Statistical considerations in economic evaluations: a guide for cardiologists

Shelby D. Reed*

Duke Clinical Research Institute and Department of Medicine, Duke University School of Medicine, PO Box 17969, Durham, NC 27715, USA

Received 15 November 2013; revised 19 March 2014; accepted 2 April 2014; online publish-ahead-of-print 6 May 2014

The author reviews statistical methods commonly applied in economic evaluations that rely on individual patient-level data. The paper includes a review of foundational concepts, unique characteristics of health economic data, and methods developed to address them. The paper then highlights issues that should be considered in the interpretation of findings from economic evaluations.

Keywords
Cost–benefit analysis • Costs and cost analysis

Introduction

As the need for high-quality evidence on the value of healthcare interventions has increased, economic evaluations have grown in sophistication. It is also now common to see health economic outcomes integrated into protocols for clinical trials, prospective registries, and retrospective studies of secondary data, all of which include individual patient-level data on costs and health outcomes. The availability of such data allows health economists to draw on conventional and innovative statistical methods to make inferences to inform decision-makers about the relative value of new medical interventions.

The aim of this paper was to provide non-statistician clinicians with a practically oriented review of statistical methods commonly applied in economic evaluations that rely on individual patient-level data. The paper reviews some foundational concepts, the unique characteristics of these data, and the methods that have been developed to address them. The paper then highlights issues that should be considered in the interpretation of findings from economic evaluations.

Components of an economic evaluation

Economic evaluations that use individual patient-level data rely on empirical data representing medical resource use, costs, and health outcomes typically collected within a randomized clinical trial. Economic evaluations can be limited to comparisons of resource use and costs. In cases when an intervention is associated with significantly lower medical resource use or costs, the analysis may be sufficient to inform clinical or policy decisions. However, when there is a net increase in costs after accounting for intervention-related costs, most analysts will incorporate comparisons of health outcomes in a formal cost-effectiveness analysis wherein health outcomes are typically measured in terms of survival or quality-adjusted survival. There are a myriad of statistical issues associated with analysing medical resource use, costs, and long-term health outcomes. When these components are combined in a cost-effectiveness analysis, additional analytical issues arise.

Distribution issues

Cost data are always non-negative and typically exhibit right-skewness attributable to a small proportion of patients with atypical medical journeys (i.e. outliers). In examinations of total medical costs, zero costs are rare. However, in examinations of specific types of costs, such as inpatient costs, large spikes at zero are common (Figure 1). As costs increase, they typically exhibit greater variation, indicating that the variance is heteroscedastic.

For descriptive purposes, measures of central tendency (i.e. means and medians) should be reported, as should statistics that describe distributional characteristics (e.g. inter-quartile range) and variation (e.g. standard deviation). A median cost estimate represents the cost at the 50th percentile, often considered the cost of the ‘typical’ patient. However, because healthcare budgets must also account for atypical patients (i.e. outliers), mean costs are of central interest and should always be reported in economic evaluations. The same logic applies to health outcomes, for which mean...
estimates are necessary if the aim is to maximize health outcomes within a given budget.

Cost comparisons with complete data

Given the ubiquity of skewed cost distributions, it is commonly suggested that non-parametric statistical tests (e.g. Wilcoxon rank-sum tests) should be used to compare costs between groups. However, non-parametric tests that rely on rank ordering of observations rather than their magnitude are not appropriate for comparing mean costs between treatment groups, the primary statistic of interest. Another common suggestion is to ‘normalize’ costs by log-transforming the data. However, inferences about costs on the transformed scale (e.g. log-euros) are not meaningful to decision-makers, and methods for retransformation to the original monetary scale are frequently biased due to the heteroscedasticity present in cost data. To address concerns about distributional assumptions, many analysts use the non-parametric bootstrap method to compare mean costs between treatment groups. This approach subjects the original data to resampling to generate a sampling distribution of the statistic of interest (i.e. the mean). For each bootstrap sample, the analyst computes the difference in estimated means between treatment groups. Even with relatively small samples, the sampling distribution of the difference in mean costs between groups does not exhibit the skewness or kurtosis that may be present in the raw data unless extreme outliers are present. With the bootstrap method, analysts can examine whether the distribution of the difference in mean costs is normally distributed and correspondingly report $(100 - \alpha\%)$ confidence intervals. When the confidence interval excludes zero, the difference is considered statistically significant.

Although the bootstrap method provides analysts with a transparent approach to comparing mean unadjusted costs, it is often desirable to apply a regression model to adjust for potential imbalances between groups and to generate more precise estimates by accounting for the variance in outcomes that is attributable to prognostic baseline variables. Simulation studies indicate that mean costs can be estimated without bias using ordinary least squares regression to provide valid estimates of incremental costs associated with a given treatment. However, other regression modelling strategies may be more precise, and may thereby generate tighter confidence intervals and smaller $P$-values.

One approach that has become ubiquitous throughout the methodological and applied literature for analysing cost data is the use of generalized linear models. These models estimate how a linear predictor (e.g. treatment group) is associated with the expectation of the dependent variable (e.g. costs) through application of a link function. The most widely applied generalized linear models to cost data are specified with a log-link and variance structure that follows a gamma error distribution in which the variance is assumed to be proportional to the mean. From an interpretative perspective, one must be aware that the exponentiated parameter estimates are multiplicative. For example, an analysis may reveal that a heart failure disease management programme is associated with a 0.8 multiplicative effect (i.e. 20% decrease) on costs. Because its interpretation is relative to the baseline cost (i.e. intercept), the extent to which baseline costs vary across other model parameters is important. If baseline costs are $10,000 per year in patients with New York Heart Association (NYHA) class II heart failure, the programme would be associated with a $2000 decrease in costs. However, among patients with NYHA class IV with mean costs of $30,000 per year, the model predicts a $6000 decrease. By comparison, an ordinary least squares regression model would directly estimate the incremental cost savings with the programme where a constant effect is assumed across NYHA classes unless statistical interactions are included.

Censoring

Unless a limited time horizon is chosen for the analysis or all patients are followed until death, censoring arises with patient-level data. If analysts do not account for censoring, resulting mean cost estimates will be biased because accumulated costs at the time of censoring are usually positively correlated with total costs at death.

Inverse probability weighting has been widely adopted to handle censored cost data. Conceptually, the approach begins by dividing the data into small time intervals. Then, for each time interval over which one wishes to estimate costs, each patient is coded as (i) having cost data available or being dead or (ii) being alive but no longer having cost data available (censored). Each time interval is then inversely weighted according to the probability that patients are not censored during the corresponding interval. When more patients in a time interval are censored, the data from the patients...
with uncensored cost data are weighted more heavily. Then, costs across the time intervals are combined to compute total costs.

Hierarchical data structures
Data used in economic evaluations often represent care patterns and outcomes from multiple clinical sites and countries. Although pooling the data allows investigators to have greater statistical power, the underlying differences in practice patterns, patient characteristics, availability of resources, and unit costs across sites create multiple analytic and interpretative challenges.

Given that coverage decisions are typically made at the local level, methodological efforts have focused on examining whether heterogeneity exists between countries. Early approaches centred on tests of homogeneity or inclusion of fixed effects (for each country or region) in regression models. However, both of these strategies are vulnerable to small samples within individual countries. Thus, a lack of statistical significance should give one little confidence that potentially meaningful differences have been ruled out. More recently, methodologists have turned their focus to hierarchical (or multilevel) models. These models are designed to address statistical correlations that exist between patients enrolled within sites and/or countries owing to similarities in practice patterns, costs for medical supplies and labour, and other factors. As applied to the ASCEND-HF and ASSENT-3 trials, hierarchical models also allow analysts to compute country-specific estimates.

Extrapolate beyond the observed data
For formal cost-effectiveness analyses, complete reliance on observed data is often not possible because a (remaining) lifetime time horizon is typically needed to incorporate expected long-term costs and health outcomes. Therefore, some means of modelling is necessary. When a considerable portion of the survival curve is available from the trial data alone, parametric survival models can be used to extend survival curves beyond the observed follow-up period. A broad range of statistical distributions provides analysts with numerous options, but the choice can significantly impact estimated survival differences.

Other strategies for extrapolating outcomes include ‘age-based’ modelling, wherein a treatment effect is modelled as a function of age instead of a function of time, or application of a series of risk equations estimated from long-term epidemiological studies to predict disease-related complications, costs, and survival.

Regardless of the approach for extrapolation, analysts should provide a rationale for their choice. Because external validation is rarely possible, good practice includes sensitivity analyses to evaluate the extent to which findings change when extrapolation methods are altered.

Measures of cost-effectiveness
Incremental cost-effectiveness ratio
In a cost-effectiveness analysis, in which, the analyst wishes to evaluate trade-offs between costs and benefits, the statistic of interest is the incremental cost-effectiveness ratio (ICER):

\[
\text{ICER} = \frac{\hat{C}_{\text{Tx}} - \hat{C}_{\text{StdCare}}}{\hat{E}_{\text{Tx}} - \hat{E}_{\text{StdCare}}} = \frac{\Delta C}{\Delta E}.
\]

where \(\hat{C}\) and \(\hat{E}\) represent mean costs and mean measures of effectiveness and subscripts \(\text{Tx}\) and \(\text{StdCare}\) represent the treatment of interest and standard care, respectively. Measures of effectiveness can include specific clinical outcomes such as the number of myocardial infarctions averted, but most often represent incremental survival or quality-adjusted survival. Although not without controversy, quality-adjusted life-years (QALYs) are widely recommended for use in cost-effectiveness analyses. Quality-adjusted life-years account for periods of time that patients spend in better vs. worse health states as measured by health utilities on a 0–1 scale, where 0 is equivalent to dead and 1 is equivalent to perfect health (Table 1).

Distributional characteristics of incremental cost-effectiveness ratios
With individual patient-level data on costs and effectiveness, one can make statistical inferences about the ICER. As with cost comparisons, the bootstrap method is a widely used approach. Analysts frequently plot the resulting pairs of estimates (\(\Delta E, \Delta C\)) from the bootstrap samples as a scatterplot on the cost-effectiveness plane (Figure 2). Although these plots provide useful visual images, analysts wishing to report confidence intervals frequently encounter difficulties due to the ICER’s peculiar distributional characteristics.

In most cases, the numerator (\(\Delta C\)) and the denominator (\(\Delta E\)) of the ICER will both be normally distributed, but can include positive and negative numbers. If the two distributions are independent, the ICER is considered to have a Cauchy distribution, which has an

![Figure 2](bootstrap_plot.png) Bootstrap scatterplot on a cost-effectiveness plane. Note: The y-axis represents the difference in mean costs (\(\Delta C\)) between the treatment of interest and standard care groups. The x-axis represents the differences in mean QALYs (\(\Delta QALYs\)) between the treatment of interest and standard care groups. The points in the scatterplot represent pairs of (\(\Delta C\)) and (\(\Delta QALYs\)) generated from 1000 bootstrap samples. The diagonal line corresponds to a hypothetical willingness-to-pay threshold of $50 000 per QALY. Of the 1000 bootstrap samples, 732 fall below this line, demonstrating acceptable cost-effectiveness.
Table 1  Key health economics concepts

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICER</td>
<td>Incremental cost-effectiveness ratio</td>
<td>The primary result of a cost-effectiveness analysis, the ICER represents the ratio of the incremental difference in mean costs between the intervention of interest vs. standard care by the difference in estimated effectiveness between the same two groups: ICER = $\hat{C}<em>{Tx} - \hat{C}</em>{StdCare}$/$\hat{E}<em>{Tx} - \hat{E}</em>{StdCare}$</td>
</tr>
<tr>
<td>QALY</td>
<td>Quality-adjusted life-year</td>
<td>A measure commonly used in cost-effectiveness analysis, the QALY combines survival time with an adjustment for quality of life. One year in perfect health represents 1 QALY, and 1 year in a health state valued at 0.7 on a scale from 0 (death) to 1 (perfect health) is equal to 0.7 QALYs</td>
</tr>
<tr>
<td>WTP</td>
<td>Willingness to pay</td>
<td>WTP is a monetary value that a stakeholder (e.g. society, government, patient) is willing to pay for a given outcome, such as an additional QALY</td>
</tr>
<tr>
<td>INMB</td>
<td>Incremental net monetary benefit</td>
<td>The INMB is the net monetary benefit resulting from adopting the intervention of interest rather than standard care, conditional on the maximum willingness to pay for an additional unit of effectiveness. If the maximum willingness to pay is $35 000 per QALY and the ICER is &lt;$35 000 per QALY, the INMB will be &gt;$0, indicating good value</td>
</tr>
<tr>
<td>INHB</td>
<td>Incremental net health benefit</td>
<td>The INHB is the net health benefit resulting from adopting the intervention of interest rather than standard care, conditional on the maximum willingness to pay for an additional unit of effectiveness. If the maximum willingness to pay is $35 000 per QALY and the ICER is &lt;$35 000 per QALY, the INHB will be &gt;0 QALYs, indicating good value</td>
</tr>
<tr>
<td>CEAC</td>
<td>Cost-effectiveness acceptability curve</td>
<td>A CEAC provides a means to summarize uncertainty associated with the results of a cost-effectiveness analysis. Because maximum WTP thresholds are generally not firmly established, points along the CEAC correspond to the (Bayesian) probability that the intervention of interest is cost-effective (compared with standard care) across a range of WTP thresholds</td>
</tr>
</tbody>
</table>

Net benefits

Stinnett and Mullahy\(^{19}\) suggested circumventing these issues with a ‘net benefits’ approach in which the ICER is ‘linearized’ by the addition of the cost-effectiveness (or willingness-to-pay) threshold, $\lambda$, which represents the maximum amount of money one would pay for an additional unit of effectiveness. Incremental net benefits (INB) can be computed as monetary (INMB) or health benefits (INHB):

$$\text{INMB} = \hat{\Delta}E\lambda - \hat{\Delta}C \text{ or } \text{INMB} = \hat{\Delta}E - \frac{\hat{\Delta}C}{\lambda}.$$  

Based on point estimates, when INB $\geq 0$, the intervention of interest is considered cost-effective relative to standard care, and when INB $<0$, the intervention is not cost-effective. In addition to its simple application to decision rules (i.e. cost-effective vs. not cost-effective), the variance of the INB statistic can be readily calculated, and analysts can apply standard methods (i.e. $\text{INB} \pm Z_{\alpha/2} \sqrt{SD_{\text{INB}}}$) to compute a (100-\(\alpha\)) confidence interval to examine whether it excludes zero.

Regardless of whether one uses ICERs or INBs, cost-effectiveness acceptability curves (CEACs) are widely used to represent uncertainty associated with results from cost-effectiveness analyses (Figure 3). The y-axis represents the probability that the intervention is cost-effective and the x-axis represents increasing values of $\lambda$.

Figure 3  Cost-effectiveness acceptability curve.

Note: The y-axis represents the probability that the treatment of interest is cost-effective relative to standard care. The x-axis represents a range of willingness-to-pay thresholds, $\lambda$. The cost-effectiveness acceptability curve corresponds to the proportion of bootstrap replicates across different values of $\lambda$. There is direct correspondence with the scatterplot in Figure 2. At $\lambda = 0$ per QALY, the CEAC corresponds to the 16.9% of bootstrap samples that are in the lower 2 quadrants of the cost-effectiveness plane in Figure 2. At $\lambda = 50 000$ per QALY, the CEAC corresponds to the 73.2% of the bootstrap samples that fall below the willingness-to-pay threshold of $50 000 per QALY.

At $\lambda = 0$, the point on the y-axis represents the probability that the intervention is cost-saving relative to the standard care. As willingness-to-pay increases, the probability that an intervention is cost-effective will generally increase.
Other issues to consider

Although some of the more sophisticated analytical techniques may be daunting to many clinicians and clinical researchers, issues pertaining to study design, study execution, and cost estimation can have a greater impact on the overall usefulness of an economic evaluation to guide decision-making. As with clinical outcomes, elements of a strong study design include randomization and accurate and complete data collection.

With economic outcomes, methods used for cost assignment should be critically considered. The extent to which cost outliers are present may not so much represent the proportions of patients who experience complicated clinical courses, but the manner in which costs are assigned to information on resource use available to the analyst. For example, if an analyst applies a fixed unit cost [e.g. diagnosis-related group (DRG) payment] to each hospitalization for heart failure, the variance in costs for inpatient care is suppressed and the method fails to account for outliers. As a result, mean costs will be underestimated and cost comparisons will have inflated statistical power. If an analyst applies a costing approach that relies on resource use, costs, and quality of life in patients with acute compensated heart failure: findings from ASCEND-HF.

Another critical issue is the extent to which cost estimates can be transferred from one location to another. As recently reported by Quentin et al. in the European Heart Journal, cost assignment based on DRGs can vary markedly across countries. For a complicated case of acute myocardial infarction, costs assigned using DRG-based payments in Estonia would have been 11 times higher in England. All else equal, a cost-effectiveness analysis of a strategy designed to limit complications of acute myocardial infarctions would appear more cost-effective in Estonia than in England.

Conclusions

Statistical methods used in economic evaluations have become more sophisticated. Regardless of the complexity of the statistical methods employed, careful interpretation of a study will require scrutiny of the study design, data elements, and cost estimation methods used.

Funding

This work was supported internally by the Duke Clinical Research Institute.

Conflict of interest: none declared.

References