Transferability of Economic Evaluations in Clinical Trials

Henry Glick
Institutt for helseledelse og helseøkonomi
November 25, 2008

The Problem

- Multicenter and multinational studies are the norm for the evaluation of new medical therapies
- They speed the development process, broaden the representativeness of the patients who receive the therapy, and familiarize clinical decision makers with the therapy prior to its approval by regulatory agencies
- They can also increase the heterogeneity of patients, disease presentation and severity, medical service use, price weights, and the like that are observed in the study
- In this lecture focus on these issues as they arise in multinational studies

The Concern

- The presence of between-country heterogeneity in studies has led to a growing concern that the pooled or average economic results from multinational studies may not be reflective of the results that would be observed in individual countries that participated in the study
  - Referred to as "transferability" or "generalizability"
- Common sources for concern include transnational differences in morbidity/mortality patterns, practice patterns, 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 are evaluated in multinational studies
Morbidity / Morality Patterns

- Cost-effectiveness can vary due to variations in underlying morbidity/mortality patterns
- For example, the cost-effectiveness of cholesterol-lowering agents may vary if the underlying risk for coronary heart disease differs
- Potential sources of variation include:
  - Naturally occurring differences
  - Differences in disease identification and treatment
  - Variation in severities of disease of enrolled participants between countries

Practice Patterns (Medical Service Use)

- Effectiveness of a therapy may be related to the other care received by participants
- If the therapy is complementary with other supportive care, i.e., if it is more effective when combined with other supportive care, it will be more cost-effective in countries that provide more supportive care and less cost-effective in countries that provide less of it
- If the therapy substitutes for other supportive care, i.e., if it is more effective when this other care is absent, the reverse will be true

Relative Prices (Price Weights)

- One of the most common concerns
- When addressing transferability of data from other countries, Australian, Canadian, and British guidelines for cost-effectiveness all request that local price weights be used in evaluations submitted to the government, and at least one has a preferred schedule for such submissions
- However, there is little evidence about the impact of simply valuing study-wide medical service use by use of country-specific price weights
But Can't Repeat Studies Everywhere!

• As O'Brien has noted, it is impossible to replicate studies in all settings where therapy choices are made

• There thus is a need to use the data we collect in multinational studies to provide decision makers in individual countries with information they can use in making adoption decisions

Outline

• Relative Impact of medical service use or price weights on transferability

• Three analytic methods for evaluating transferability
  – Tests of homogeneity
  – Fixed effect models
  – Random effect models

• Limitations common to all 3 methods

Importance of Price Weights?

• Common source of concern for transferability

• When generalizing among countries with similar levels of economic development, not clear how important this issue is

• In their study of the impact of price weights in a European stroke study, Willke et al. found little variability in cost-effectiveness ratios when one multiplied study-wide resource use by 5 different countries' price weights
Commonly Proposed, Potentially Inadequate Solutions

- Study-wide clinical results, study-wide medical service use, and price weights from a single country
- Study-wide clinical results, and country-specific medical service use and price weights
- Both approaches ignore the fact that clinical and economic outcomes may influence one another
  - Differences in cost may affect practice patterns, which in turn may affect outcome
  - Differences in practice pattern may affect outcome, which in turn may affect cost

Relative Impact of Medical Service Use and Price Weights

<table>
<thead>
<tr>
<th>Country</th>
<th>Trial-Wide Effect</th>
<th>Country-Specific Cost and Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Country-Price Weights</td>
<td>Country-Specific Cost</td>
</tr>
<tr>
<td>1</td>
<td>46,800</td>
<td>5,900</td>
</tr>
<tr>
<td>2</td>
<td>53,900</td>
<td>90,500</td>
</tr>
<tr>
<td>3</td>
<td>57,600</td>
<td>91,900</td>
</tr>
<tr>
<td>4</td>
<td>65,800</td>
<td>**</td>
</tr>
<tr>
<td>5</td>
<td>69,100</td>
<td>93,300</td>
</tr>
<tr>
<td>Overall</td>
<td>45,900</td>
<td>45,900</td>
</tr>
</tbody>
</table>

Source: Willke et al.
† Country-specific resource use × Country-specific price weights
** New therapy dominates

Barbieri et al.

- Barbieri et al. attempted to identify the major causes of variation in study results between countries.
- Conducted a literature search to identify economic evaluations of pharmaceuticals that were conducted in two or more European countries
Barbieri (II)
- Classified studies by whether price weights, price weights and medical service use, or price weights, medical service use, and clinical effect were allowed to vary in the analysis
- Also classified the studies by the level of variability of the results between countries
  - "Low variability": observed differences in incremental cost-effectiveness ratios were unlikely to change adoption decisions
  - "High variability": observed differences in incremental cost-effectiveness ratios were likely to change adoption decisions

Barbieri (III)

<table>
<thead>
<tr>
<th>PW only</th>
<th>PW &amp; MSU</th>
<th>PW, MSU, &amp; CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

PW = Price weight; MSU = Medical service use; CO = Clinical outcome

Barbieri Conclusions
- The authors concluded that the amount of variation in study results that was observed across countries depended upon the amount of variation that was allowed by the studies' authors
- A key factor they identified was whether the authors of study-based studies allowed medical service use to vary across countries
- Not clear if the conclusions of either the Willke or Barbieri studies apply when transferring between developed and developing countries
Potential Weakness of Willke and Barbieri Studies

- The different methods for estimating cost-effectiveness use different sample sizes
  - e.g., Study-wide medical service use and country-specific price weights use the entire sample; country-specific medical service use and price weights use the country-specific sample
- Whenever we subdivide a larger group into smaller groups, we should expect to observe increases in the variability in the point estimates, even if medical service use were identical in all countries

Impact of Differing Sample Sizes

- When countries’ medical service use has the same mean and variance and differ in their price weights alone, the expected difference in variability associated with use of country-specific rather than study-wide medical service use equals \( N_0^{0.5} / N_i^{0.5} \) (\( N_i \) represents the total number of participants in the study and \( N_i \) represents the number of participants in country \( i \))
- If there are 100 participants in each of two countries, the standard errors for the estimates derived from country-specific medical service use and price weights would differ by 41%
- If the mean and variances differ, the difference can be greater

Summary, Relative Impact

- If we are working in countries with similar economic conditions (e.g., Western Europe), multiplication of study-wide medical service use times a country’s price weights tends to shrink estimates for all countries towards a common result
- Whether or not medical service use matters more than price weights is a matter of conjecture, although the current evidence does not rule out this possibility
- Better evaluations, which overcome the problem of the different sample sizes, are needed before we can draw reliable conclusions
Analytic Approaches to Transferability

- Address two sets of objectives:
  - Evaluate whether there is evidence of heterogeneity across countries (or jurisdictions) and to explore potential sources of heterogeneity
  - Estimate measures of cost-effectiveness that are appropriate for decision making within a particular country
- The use of decision analytic models will continue to be a mainstay
- However, statistical modeling of individual patient-level data from multinational studies represents a growing body of literature that addresses transferability of clinical and economic information from multinational trials

Three Statistical Methods

- Three methods
  - Tests of homogeneity
  - Fixed effects model
  - Multilevel, or hierarchical, models
- They share the common approach of using country-specific data for the estimates, although the random effects models also borrow information from the pooled study result to overcome the problem of the expected increase in variability that one observes when subsetting the data by country

Detection of Heterogeneity

- The identification of heterogeneity in measures of cost-effectiveness across countries participating in a multinational study can be handled in the same manner as the identification of heterogeneity in the clinical measures from the trial
- One approach involves tests of qualitative and quantitative interactions developed by Gail and Simon
  - where a quantitative interaction represents a situation where the treatment effect is consistent in direct but not magnitude
  - a qualitative interaction represents a situation where the direction of the treatment effect is different
Detection of Heterogeneity (II)

- Heterogeneity tests should be applied to a treatment's net benefits
  - Conclusions about heterogeneity should not be based on separate heterogeneity tests of cost and effect

Why Not Separate Tests for Cost and Effect?

- Homogeneity of a therapy’s effects on 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
  - 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 and 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{adj}} )</th>
<th>OR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial #1</td>
<td>( 0.2 )</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>( 0.1 )</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>( &lt;0.05 )</td>
<td>( &lt;1.0 )</td>
</tr>
<tr>
<td>Trial #2</td>
<td>( 0.2 )</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>( 0.1 )</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>( &lt;0.05 )</td>
<td>( &lt;0.05 )</td>
</tr>
</tbody>
</table>

- Will the country-specific absolute differences in be more similar to one another in trials #1 or #2?
Absolute Difference the Same When OR Different

<table>
<thead>
<tr>
<th>Country</th>
<th>( p_{\text{iloc}} )</th>
<th>OR</th>
<th>( p_{\text{act}} ) †</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial #1</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>Trial #2</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>

† \( p_{\text{act}} = (p_{\text{iloc}} \times \text{OR}) / ((p_{\text{iloc}} \times \text{OR}) + (1-p_{\text{iloc}})) \)

Interpreting the Results

- Conclusion from test limited to “Failure to reject homogeneity”
  - Can’t CONCLUDE “Homogeneity”
- If there is no evidence of heterogeneity, and if the test is powerful enough to rule out economically meaningful differences, it is recommended that the pooled or fixed treatment effect across all countries can be used to summarize the study’s results
- If the test reveals evidence of heterogeneity, a pooled estimate of cost-effectiveness is not transferable across countries and should not be used to represent the trial-wide results

Limitations

- Tests of homogeneity are typically underpowered, especially when relatively small numbers of patients are enrolled in individual countries
  - Can be partially addressed by combining patients across a subset of countries with similar practice characteristics (e.g., types of centers and providers or characteristics of reimbursement systems)
  - Can also be addressed by using larger \( \alpha \) levels (e.g., 0.1 or 0.2)
- They don’t offer a natural extension to generate country-specific estimates of cost-effectiveness
Heart Failure Example

- Data derived from a randomized, double-blinded, placebo-controlled study evaluating a drug for severe heart failure (usual care plus active intervention versus usual care plus placebo)
- A total of 1663 participants enrolled in 16 countries were used in our analysis
- We separately analyzed four countries that enrolled more than 100 participants and for which price weights were available (Ns = 130, 372, 382, and 254) and a miscellaneous category made up of the remaining 12 countries

Cost Data

- Cost was estimated for hospitalization, active drug therapy, and ambulatory care
- Price weights for hospitalization were obtained from four countries that enrolled more than 100 participants, and from one that enrolled fewer
  - We used the average of price weights collected from four developed countries to value medical service use in seven developed countries for which they were unavailable
  - We used price weights collected from one developing country to value medical service use in the three developing countries for which they were unavailable

Analysis

- We report arithmetic mean cost, QALYs, and NMB plus 95% confidence intervals
  - CI estimated by use of nonparametric bootstrap analyses for both the pooled and country-specific data
- We report a pooled result, results for the four countries that enrolled more than 100 participants and for which price weight data were available, and for the miscellaneous category
- For NMB, we calculated point estimates and 95% confidence intervals for two values of willingness to pay, 20,000 and 80,000
- We also report statistical tests of the homogeneity
Pooled and Country-Specific Estimates of Incremental Cost and QALYS

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean</th>
<th>95% CI</th>
<th>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>-932</td>
<td>-2647 to 654</td>
<td>0.08</td>
<td>0.01 to 0.16</td>
</tr>
<tr>
<td>C2</td>
<td>-802</td>
<td>-4895 to 2872</td>
<td>0.09</td>
<td>-0.02 to 0.22</td>
</tr>
<tr>
<td>C3</td>
<td>-452</td>
<td>-3195 to 2606</td>
<td>0.15</td>
<td>0.05 to 0.25</td>
</tr>
<tr>
<td>C4</td>
<td>-2457</td>
<td>-6056 to 945</td>
<td>0.06</td>
<td>-0.03 to 0.14</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>

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 for Selected Ceiling Ratios

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean</th>
<th>95% CI</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>2429</td>
<td>1157 to 3793</td>
<td>7174</td>
<td>4171 to 10,389</td>
</tr>
<tr>
<td>C1</td>
<td>2509</td>
<td>400 to 4743</td>
<td>7242</td>
<td>1118 to 13,094</td>
</tr>
<tr>
<td>C2</td>
<td>2570</td>
<td>-1959 to 7432</td>
<td>7877</td>
<td>-1876 to 19,292</td>
</tr>
<tr>
<td>C3</td>
<td>3407</td>
<td>-176 to 7029</td>
<td>12,270</td>
<td>3778 to 21,382</td>
</tr>
<tr>
<td>C4</td>
<td>3583</td>
<td>26 to 7445</td>
<td>6960</td>
<td>-557 to 14,595</td>
</tr>
<tr>
<td>Other</td>
<td>556</td>
<td>-1386 to 2461</td>
<td>3072</td>
<td>-1639 to 7888</td>
</tr>
</tbody>
</table>

Tests for interaction

Qual P = 1.00 p = 1.00
Quan p = 0.51 p = 0.43

Summary

• The tests of homogeneity provided no evidence of either qualitative or quantitative interaction
  – P values ranged from a low of 0.5 for a quantitative interaction for QALYs to a high of 1.0 for a qualitative interaction for QALYs
• Thus, we have no evidence that these outcomes differ between the countries
Limitations Redux

• These results do not rule out a finding of country-by-treatment heterogeneity for NMB
• Depending upon the willingness to pay, the confidence intervals for the pooled result and for two or three of the five countries confirmed this finding. For the remaining one or two countries, the confidence intervals indicated we could not be confident of savings
• If we do not believe the test is sufficiently powered, we may not be impressed by the fact that we cannot reject that the pooled result

Fixed Effect Models

• Simplest means of addressing center effects is to include a fixed effect for country
• Willke et al. used of fixed effects models to separate the direct effect of a study treatment on costs through changes in resource use independent of the patient’s clinical outcome and the indirect effect of a study treatment on costs through changes in the patient’s clinical status
• Through the use of multiple therapy-by-country and country-by-outcome interaction terms, their approach estimates country-specific direct and indirect effects of study treatment on costs

Advantages

• The approach uses patient-level data and standard statistical procedures to estimate country-specific differences of both costs and effects
• It also enables sensitivity analyses, particularly with respect to country-specific treatment outcomes
Limitations

- Fixed effect regression models do not account for the inherently hierarchical data structure in multinational clinical trials
  - Failure to incorporate the clustering that can exist within each country will result in overestimates of the precision and biased standard errors, leading to confidence intervals that are too narrow

Multilevel Models

- Multilevel models account for the hierarchical structure of data
  - i.e., the fact that patients are treated by physicians in centers which means that these data may violate the assumption that the random errors are [identically and] independently distributed

Multilevel Models

- These models have been widely used in fields such as education and health services research
- They have been used more recently to analyze cost data from multinational randomized studies, to estimate incremental net benefits across centers, and to compare medical service use and cost across health care providers
- These models specifically account for the fact that when we subdivide a larger group into subgroups, we expect that the point estimates will vary even when all of the subgroups have been drawn from the same underlying population (i.e., homogeneity)
Multilevel Models (II)

- Multilevel random effects models provide more precise estimates of country-specific results than those resulting from analysis of each country’s data separately and provide unbiased standard errors
  - This occurs because each country’s estimate of costs or cost-effectiveness is shrunk toward the summary estimate by ‘borrowing strength from across countries’
  - The extent of the “shrinkage” is dependent on the variability between country-specific estimates, the variability within a country, and the number of trial participants from the country of interest (more patients, less shrinkage)

Role of Normality

- Analysts have tended to assume normal distributions with a mean and variance specific to each country
- Can relax this assumption:
  - Nixon and Thompson and Thompson et al. have proposed that analysts explicitly address the skewness of cost or cost-effectiveness data in hierarchical or multilevel models
  - Thompson, et al. used generalized linear multilevel models to estimate the effects of patient and national characteristics on costs

Exchangeability

- Usually based on the assumption that the parameters are drawn from a common distribution, or ‘exchangeable’
  - i.e., There are no other a priori reasons why one country may have more or less favorable measures of costs or cost-effectiveness than another
  - Prejuding the transferability of cost-effectiveness results in clinical trials?
- Use of country-level covariates may relaxation of this assumption by use of conditional independence
  - Allows for shrinkage towards two (or more) separate pooled means
Potential Problems Common to All Three Methods

- While these three methods differ in their approach to answering questions of transferability, there are at least five potential problems that are common to all of them:
  - Protocol effects and transferability
  - Influential observations
  - Transferability and censored data
  - Sample size
  - Countries that did not participate in the study

Protocol Effects and Transferability

- All three approaches assume that observed similarities and differences in country-specific results arise naturally
- Study protocols’ attempts to standardize care might be responsible instead
- Should be most confident that the protocol is not artificially driving a finding of transferability when we observe between country variability in medical service use in the placebo group, yet we are able to observe common cost-effectiveness across the countries
- Should be least confident when there is little or no between-country variability in medical service use in the placebo group

Influential Observations

- Evaluation of the pooled average effect of therapy on cost is complicated by the existence of a small number of observations with extreme values for cost
  - The problems these observations pose to the analysis of the pooled effect are only magnified in the analysis of country-specific treatment effect
- These few observations will often be spread, possibly one each, across several countries
  - Their existence will make it difficult for us to determine whether observed homogeneity or country-specific differences are due to these individual observations or whether they are due to underlying differences between the countries
Transferability and Censored Data

• The principle that underlies most methods for addressing censored data is to "borrow" data – either by inverse probability weighting or multiple imputation – from participants who are "like" participants whose data are censored

• Transferability, on the other hand, questions whether the data from participants from one country are representative of data from participants from another country

• If we "borrow" data from different countries to address censoring, will we undermine our ability to detect country-specific differences

Sample Size

• Many multinational studies now enroll participants in 40 or 50 countries, and for many of these countries only 10 or 15 participants are enrolled

• These designs pose problems for all three methods

• For example, when we test homogeneity, small samples in each country will generally yield large country-specific confidence intervals for NMB, which make rejection of homogeneity of the treatment effect difficult

• Similarly, small sample sizes predisposes one towards large amounts of shrinkage to the mean

Sample Size (II)

• One means of overcoming this problem is to modify the study design so that more participants are enrolled in fewer countries

• If it is infeasible to change the study design, we may want to change the level of analysis for the homogeneity test

  – e.g., aggregate across "similar" countries, where similarity might be defined by:
    • Level of economic development
    • Practice pattern
    • Characteristics of the health care reimbursement system
Countries That Did Not Participate in the Study

- None of three methods directly provide information for countries that did not participate in the study
- There has been little published methodologic work for addressing this problem, and without knowing the specifics of any particular country, it is hard to generalize
- Decision analysis is one option for attempting to address this issue
- Another approach may be to identify countries that participated in the study that are similar to those that did not, and use the information from the one to inform decisions for the other

Recommendations

- Issues of transferability are of growing importance as decision makers in more countries seek information about the likely economic impact of the adoption of new therapies in their countries
  - We have tended to do a poor job of estimation of transferability
  - Although we seem to focus most on adjusting price weights, they likely are not the primary source of variability when countries have similar economic conditions

Recommendations (II)

- Three recently proposed analytic techniques provide improved methods for addressing issues of transferability:
  - Tests of homogeneity
  - Fixed effects models
  - Multilevel random effects models
- Each method uses data from the study to address these issues. Our early understanding is that the methods trade-off problems of power for problems of exchangeability
Recommendations (III)

• In many studies, small sample sizes may limit our ability to detect country-specific effects, no matter which method we pursue in the evaluation of transferability

• We may be better served to look at major contrasts that provide some of the information sought by decision makers
  – Does the therapy compliment other supportive care or does it or substitute for other supportive care
  – Is the therapy more or less cost-effective in developed countries than it is in developing countries

• Simply knowing these relationships should improve our understanding of the transferability of the economic results from multinational studies