected.

When there is evidence of heterogeneity, one should explore the possible reasons for it, and determine whether it is qualitative or quantitative in nature.

Use of pooled estimates to represent the ratio for all of the countries in the trial is less problematic in the presence of quantitative interaction, because the direction of the effect (i.e., its acceptability) is the same in all of the countries.

In the face of qualitative interaction, pooled estimates should not be used. Instead, one should attempt to investigate why the therapy is cost-effective in some countries, but not in others.
WHY ARE COVARIATES IMPORTANT?

! Given that patients were not randomized to the country in which they received treatment, one should compare the characteristics of patients in the different countries to determine if severity levels were similar in the different countries

! One should also adopt analytic strategies that allow one to control for potential transnational imbalances in the covariates

! These analyses will allow one to determine whether differential treatment effects are due to observable variations in the types of patients treated in the different countries or whether they are independent of such observable variations

DEEP ASSUMPTIONS

! These techniques assume that what we observed is a guide to what we did not observe

- We use evidence of heterogeneity or the lack thereof across countries, practice patterns, levels of illness severity, etc., to determine whether the pooled results are applicable across countries, practice patterns, levels of illness severity, etc.

! These techniques will fail if:

- What we observed is so transformed by the study protocol that it is not a guide to what we would have observed had the therapy been used outside of a trial

- We desire to know about cost effectiveness in settings or populations that are unlike anything observed in the trial (e.g., the trial was conducted in tertiary care hospitals but the therapy actually will be delivered in community hospitals)

! NOTE: If the heterogeneity test is thought to be inapplicable, that probably means we have no evidence of the therapy’s affect for use in decision analytic models that might be used to assess generalizability

CONCLUSION

! The growing policy demand for health technology assessment often requires that limited clinical and economic data be applied to a variety of different populations

! Tests used to assess country-by-treatment interactions in clinical endpoints can be adapted to assess the presence of an interaction in economic outcomes, including the cost-effectiveness ratio

! The use of the Gail and Simon tests for quantitative and qualitative interaction can be useful in determining if and how country-specific results can be pooled to obtain a trial-wide estimate of the economic impact or cost effectiveness of treatment

! Tests of homogeneity should partially offset difficulties decision makers in specific countries have in drawing useful conclusions about the value for the cost of the therapies from trials

! In the future, it will be important to consider design issues such as sample size per country when economic assessments are planned in multinational clinical trials
**“STANDARDIZED” ANGLES**

Cost-effectiveness angles ranging between -180° and 180°, where 45° equals a cost effectiveness ratio of 50,000 / unit of outcome

---

**DISTRIBUTIONAL PROPERTIES OF ANGLES VS RATIOS**

<table>
<thead>
<tr>
<th></th>
<th>Skewness Ratio</th>
<th>Skewness Angle</th>
<th>Kurtosis Ratio</th>
<th>Kurtosis Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full sample</td>
<td>2.54</td>
<td>-0.05</td>
<td>25.37</td>
<td>3.30</td>
</tr>
<tr>
<td>Costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>2.34</td>
<td>-0.06</td>
<td>20.71</td>
<td>3.26</td>
</tr>
<tr>
<td>Log normal</td>
<td>2.74</td>
<td>-0.04</td>
<td>30.03</td>
<td>3.35</td>
</tr>
<tr>
<td>Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25-0.15</td>
<td>0.92</td>
<td>-0.06</td>
<td>6.08</td>
<td>3.20</td>
</tr>
<tr>
<td>0.55-0.45</td>
<td>4.17</td>
<td>-0.04</td>
<td>44.67</td>
<td>3.41</td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.50</td>
<td>4.17</td>
<td>-0.24</td>
<td>42.26</td>
<td>3.23</td>
</tr>
<tr>
<td>-0.25</td>
<td>4.07</td>
<td>-0.18</td>
<td>44.97</td>
<td>3.28</td>
</tr>
<tr>
<td>0.00</td>
<td>2.58</td>
<td>-0.11</td>
<td>23.67</td>
<td>3.30</td>
</tr>
<tr>
<td>0.25</td>
<td>1.48</td>
<td>0.00</td>
<td>11.23</td>
<td>3.25</td>
</tr>
<tr>
<td>0.50</td>
<td>0.41</td>
<td>0.28</td>
<td>4.76</td>
<td>3.46</td>
</tr>
</tbody>
</table>
APPENDIX

Definitions

Treatment Effects: \( D_i \), where \( i = \text{countries } 1 \text{ to } K \), and are estimates of the actual treatment effects (\( \delta_i \)).

Variance of the treatment effects: \( S_i^2 \), where \( i = \text{countries } 1 \text{ to } K \).

Statistical Test for Qualitative Interactions

Null hypothesis: All \( \delta_i \) are either greater than 0 or they all are less than 0.

Compute the following quantities:

\[
Q^+ = \sum_{i=1}^{K} (D_i^2 / S_i^2) \forall D_i \geq 0
\]

\[
Q^- = \sum_{i=1}^{K} (D_i^2 / S_i^2) \forall D_i < 0
\]

The likelihood ratio is expressed by:

\[
Q = \min(Q^+, Q^-) > c
\]

Critical values for \( c \) are given in Gail and Simon (Gail, 1985).

Statistical Test for Quantitative Interactions

Null hypothesis: All \( \delta_i \) are equal.

Compute the following quantity:

\[
H = \sum_{i=1}^{K} (D_i - \overline{D})^2 / S_i^2
\]

where

\[
\overline{D} = \frac{\sum_{i=1}^{K} D_i^2 / S_i^2}{\sum_{i=1}^{K} 1 / S_i^2}
\]

Computed values of \( H \) are compared to critical values of the \( \chi^2 \) distribution with \( K-1 \) degrees of freedom. Large values of \( H \) indicate heterogeneity.

Critical Values (c) for the Likelihood Ratio Test: \( \min (Q^+, Q^-) > c \)

<table>
<thead>
<tr>
<th>No. of Groups</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.64</td>
<td>2.71</td>
<td>3.84</td>
<td>9.55</td>
</tr>
<tr>
<td>3</td>
<td>2.95</td>
<td>4.23</td>
<td>5.54</td>
<td>11.76</td>
</tr>
<tr>
<td>4</td>
<td>4.01</td>
<td>5.43</td>
<td>6.86</td>
<td>13.47</td>
</tr>
<tr>
<td>5</td>
<td>4.96</td>
<td>6.50</td>
<td>8.02</td>
<td>14.95</td>
</tr>
<tr>
<td>6</td>
<td>5.84</td>
<td>7.48</td>
<td>9.09</td>
<td>16.32</td>
</tr>
<tr>
<td>7</td>
<td>6.67</td>
<td>8.41</td>
<td>10.09</td>
<td>17.85</td>
</tr>
<tr>
<td>8</td>
<td>7.48</td>
<td>9.29</td>
<td>11.05</td>
<td>18.78</td>
</tr>
<tr>
<td>9</td>
<td>8.26</td>
<td>10.15</td>
<td>11.98</td>
<td>19.93</td>
</tr>
<tr>
<td>10</td>
<td>9.02</td>
<td>10.99</td>
<td>12.87</td>
<td>21.03</td>
</tr>
<tr>
<td>12</td>
<td>10.50</td>
<td>12.60</td>
<td>14.57</td>
<td>23.13</td>
</tr>
<tr>
<td>14</td>
<td>11.93</td>
<td>14.15</td>
<td>16.25</td>
<td>25.17</td>
</tr>
<tr>
<td>16</td>
<td>13.33</td>
<td>15.66</td>
<td>17.85</td>
<td>27.10</td>
</tr>
<tr>
<td>18</td>
<td>14.70</td>
<td>17.13</td>
<td>19.41</td>
<td>28.96</td>
</tr>
<tr>
<td>20</td>
<td>16.04</td>
<td>18.57</td>
<td>20.94</td>
<td>30.79</td>
</tr>
<tr>
<td>25</td>
<td>19.34</td>
<td>22.09</td>
<td>24.65</td>
<td>35.15</td>
</tr>
<tr>
<td>30</td>
<td>22.55</td>
<td>25.50</td>
<td>28.23</td>
<td>39.33</td>
</tr>
</tbody>
</table>
REFERENCES


