Designing Economic Evaluations in Clinical Trials

Statistical Considerations in Economic Evaluations

ISPOR 17th Annual International Meeting

June 3, 2012

Henry Glick and Jalpa Doshi
www.uphs.upenn.edu/dgimhsr

Good Value for the Cost

- Economic data collected as primary or secondary endpoints in randomized trials are commonly used in the evaluation of the value for the cost of medical therapies
  - Short-term economic impacts directly observed
  - Longer term impacts potentially projected by use of decision analysis
  - Reported results: point estimates and confidence intervals for estimates of:
    - Incremental costs and outcomes
    - Comparison of costs and effects
  - Impact of sensitivity analysis judged by its impact on both the point estimates and the confidence intervals of the ratios

Example

Analysis	Point Estimate	95% CI
Incremental Cost	-713	-2123 to 783
Incremental QALYs	0.13	0.07 to 0.18
Cost-Effectiveness Analysis		
Principal Analysis	Dominates	Dom to 6650
Survival Benefit		
-33%	Dominates	Dom to 9050
+33%	Dominates	Dom to 5800
Hospitalization Cost		
-50%	Dominates	Dom to 5300
+50%	Dominates	Dom to 8400
Drug Cost		
-50%	Dominates	Dom to 4850
+50%	Dominates	Dom to 8750
Discount rage		
0%	Dominates	Dom to 6350
7%	Dominates	Dom to 7000

Outline

- Steps in economic evaluation
- The gold standard and its tensions
- 4 strategic issues
  - What medical service use should we collect?
  - How should we value medical service use
  - What is the appropriate sample size?
  - How should we interpret results from multicenter studies?

Steps in Economic Evaluation

Step 1: Quantify the costs of care

Step 2: Quantify outcomes

Step 3: Assess whether and by how much average costs and outcomes differ among the treatment groups

Step 4: Compare magnitude of difference in costs and outcomes and evaluate "value for costs" (e.g. by reporting a cost effectiveness ratio or the probability that the ratio is acceptable

- Potential hypothesis: The cost per quality-adjusted life year saved is significantly less than $60,000

Step 5: Perform sensitivity analysis

Ideal Economic Evaluation Within a Trial

- Conducted in naturalistic settings
  - Compares the therapy with other commonly used therapies
  - Studies the therapy as it would be used in usual care
- Well powered for:
  - Average effects
  - Subgroup effects
- Designed with an adequate length of follow-up
  - Allows the assessment of the full impact of the therapy
- Timely
  - Can inform important decisions in the adoption and dissemination of the therapy

Ideal Economic Evaluation Within a Trial (II)

- Measures all costs of all participants prior to randomization and for the duration of follow-up
  - Costs after randomization—cost outcome
  - Costs prior to randomization—potential predictor
- Independent of the reasons for the costs
- Most feasible when:
  - Easy to identify when services are provided
  - Service/cost data already being collected
  - Ready access to data

Design Issues Not Unique To Trials

- A number of design issues apply equally to economic evaluations that are incorporated within clinical trials and to other economic evaluations:
  - The type of analysis that will be conducted (e.g. cost-benefit, cost-effectiveness, or cost minimization analysis)
  - The types of costs that will be included (e.g. direct medical, direct nonmedical, productivity, and intangible)
  - The perspective from which the study will be conducted
- These issues have been well addressed in the literature

Difficulties Achieving an Ideal Evaluation

- Settings often controlled
- Comparator isn't always the most commonly used therapy or the currently most cost-effective
- Investigators haven't always fully learned how to use the new therapy under study
- Sample size required to answer economic questions may be greater than sample size required for clinical questions
- Ideal length of follow-up needed to answer economic questions may be longer than follow-up needed to answer clinical questions

Trade-off

- These trials may be the only source of information needed for important early decisions about the adoption and diffusion of the therapy
- TRADE-OFF: Ideal vs best feasible

Issue #1: What Medical Service Use Should We Collect?

- Real/perceived problems
  1. Don't have sufficient resources to track all medical service use
  2. Don't expect to affect all medical service use, just that related to the disease in question
- Implication: given sample size in trial, collection of all medical services, independent of the reason for these services, may swamp the "signal" with "noise"

  → Why not limit data to disease-related services?

Limited Data Collection Resources

- Availability of administrative data may reduce costs related to tracking all medical service use
- If administrative data are unavailable:
  - Measure services that make up a large portion of the difference in treatment between patients randomized to the different therapies under study
    - Provides an estimate of the cost impact of the therapy
  - Measure services that make up a large portion of the total bill
    - Minimizing unmeasured services reduces the likelihood that differences among them will lead to biased estimates
    - Provides a measure of overall variability

Measure as Much as Possible

- Best approach: measure as many services as possible
  - No a priori guidelines about how much data are enough
  - Little or no data on the incremental value of specific items in the economic case report form

Document Likely Service Use During Trial Design

- Decisions improved by documenting types of services used by patients who are similar to those who will be enrolled in the trial
  - Review medical charts or administrative data sets
  - Survey patients and experts about the kinds of care received
  - Have patients keep logs of their health care resource use
- Guard against possibility that new therapy will induce medical service use that differs from current medical service use

Account for Data Collection Expense

- Decisions about the services to measure should take into account the expense of collecting particular data items
  - e.g., frequently performed, low cost items?
    - 6,700 blood gas tests equaled 1.8% of procedure and diagnostic test costs
    - 420 angiocardiopneumographies equaled 4.3%

Limit Data to Disease-Related Services?

- Little if any evidence exists about the accuracy, reliability, or validity of such judgments
- Easy for judgments to be flawed

Limit Data to Disease-Related Services (II)

- Investigators routinely attribute AEs to the intervention, even when participants received vehicle/placebo
- Medical practice often multifactorial: modifying disease in one body system may affect disease in another body system
  - In the Studies of Left Ventricular Dysfunction, hospitalizations "for heart failure" and death reduced by 30% ($p<0.0001$)
  - Hospitalizations for noncardiovascular reasons reduced 14% ($p = 0.006$)
- If a patient has an automobile accident, how does the clinician determine whether or not it was due to a hypotensive event caused by therapy?

Limit Data to Disease-Related Services (III)

- Potential biases more of a problem in unblinded studies, but need not "balance out" in double-blinded studies

Other Types of Costs?

- Other types of costs that sometimes are documented within economic evaluations include:
  - Time costs: Lost due to illness or to treatment
  - Intangible costs
- Types of costs that should be included in an analysis depend on:
  - What is affected by illness and its treatment
  - What is of interest to decision makers
    - e.g., the National Institute for Clinical Excellence (U.K.) and the Australian Pharmaceutical Benefits Scheme has indicated they are not interested in time costs

General Recommendations

- General Strategy: Identify a set of medical services for collection, and assess them any time they are used, independent of the reason for their use
- Decision to collect service use independent of their reason for use does not preclude ADDITIONAL analyses testing whether designated “disease-related” costs differ

General Recommendations (2)

- If data collection is limited to a single page in the CRF:
  - First impression: Collect big-ticket items, (e.g., hospitalization, long term care, etc); don't sweat smaller ticket items
    - Heart failure: hospitalization costs, number of outpatient visits
    - Hospitalized infections: ICU, stepdown, and routine care days; major procedures
    - Asthma: ER visits, Hospitalizations, comedications

Better Approach

- Prior to the study, invest in determining which services will likely make up a large portion of the difference in costs between the treatment groups
  - If the therapy is likely to affect the number of hospitalizations, collect information that will provide a reliable estimate of the cost of these hospitalizations
  - If the therapy is likely to affect days in the hospital and location in the hospital, collect this information
  - If the therapy is principally likely to affect outpatient care, collect measures of outpatient care, etc.

Specific Recommendations, Which Services?

- Identify common patterns of medical service use in countries that will participate in the trials
  - Speak with experts in multiple countries
  - Focus groups, etc.
- Design case report forms to collect important, common medical service use
- Collect the services independent of the reason for their use
- Pilot test forms (if appropriate, in multiple countries)
- Consider collecting costs other than medical service use

Issue #2. How Should We Value Medical Service Use?

- Availability of billing data may simply valuation
- If billing data aren't available, collect price weights for a selected set of medical services from a selected set of countries
  - For international studies, most often derived from a national data or a single center per country
- Sample sources of data:

http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Managingyourorganisation/Financeandplanning/NHScostingmanual/index.htm

Outpatient: http://www.cms.hhs.gov/Medicare/Medicare-Fee-for-Service - Payment/FeeScheduleGenInfo/ index.html?redirect=/FeeSchedule GenInfo/

Inpatient: http://www.cms.hhs.gov/Medicare/Medicare-Fee-for-Service-Payment/ProspMedicareFeeSvcPmtGen/index.html?redirect=/ProspMedicareFeeSvcPmtGen

Price Weights from Which Centers / Countries

- The centers/countries from which price weights are collected might be ones:
  - That enroll a large number of patients
  - That represent the spectrum of economic conditions
  - In which regulators require a submission
  - For which price weights are readily available
  - In which the sponsor wishes to make economic claims

Estimating Missing Price Weights

- Eventually, we will need to identify price weights for all medical services recorded in the case report form
- Because collecting price weights for all services may be expensive, we commonly:
  - Collect price weights for service use that:
    - Occurs most frequently in the trial
    - Is considered likely to be affected by the intervention
    - Has particularly high or low costs
  - Develop a method of imputation to estimate price weights that haven't been collected

More / Fewer Countries or the Reverse?

- Presuming we are using a reliable method for imputing price weights (e.g. DRG weights), do we know anything about how we should trade-off number of centers/ countries sampled versus number of price weights per center/country?

More / Fewer Countries or the Reverse (II)

- In simulations based on data from 4 countries:
  - If the number of price weights we plan to collect is fixed:
    - Better to sample a smaller number of price weights in more centers than to sample a larger number of price weights in fewer centers
    - e.g., in simulations the imputation error was smaller when 12 price weights were collected in each of 4 countries than when 47 were collected in a single country

Glick HA, Orzol SM, Tooley JF, Polsky D, Mauskopf JM. Design and Analysis of Unit Cost Estimation Studies: How Many Hospital Diagnoses? How Many Countries? Health Economics. 2003;12:517-27.

Center/Country-Specific vs Averaged Price Weights

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

Center/Country-Specific vs Averaged Price Weights (II)

- Ideal: Because relative prices can affect quantities of services provided, where ever feasible, multiply country-specific price weights times times country-specific counts of medical services
- For countries for which price weights aren't available:
  - Use (averages of) price weights from similar countries
  - e.g., in a trial that enrolls patients in Western and Eastern Europe and Latin America, we might average price weights from other Western European countries to value service use in Germany, but wouldn't want to use this average for Eastern Europe or Latin America

Center/Country-Specific vs Averaged Price Weights (III)

- Corollary: If we have a set of price weights for each country that participated in the trial, we should not average them and use this average for all services measured in the trial
  - The most common reasons suggested for such a strategy are that
    - Reducing variability in the price weights reduces variability in the estimated costs and
    - An average set of price weights may be more representative

Center/Country-Specific vs Averaged Price Weights (IV)

  - However:
    - Empirically, use of a single set of price weights need not reduce variance
    - If substitution effects are strong, this strategy may introduce bias in the estimates of cost differences
    - Why is it more “representative” to use a set of price weights that no one faces?

What is the appropriate sample size?

- Sample size and power calculations allow us to conduct experiments with an expected likelihood that at the conclusion of the experiment we will be able to be confident in the resulting comparison of costs and effects
  - e.g., May hypothesize that the point estimate for the cost-effectiveness ratio will be 20,000 per quality-adjusted life year (QALY)
  - May want to identify a sample size that will provide an 80% chance (i.e., power) to be 95% confident that the therapy is good value when we are willing to pay at most 75,000 per QALY

Sample Size Formula, Common SDs

- Assuming equal standard deviations for cost and effect and equal sample sizes, the sample size formula is:

$$n = \frac{2\left(z_{\alpha}+z_{\beta}\right)^{2}\left(sd_{c}^{2} + \left(W\ sd_{q}\right)^{2} - \left(2\ W\ \rho\ sd_{c}\ sd_{q}\right)\right)}{\left(W\Delta Q - \Delta C\right)^{2}}$$

where n = sample size/group; z_{α} and z_{β} = z-statistics for α (e.g., 1.96) and β (e.g., 0.84) errors; sd = standard deviation for cost (c) and effect (q); W = maximum willingness to pay we wish to rule out; and ρ = correlation of the difference in cost and effect

www.uphs.upenn.edu/dgimhsr/stat-samps.htm

Similarities With Clinical Sample Size Formulas

Error Rates

Variance

$$n = \frac{2 \quad (z_\alpha + z_\beta)^2 \quad \left(sd_c^2 + (W^2 \; sd_q^2) - (2 \; W \; \rho \; sd_c \; sd_q)\right)}{\Delta NMB^2}$$

$$n = \frac{2 \quad (z_\alpha + z_\beta)^2 \quad \left(\qquad sd_q^2 \qquad \right)}{\Delta Q^2}$$

Difference^2

Differences in Formulas

$$Var_{NMB} = sd_c^2 + (W^2 \; sd_q^2) - (2 \; W \; \rho \; sd_c \; sd_q)$$

- Variance of NMB more complicated than variance for usual continuous clinical differences
  - Includes ρ, the correlation of the difference between cost and effect
  - Includes W, the decision threshold we are trying to rule out

Correlation

- The correlation of the difference in cost and effect indicates how changes in the difference in cost are related to changes in the difference in effect
  - Negative (win/win) correlation: increasing effects are associated with decreasing costs
    - e.g., asthma care
  - Positive (win/lose) correlation: increasing effects are associated with increasing costs
    - e.g., life-saving care
  - All else equal, fewer patients need to be enrolled when therapies are characterized by a positive correlation than when they are characterized by negative correlation

Ability to Shift W

- W is to cost-effectiveness analysis as 1 is to OR and RR
  - It is the decision threshold we are trying to rule out if we are to have confidence about value
- While we rarely consider comparing OR and RR to a decision threshold other than 1 (noninferiority trials may be the exception), we often choose W because there is no clear consensus on what its value is
- Moving W "nearer to" or "further away from" the expected point estimate reduces or increases the power we have to be confident of value
- Caution: "Nearer" and "further away" are not measured on the real number line

  → Sample size need NOT decrease as WTP increases

Power Formula, Common SDs

- Assuming equal standard deviations for cost and effect and equal sample sizes, the power formula is::

$$z_\beta = \sqrt{\frac{n * \left(W\Delta Q - \Delta C\right)^2}{2\left(sd_c^2 + \left(W\, sd_q\right)^2 - \left(2\, W\, \rho\, sd_c\, sd_q\right)\right)}} - z_\alpha$$

- Unlike sample size equation where result = N, result of formula is z_β, not power
- To estimate power, use the normal distribution table to identify the fraction of the tail that is to the left of z_β
  - Stata code: power = norm(zbeta)
  - E.g., -1.96 = 2.5% power; -0.84 = 20% power; 0 = 50% power; .84 = 80% power; 1.28 = 90%

"Typical" Sample Size Table, W

	Sample Size Per Group
WTP	Exp 1 *
20,000	321
30,000	273
50,000	234
75,000	214
100,000	204
150,000	194

* ΔC=-120; ΔQ=0.015; sd_c=1000; sd_q=.05; ρ= -.8; α=.05; 1-β=.8

Sample Size Can Increase with Increasing W

	Sample Size Per Group	
WTP	Exp 1	Exp 2 *
20,000	321	36
30,000	273	42
50,000	234	68
75,000	214	92
100,000	204	108
150,000	194	127

* ΔC=-120; ΔQ=0.015; sd_c=1000; sd_q=.05; ρ=0.8; α=.05; 1-β=.8

Sample Size Not Necessarily Monotonic With W

	Sample Size Per Group		
WTP	Exp 1	Exp 2	Exp 3 *
20,000	321	36	178
30,000	273	42	158
50,000	234	68	***151***
75,000	214	92	154
100,000	204	108	156
150,000	194	127	160

* ΔC=-120; ΔQ=0.015; sd_c=1000; sd_q=.05; ρ=0.0; α=.05; 1-β=.8

Where to Obtain the Necessary Data?

- When therapies are already in use: Expected differences in outcomes and standard deviations can be derived from feasibility studies or from records of patients
- Simple correlation between observed costs and effects may be an adequate proxy for the measure of correlation used for estimating sample size
- For novel therapies, information may need to be generated by assumption
  - e.g., sd from usual care will apply to new therapy, etc.

Willingness to Pay and Identification of an Appropriate Outcome Measure

- Sample size calculations require us to stipulate what we are willing to pay to obtain a unit of outcome
- In many medical specialties, researchers use disease specific outcomes
- While we can calculate a cost-effectiveness ratio for any outcome we want (e.g., cost/case detected or cost/additional abstinence day), to be convincing that a new, more costly and more effective therapy is good value, the outcome must be one for which we have recognized benchmarks of cost effectiveness
  - Argues against use of too disease-specific an outcome for economic assessment

Glick HA. Sample size and power for cost-effectiveness analysis (part 1). Pharmacoeconomics. 2011;29;189-98.

Glick HA. Sample size and power for cost-effectiveness analysis (part 2). The effect of maximum willingness to pay. Pharmacoeconomics. 2011;29:287-96.

Issue #4. How Should We Interpret Results From Multicenter (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 this 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

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
- Ignore the fact that clinical and economic outcomes may influence one another (cost affects practice which affects outcome; practice affects outcome which affects cost)

Impact of Price Weights vs Other Variation

Country	Trial-Wide Effects		Country-Specific Costs and Effects
	Price weight	Country-Specific Costs	
1	46,818	5921	11,450
2	57,636	91,906	60,358
3	53,891	90,487	244,133
4	69,145	93,326	181,259
5	65,800	**	**
Overall	45,892	45,892	45,892

* Willke RJ, et al. Health Economics. 1998;7:481-93
† Country-specific resource use × Country-specific price weights
** New therapy dominates

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

Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, Reed S, Rutten F, Sculpher M, Severens J. Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Practices ResearchTask Force Report. Value in Health. 2009;12:409-18.

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 we believe the test was powerful enough to rule out economically meaningful differences in costs, then we 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 we 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

Estimation

- Multi-level random-effects model shrinkage estimation assesses whether observed differences between countries are likely to have arisen simply because we have divided the trial-wide sample into subsets or whether they are likely to have arisen due to systematic differences between countries
  - Borrows 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

Summary

- Clinical trials may provide the best opportunity for developing information about a medical therapy's value for the cost early in its product life
- When appropriate types of data are collected and when they are analyzed appropriately, these evaluations can provide data about uncertainties related to the assessment of the value for the cost of new therapies that may be used by policy makers, drug manufacturers, health care providers and patients when the therapy is first introduced in the market