TRANSFERABILITY OF THE RESULTS FROM MULTINATIONAL TRIALS

Henry Glick, PhD

University of Pennsylvania School of Medicine
Leonard Davis Institute of Health Economics

Astra Zeneca Methodology Seminar
December 6, 2005
MULTINATIONAL TRIALS

• Multinational trials are the norm for the evaluation of new medical therapies
  - 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

• Recent trend toward inclusion of countries with heterogeneous economic conditions
  - Reduce the cost of the trials
  - ? Increase event rates ?
ECONOMICS AND INTERNATIONAL TRIAL

- 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.
OUTLINE

- Problems in interpreting economic results from multinational trials?
- Bad solutions to the problem
- Which is more important, country-specific medical service use or price weights?
- Price weight recommendations
- Two analytic approaches for the evaluation of transferability
  - Hypothesis tests of homogeneity
  - Estimation by use of multi-level random-effects model shrinkage estimators
- General comments

Technical appendix

References
INTERPRETING ECONOMIC RESULTS FROM MULTINATIONAL TRIALS?

- The Problem:
  - There has been growing concern that the pooled (i.e., average) economic results from multinational trials may not be reflective of the results that would be observed in individual countries that participated in the trial.
  - Similar issues arise for any subgroup of interest in the trial (e.g., more and less severely ill patients).
COMMON SOURCES FOR CONCERN

- Transnational differences in morbidity/mortality patterns; practice patterns (i.e., medical service use); and absolute and relative prices for medical service use (i.e., price weights)

- Thus decision makers may find it difficult to draw conclusions about the value for the cost of the therapies that were evaluated in multinational trials
DIFFERING MORBIDITY/MORTALITY PATTERNS

- Concerns about transferability 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 in disease identification and treatment
  
  - Variation in severities of disease of enrolled patients between countries (e.g., cream-skimming or reverse cream-skimming)
DIFFERING PRACTICE PATTERNS

- Concerns about transferability may arise due to variations in practice patterns in different countries
  
  - Effectiveness of intervention may be related to the other care received by the patient
    
    * e.g., therapies can be complementary with other care (i.e., more effective when combined with care), and thus less cost-effective in settings that use less other care
    
    * Or, therapies can be substitutes for other care (i.e., more effective when not combined with other care), and thus more cost-effective in settings that use less other care
DIFFERING PRACTICE PATTERNS (II)

● Practice pattern may affect and be affected by relative price

  - There is (limited) evidence that practice styles vary in the face of differing relative prices

  - If an 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

● NOTE: While a commonly expressed concern is that protocols dramatically affect and homogenize treatment patterns, in a number of trials in which I've been involved, practice variation has been observed across countries
DIFFERING PRICE WEIGHTS

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Price Weights</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>

BAD SOLUTIONS

• Use trial-wide clinical results, trial-wide medical service use, and price weights from one country
  - e.g., to tailor the results to the U.S., just use U.S. price weights, and conduct the analysis as if all participants were treated in the U.S.

• Use trial-wide clinical results and use costs derived from the subset of 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)
PRICES VS SERVICES?

- When addressing transferability of data from other countries, Australian, Canadian, and British guidelines for cost-effectiveness primarily discuss variability in price weights and medical service use.

- All 3 request that local price weights be used in evaluations submitted to the government. At least one has a preferred schedule.

- All 3 also recognize that medical service use in other countries may not reflect their medical service use. None provide specific instructions for addressing the latter issue.

WHAT DO WE KNOW ABOUT THEIR RELATIVE IMPORTANCE?
Drummond and colleagues reviewed the "main causes of variation in study results from place to place"

Conducted a literature review using OHE-HEED and NHSEED to identify 46 European studies with inter-country comparisons

- In current presentation, 2 studies omitted because they had "detailed data from only one country, with negligible information about other countries"
CLASSIFICATION BY EXTENT OF "GENERALIZABILITY"

- Classified "based on size of difference in results and importance for decision-making"

  - More generalizable: Observed differences in ICERS unlikely to change adoption decision
  
  - Less generalizable: Observed differences in ICERS likely to change adoption decision
TRANSFERABILITY OF RESULTS IN 44 STUDIES

PW = Price weight; RU = Resource use; Ep = Epidemiology

Authors' conclusion: "...the amount of variation you find depends on the amount the analyst allows (in study design or analysis)"

Source: Drummond MF. Are the results of economic evaluations generalizable? iHEA 4th World Congress.
<table>
<thead>
<tr>
<th>Country</th>
<th>Price Weights</th>
<th>Country Specific Costs †</th>
<th>Country Specific Costs and Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46,818</td>
<td>5,921</td>
<td>11,450</td>
</tr>
<tr>
<td>2</td>
<td>57,636</td>
<td>91,906</td>
<td>60,358</td>
</tr>
<tr>
<td>3</td>
<td>53,891</td>
<td>90,487</td>
<td>244,133</td>
</tr>
<tr>
<td>4</td>
<td>69,145</td>
<td>93,326</td>
<td>181,259</td>
</tr>
<tr>
<td>5</td>
<td>65,800</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Overall</td>
<td>45,892</td>
<td>45,892</td>
<td>45,892</td>
</tr>
</tbody>
</table>


† Country-specific resource use \times\ Country-specific price weights

** New therapy dominates
A SECOND EXPERIMENT

- One problem with last experiment is that when we assessed price weights alone, we multiplied them times trial-wide medical service use; when we assessed country-specific medical service use, the Ns for each mean were the Ns from each country

- When one subdivides a larger group into smaller groups, one expects variability to grow!
DESIGN OF SECOND EXPERIMENT

To address this problem we used data from 2 RCTs to design a different experiment in which the N's per country and per treatment group were maintained.

1) To assess the relative contribution of price weights, we randomly assigned patients to countries' price weights and assessed results for each set of price weights (pseudo-countries based on prices).

2) To assess the relative contribution of service use, we randomly assigned countries' price weights to participants in each country and assessed results for each set of countries' participants (pseudo-countries based on medical service use).
INTERPRETING SECOND EXPERIMENT

- If results from all pseudo-countries shrink towards a common mean, suggests variable that is held constant (e.g., groups of price weights) isn't that important

- If results from all pseudo-countries have patterns of variability similar to what was observed in the trial, suggests variable that is held constant (e.g., practice pattern) is important

- Complexity of relationship between price and service use may lead to other patterns of results

- When we calculate costs for patients who were all treated in the same country (i.e., when we randomize price weights), what look like country-specific effects may be due to differences in either practice pattern or severity of illness
  
  - Used multivariable analysis in an attempt to control for differences in severity that may exist between countries
TAKE HOME MESSAGES FROM EXPERIMENTS

- Unlikely that multiplication of trial-wide service use times a country's price weights provides information about the likely impact of the therapy when it is used in that country

- In general, dramatically shrinks estimates for all countries to a common result
TAKE HOME MESSAGES (cont.)

- While there is evidence that price weights matter, among countries with similar levels of economic development, accounting for differences in medical service use appears to be more important than differences in price weights.

  - Limited evidence that when working with dramatically different price weights (e.g., from the developing and the developed world), price weight differences more important (see LA1 in experiment 2B).

- Should perform experiment in more trials -- particularly those with evidence of heterogeneity -- to evaluate consistency of these findings across trials.
PRICE WEIGHTS FROM WHICH COUNTRIES?

- The countries from which price weights are collected might be ones:
  - That represent the spectrum of economic development among countries that participated in the trial
  - That enroll a large number of patients in the trial
  - In which the countries’ regulators require a submission for reimbursement
  - For which price weights are readily available
  - In which the study sponsor wishes to make economic claims
MORE PRICE WEIGHTS/FEWER COUNTRIES VS MORE COUNTRIES/FEWER PRICE WEIGHTS

● Presuming one is using a reliable method for imputing price weights, do we know anything about how we should trade-off number of countries sampled vs number of price weights per country?

● In simulations we performed to evaluate whether it was better to collect fewer price weights in more countries or more price weights in fewer countries*, we found:

   - Imputation error decreased as the number of medical services and countries sampled increased, but the rate of reduction in this error shrank

PRICE WEIGHTS / COUNTRIES? (cont.)

- Error was minimized by obtaining estimates for fewer types of hospitalization -- in our experiment, at least 25 -- from as many countries as is feasible, rather than by obtaining estimates for more types of hospitalization in fewer countries.

- We did not evaluate such a strategy, but it may be more efficient to measure a core group of services in all countries plus a set of services that differ for different countries.
COUNTRY-SPECIFIC VS. AVERAGED PRICE WEIGHTS

Once one has a number of different sets of price weights (e.g., weights from 5 countries that participated in the trial), how should they be used to construct the cost outcome of the trial?

- Ideal: Because relative prices can affect quantities of services provided, where ever feasible, one should multiply country-specific price weights times country-specific counts of medical service use.
COUNTRY-SPECIFIC VS. AVERAGED PRICE WEIGHTS (cont.)

- For countries for which price weights are not available

* Use (averages of) price weight estimates from similar countries

* e.g., in a trial that enrolls patients in Western Europe, Eastern Europe and South America, one might average price weights from the UK, Sweden, and France to value medical service use in Germany, but one wouldn't want to do so to value it in Eastern Europe and Latin America
TWO ANALYTIC APPROACHES TO TRANSFERABILITY

- Two approaches -- which rely principally on data from the trial to address these issues -- have made their way into the literature
  - Hypothesis tests of homogeneity (Cook et al.)
  - Multi-level random-effects model shrinkage estimators
HYPOTHESIS TESTS OF HOMOGENEITY

- Evaluate the homogeneity of the results from the different countries

  - If there is no evidence of heterogeneity (i.e., a nonsignificant p-value for the test of homogeneity), and if one believes the test was powerful enough to rule out economically meaningful differences in costs, then one cannot reject that the pooled economic result from the trial applies to all of the countries that participated in the trial

  - If there is evidence of heterogeneity, then the method indicates one should not use the pooled estimate to represent the result for the individual countries, but this method is less clear about the result that should be used instead
SEPARATE HOMOGENEITY EVALUATIONS FOR COSTS AND EFFECTS?

- Homogeneity of a therapy’s costs and outcomes need not guarantee the homogeneity of the resulting cost-effectiveness ratios or net monetary benefit (NMB) associated with the therapy.

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

- Economic outcomes are the result of absolute differences.

- Heterogeneity in absolute treatment effects (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 you had data from two trials

<table>
<thead>
<tr>
<th>Country</th>
<th>$p_{\text{plac}}^*$</th>
<th>OR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial #1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.05</td>
<td>~1.0</td>
</tr>
<tr>
<td>Trial #2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.72</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

- Will the country-specific absolute differences in outcome be more similar to one another in trials #1 or #2?
### EXAMPLE (cont.)

<table>
<thead>
<tr>
<th>Country</th>
<th>$p_{plac}$</th>
<th>OR</th>
<th>$p_{act}$ †</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trial #1</strong></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>p-value</td>
<td>&lt;0.05</td>
<td>~1.0</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><strong>Trial #2</strong></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>
<tr>
<td>p-value</td>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

† $p_{act} = (p_{plac} \times OR) / \{(p_{plac} \times OR) + (1-p_{plac})\}$
EVALUATING HOMOGENEITY OF VALUE FOR THE COST

- The complexities related to assessing the homogeneity of the country-specific cost-effectiveness / NMB by evaluating the homogeneity of the country-specific treatment effects on cost and effects suggest an alternative approach:
  
  - Estimate country-specific NMB and evaluate their homogeneity directly
EVALUATING HOMOGENEITY OF VALUE FOR THE COST (cont.)

- One would use the more precise pooled (average) ratio / NMB for the overall study to represent these countries’ ratios only if:
  
  1) It appears that there is no country-by-ratio (country-by-NMB) interaction and

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

- In what follows, I am focus on homogeneity of NMB
TESTS OF HOMOGENEITY

- Evaluation of homogeneity of clinical outcomes common
  - Similarity of treatment effect in men and women; young and old; more and less severely ill

- Heterogeneity of treatment effects (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 homogeneity
  - e.g., F or $\chi^2$ tests to assess significance of a set of country-by-outcome interaction terms
GAIL AND SIMON TESTS OF HOMOGENEITY

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

  - Qualitative or crossover interaction: The treatment effect is positive in some countries and negative in other countries (i.e., inconsistent in both direction and magnitude)

  - Non-crossover interaction: There is variation in the magnitude of the effect, but not in its direction. 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, treatment is effective in some countries and ineffective in others
  - For the cost outcome, treatment saves money in some countries and adds costs in others
  - For NMB, treatment has an NMB greater than 0 in some countries and less than 0 in others
A finding of qualitative interaction suggests:

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

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

- For NMB, treatment is net monetary beneficial (non-net monetary beneficial) in all countries, but differs in its magnitude
TEST FOR QUALITATIVE INTERACTION

- The formal test for qualitative treatment-by-country interaction uses estimates of NMB 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
TEST FOR QUANTITATIVE INTERACTION

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

- 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., NMB) and variance for each country should be derived separately and independently

• 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)

  - Impact of "outlying" observations may be greater in evaluation of homogeneity than it is in evaluation of overall treatment effect
EXAMPLE, TEST OF HOMOGENEITY

- 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 11 countries enrolled 289 patients (80 from the developing world and 209 from the developed world).
COSTING

- Costs estimated for hospitalization, active drug therapy, and ambulatory care

- Price weight for hospitalizations obtained from 4 countries that enrolled more than 100 patients, and from 1 that enrolled fewer

- Used average of price weights collected from 4 developed countries to value medical service use in 7 developed countries for which they were unavailable

- Used price weights collected from 1 developing country to value medical service use in the 3 developing countries for which they were unavailable
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 price weight data are available, and results for a category made up of the remaining 11 countries (N = 525)

  - Results reported for costs, QALYs and NMB
RESULTS REPORTED FOR NMB

- We report:
  - NMB and 95% CI for three ceiling ratios, $20,000, $50,000, and $80,000
  - Probability that NMB is greater than $0 for ceiling ratios ranging from $0 to $100,000 (acceptability curves)
  - Statistical tests of the homogeneity of NMB between the four individual countries for which the analyses were performed and the remaining pool of countries
## Pooled and Country-Specific Estimates of Incremental Costs and QALYS

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

Tests for interaction
- Qual. p > 0.50 p = 1.00
- Quant. p = 0.65 p = 0.50
### Pooled and Country-Specific Estimates of Net monetary Benefits Selected Ceiling Ratios

<table>
<thead>
<tr>
<th>Cntry</th>
<th>Mean</th>
<th>95% CI</th>
<th>Mean</th>
<th>95% CI</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$20,000</td>
<td></td>
<td>$50,000</td>
<td></td>
<td>$80,000</td>
<td></td>
</tr>
<tr>
<td>Pooled</td>
<td>2429</td>
<td>1157 to 3793</td>
<td>4801</td>
<td>2741 to 6953</td>
<td>7174</td>
<td>4171 to 10,389</td>
</tr>
<tr>
<td>C1</td>
<td>2570</td>
<td>-1959 to 7432</td>
<td>5224</td>
<td>-1521 to 13,073</td>
<td>7877</td>
<td>-1876 to 19,292</td>
</tr>
<tr>
<td>C2</td>
<td>2509</td>
<td>400 to 4743</td>
<td>4876</td>
<td>768 to 8697</td>
<td>7242</td>
<td>1118 to 13,094</td>
</tr>
<tr>
<td>C3</td>
<td>3583</td>
<td>26 to 7445</td>
<td>5271</td>
<td>67 to 10,849</td>
<td>6960</td>
<td>-557 to 14,595</td>
</tr>
<tr>
<td>C4</td>
<td>3407</td>
<td>-176 to 7029</td>
<td>7838</td>
<td>2037 to 13,836</td>
<td>12,270</td>
<td>3778 to 21,382</td>
</tr>
<tr>
<td>Other</td>
<td>556</td>
<td>-1386 to 2461</td>
<td>1814</td>
<td>-1470 to 4983</td>
<td>3072</td>
<td>-1639 to 7888</td>
</tr>
</tbody>
</table>

**Tests for interaction**

- **Qual**: P = 1.00, p = 1.00, p = 1.00
- **Quan**: p = 0.51, p = 0.40, p = 0.43
EXAMPLE DISCUSSION (I)

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

- 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 - 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 equaled $3865.
EXAMPLE DISCUSSION (II)

- 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 price weight data were not available, all of the country-specific evidence suggests that the therapy saved money and increased QALYS
ESTIMATION

- An alternative to homogeneity tests is to use multi-level random-effects model shrinkage estimation to provide more precise estimates of the country-specific results than are yielded by separate -- and naive -- analysis of each country's costs and effects
  
  - Borrow information from the mean estimate to add precision to the country-specific estimates
  
  - These methods have the potential added advantage of providing better estimates of the uncertainty surrounding the pooled result than naive estimates of the trial-wide result
CURRENT DEBATE

- "Exchangability"
  
  - One must make assumptions about whether:
    
    * The data should be shrunk to a single mean (a predisposition towards homogeneity) or
    
    * The data from some countries should be shrunk to one mean while the data from others should be shrunk to a second mean (a predisposition towards heterogeneity)
IS WHAT WE OBSERVE REPRESENTATIVE?

• 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.
IS WHAT WE OBSERVE REPRESENTATIVE? (cont.)

- 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 transferability
PROBLEMS / POTENTIAL SOLUTIONS

- Trials often have too few subjects per country
  - Modify designs so that more patients are enrolled in fewer countries
  - If infeasible, pool among similar types of countries, where similarity might be defined based on price weights, event rate in placebo group, etc.

- Price weights
  - While one need not obtain price weights for every country that participates in the trial, one should obtain price weights from a subset of countries stratified by level of economic development
TESTS HAVE TOO LITTLE POWER

- Use larger $\alpha$ levels (e.g., 0.1 or 0.2) and look for larger differences between groups

- Pool across other relevant attributes, e.g.,
  - Characteristics of medical practice (e.g., types of centers and providers); characteristics of reimbursement patterns
  - Ex post: Pool countries with most similar results

- Principle: Given all the tests tend toward a finding of homogeneity, try to identify any evidence of heterogeneity that may exist
SAMPLE SIZE AND SHRINKAGE MODELS

- Shrinkage models may not be as affected by small sample sizes per country (although small sample sizes may again predispose one towards findings of homogeneity)

  - Small sample sizes may make it difficulty to determine whether one should shrink to a common mean or to multiple means
TRIALS DON’T PROVIDE INFORMATION FOR COUNTRIES THAT DID NOT PARTICIPATE

- Characterize countries that participated in trials by their practice patterns and match nonparticipating countries to these patterns
CONCLUSIONS

- Issues of transferability are of growing importance as decision makers in more countries desire information about the likely economic impact of the adoption of new therapies.

- Historically, have tended to do a poor job of estimation of transferability: price weights likely aren’t the primary source of variability when countries have similar economic conditions.

- Two recently proposed analytic techniques better address transferability issues:
  - Each uses data from the trial to address these issues.
  - Trade-off problems with power for problems with exchangability.
APPENDIX

Definitions

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

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

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:

$$Q^- = \sum_{i=1}^{k} \left( \frac{D_i^2}{S_i^2} \right) \quad \forall D_i > 0$$

$$Q^+ = \sum_{i=1}^{k} \left( \frac{D_i^2}{S_i^2} \right) \quad \forall D_i < 0$$

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:

$$H = \sum_{i=1}^{k} \frac{(D_i - \bar{D})^2}{S_i^2}$$

where
\[
\hat{D} = \left[ \sum_{i=1}^{k} \frac{D_i}{S_i^2} \right] / \left[ \sum_{i=1}^{k} \frac{1}{S_i^2} \right]
\]

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


<table>
<thead>
<tr>
<th>Number of Hospital Diagnoses Collected</th>
<th>Imputation Error Per Imputed Unit Cost Estimate *</th>
<th>% Reduction in Error †</th>
<th>Imputation Error Per Observation *</th>
<th>% Reduction in Error ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>3335</td>
<td>--</td>
<td>3075</td>
<td>--</td>
</tr>
<tr>
<td>12</td>
<td>2595</td>
<td>22.2</td>
<td>2240</td>
<td>27.2</td>
</tr>
<tr>
<td>17</td>
<td>2405</td>
<td>27.9</td>
<td>1920</td>
<td>37.6</td>
</tr>
<tr>
<td>22</td>
<td>2300</td>
<td>31.0</td>
<td>1675</td>
<td>45.5</td>
</tr>
<tr>
<td>27</td>
<td>2240</td>
<td>32.8</td>
<td>1460</td>
<td>52.5</td>
</tr>
<tr>
<td>32</td>
<td>2195</td>
<td>34.2</td>
<td>1240</td>
<td>59.7</td>
</tr>
<tr>
<td>37</td>
<td>2160</td>
<td>35.2</td>
<td>995</td>
<td>67.6</td>
</tr>
<tr>
<td>42</td>
<td>2135</td>
<td>36.0</td>
<td>700</td>
<td>77.2</td>
</tr>
</tbody>
</table>

* Error rounded to nearest 5
† (3335 - Error) / 3335
‡ (3075 - Error) / 3075
ALTERNATIVE EXAMPLE:
Net Monetary Benefit Analysis in 4S (assuming $75,000 per Additional Survivor)

<table>
<thead>
<tr>
<th>Country</th>
<th>Placebo N</th>
<th>MB0</th>
<th>Active N</th>
<th>MB1</th>
<th>NMB</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>355</td>
<td>60448</td>
<td>358</td>
<td>61656</td>
<td>1208</td>
<td>-2648 to 5066</td>
</tr>
<tr>
<td>Finland</td>
<td>433</td>
<td>66811</td>
<td>435</td>
<td>65961</td>
<td>-850</td>
<td>-3414 to 1714</td>
</tr>
<tr>
<td>Iceland</td>
<td>79</td>
<td>64938</td>
<td>78</td>
<td>64301</td>
<td>-637</td>
<td>-7359 to 6086</td>
</tr>
<tr>
<td>Norway</td>
<td>511</td>
<td>62008</td>
<td>514</td>
<td>63379</td>
<td>1371</td>
<td>-1546 to 4288</td>
</tr>
<tr>
<td>Sweden</td>
<td>845</td>
<td>63148</td>
<td>836</td>
<td>63690</td>
<td>542</td>
<td>-1626 to 2711</td>
</tr>
</tbody>
</table>

TESTS FOR INTERACTION

ANOVA (interaction term) P = 0.85
Quantitative Interaction
  Gail and Simon: (H = 1.62; D.F.=4) P = 0.81
Qualitative Interaction
  Gail and Simon: (Q = Min[Q-, Q+]) = Min (1.47; 0.46) P = 0.72
Piantadosi and Gail: Since all country-specific 95% CI overlap, the range test would not be significant for any threshold value. Use of the 95% CI is conservative because the range test would be based on the wider 97.5% CI.
Relative Impact of Medical Service Use and Price Weights on Differences in Costs

<table>
<thead>
<tr>
<th>Country</th>
<th>Experiment #2A</th>
<th>Experiment #2B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed Mean SD</td>
<td>Randomize Obs to Pr Mean SD</td>
</tr>
<tr>
<td>Pooled</td>
<td>3698 1129</td>
<td>4434 1435</td>
</tr>
<tr>
<td>EU1</td>
<td>3792 2307</td>
<td>3962 2743</td>
</tr>
<tr>
<td>EU2</td>
<td>4575 1756</td>
<td>4554 2295</td>
</tr>
<tr>
<td>EU3</td>
<td>6222 1733</td>
<td>4432 1793</td>
</tr>
<tr>
<td>EU4</td>
<td>4628 2505</td>
<td>4763 2201</td>
</tr>
<tr>
<td>EU5</td>
<td>-1871 3131</td>
<td>4409 6406</td>
</tr>
<tr>
<td>Max-Min</td>
<td>8093 3573</td>
<td>801 3520</td>
</tr>
<tr>
<td>Pooled</td>
<td>715 1351</td>
<td>929 1390</td>
</tr>
<tr>
<td>DC1</td>
<td>-1135 1824</td>
<td>-324 1866</td>
</tr>
<tr>
<td>EU6</td>
<td>491 2360</td>
<td>1355 3102</td>
</tr>
<tr>
<td>EU7</td>
<td>-1475 2891</td>
<td>1413 3963</td>
</tr>
<tr>
<td>NA1</td>
<td>1416 1897</td>
<td>1034 1911</td>
</tr>
<tr>
<td>Max-Min</td>
<td>2891 3395</td>
<td>1737 4379</td>
</tr>
</tbody>
</table>