Measuring and illustrating statistical evidence in a cost-effectiveness analysis

Jeffrey S. Hoch\textsuperscript{a,b,\ast}, Jeffrey D. Blume\textsuperscript{c}

\textsuperscript{a} Centre for Research on Inner City Health, The Keenan Research Centre in the Li Ka Shing Knowledge Institute of St. Michael’s Hospital, Toronto, Ontario, Canada

\textsuperscript{b} Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

\textsuperscript{c} Center for Statistical Sciences, Brown University, Providence, RI 02912, USA

Received 11 March 2005; received in revised form 2 July 2007; accepted 18 July 2007

Available online 7 January 2008

Abstract

Recently, there has been much interest in using the cost-effectiveness acceptability curve (CEAC) to measure the statistical evidence of cost-effectiveness. The CEAC has two well established but fundamentally different interpretations: one frequentist and one Bayesian. As an alternative, we suggest characterizing the statistical evidence about cost-effectiveness using the likelihood function (the key element of both approaches). Its interpretation is neither dependent on the sample space nor on the prior distribution. Moreover, the probability of observing misleading evidence is low and controllable, so this approach is justifiable in the traditional sense of frequentist long-run behaviour. We propose a new graphic for displaying the evidence about cost-effectiveness and explore the strengths of likelihood methods using data from an economic evaluation of a Program in Assertive Community Treatment (PACT).

© 2007 Elsevier B.V. All rights reserved.

JEL classification: I18

Keywords: Cost-effectiveness analysis; Net benefit regression framework; Likelihood methods

1. Introduction

There is growing interest in improving the quality of trial-based economic evaluations (see, for example Chapter 6 in Drummond and McGuire (2001); Chapter 8 in Drummond et al. (2005) and Ramsey et al. (2005)). The International Society for Pharmacoeconomics and Outcomes Research’s Task Force on Good Research Practices (RCT-CEA Task Force) recently concluded that “conducting high quality economic analyses alongside clinical studies is desirable because they provide timely information with high internal validity” (Ramsey et al., 2005). This conclusion was framed in the context of an “increasing demand [for] evidence of economic value for health-care interventions” (Ramsey et al., 2005). An accompanying editorial stressed the need to assess evidence in the data, using the word “evidence” more than 10 times in the space of less than two pages (Backhouse, 2005). So it would
seem desirable to provide consumers of trial-based economic evaluations with an assessment of the strength of the evidence provided by the data that a new treatment or intervention is cost-effective. Ideally, this assessment would describe only “what the data say” about the cost-effectiveness of the intervention. This paper describes how Likelihood methodology can meet this need; we explain its advantages and limitations, and we contrast it with more common methods.

Depending on the methodology chosen, statistical analysis of an intervention’s cost-effectiveness can answer one of the following three questions: “Now that we have seen the data,

(1) Should we act as if the intervention is cost-effective?;
(2) Do we believe that the intervention is cost-effective?; or
(3) What is the evidence that the intervention is cost-effective?”

The answer to question (3) is derived entirely from the data, whereas the answers to the first two questions depend on more that just the data themselves (e.g., the impending action’s probable gains and losses, or what was initially believed to be true) (Berkson, 1942, 2003; Blume, 2002; Cox, 1958; Edwards, 1970; Fisher, 1959; Neyman, 1950; Royall, 1997). As a result, the three questions and their distinct answers must be considered separately. The tools and principles of frequentist Decision theory (i.e., Neyman–Pearson–Wald hypothesis testing) and Bayesian methods can be used to answer the first two questions; however, many have argued that problems arise when these frequentist or Bayesian methods are applied to question (3) (Blume and Peipert, 2003; Goodman, 1993; Goodman and Royall, 1988; Royall, 1997).

Depending on its interpretation, the cost-effectiveness acceptability curve (CEAC) (Briggs and Fenn, 1998; Fenwick et al., 2001, 2004; Lothgren and Zethraeus, 2000; van Hout et al., 1994) is often used to address either question (1) or question (2), but it is insufficient for answering question (3). A separate theory or paradigm dedicated to measuring statistical evidence is needed. Many have argued that such a paradigm must focus on the likelihood function alone (Barnard, 1949; Birnbaum, 1962; Edwards, 1970; Fisher, 1922; Royall, 1997). While the likelihood function figures prominently in both frequentist and Bayesian methodologies, it is neither the focus nor the endpoint of either methodology. In contrast, the likelihood paradigm, with the likelihood function as its centrepiece, embodies a theory of statistical evidence (Blume, 2002; Edwards, 1972, 1992; Royall, 2000a,b). Under the likelihood paradigm, the statistical evidence is characterized by the likelihood function and the strength of the evidence for one hypothesis over another is measured by their likelihood ratio. Thus, question (3) can be answered by using the likelihood function to characterize the statistical evidence about cost-effectiveness.

Empirical examples of the likelihood paradigm can be found in the scientific literature (Blume, 2002; Mellen and Royall, 1997).

This paper considers how characterizing the strength of evidence in the data through the use of likelihood methods could complement current techniques for analyzing trial-based cost-effectiveness data. As a framework for our likelihood analysis, this paper begins by briefly tracing some of the major advances in statistical cost-effectiveness analysis. Next, we examine the different interpretations of the CEAC and discuss how they address questions (1) and (2). Subsequently, we introduce the likelihood paradigm, describe why it represents a compromise between the frequentist and Bayesian approaches, discuss its frequentist behaviour (e.g., how often strong misleading evidence is observed), and show how to measure evidence of cost-effectiveness in the net benefit regression framework (Hoch et al., 2002). In addition, we propose a new graphic that displays the statistical evidence about cost-effectiveness and apply it to data from an economic evaluation of a program of assertive community treatment (ACT) for homeless people with

---

1 Significance testing, as Fisher proposed it, was intended to address question (3). However, as both Goodman (1993) and Royall (1997) point out, today’s version of significance testing is actually a combination of hypothesis testing concepts and Fisher’s original significance testing ideas, and the resulting ‘methodology’ and application is much closer to hypothesis testing. So, for the purposes of this paper, we classify such methods along with those of hypothesis testing (although we are careful to make the necessary distinctions through the paper). Note also that neither incarnation of significance testing is adequate for answering the third question (Royall, 1997), and that we do not in this context delineate between Bayesian decision theory which addresses question (1) and Bayesian belief theory which addresses question (2).

2 Notable examples are the controversies surrounding adjustment for multiple comparisons and multiple looks at data, as well as the proper use and interpretation of p-values (Royall, 1986; Goodman and Royall, 1988; Goodman, 1998; Blume and Peipert, 2003).

3 We do not exclude the possibility that competing theories of statistical evidence will be (or have been!) proposed and developed. Rather, our contention is that neither hypothesis testing nor Bayesian methods represent competing theories of statistical evidence, largely because their derivations and intended uses were for different purposes.
Considering an economic analysis where study participants receive "usual care" (Dewa, 2007). Thus, denote the population expected values of cost and effect for treatment for various patient subgroups, Eq. (1) can be enhanced with a vector of subject characteristics ($X_i$) and interaction terms between subject characteristics and the treatment indicator ($X_i \cdot TX$); for more details see (Briggs et al., 2002; Hoch et al., 2002).

The Ordinary Least Squares (OLS) estimate of $\beta_{TX}$ in Eq. (1) is simply

$$\hat{\beta}_{TX} = \hat{NB}_1(\lambda) - \hat{NB}_0(\lambda) = (\hat{E}_1 \cdot \lambda - \hat{C}_1) - (\hat{E}_0 \cdot \lambda - \hat{C}_0) = \Delta E \cdot \lambda - \Delta C$$

where $\hat{NB}(_\lambda)$ is the mean net benefit for treatment $t$ (for a sketch of the proof, see the technical appendix of Hoch and Dewa, 2007). Thus, $\hat{\beta}_{TX}$, the estimate of $\beta_{TX}$, is the incremental net benefit (INB). When the assumptions for OLS hold, one can use the probability that $\hat{\beta}_{TX} > 0$ for the $\gamma$-axis and $\lambda$ for the $\lambda$-axis to generate the CEAC (Hoch et al., 2006; Lothgren and Zethraeus, 2000).

Using the commonly reported two-sided $p$-values from computer regression programs, one can compute this as $p/2$ when $\hat{\beta}_{TX} < 0$ and $1 - p/2$ when $\hat{\beta}_{TX} > 0$. Alternatives to the $p$-value approach include using the percentage of bootstrap replicates where $\hat{\beta}_{TX} > 0$.

2. Methodological background

2.1. Stochastic cost-effectiveness analysis

2.1.1. The incremental cost-effectiveness ratio (ICER)

The goal of a cost-effectiveness analysis (CEA) is to provide an estimate of the trade-off between additional costs and additional outcomes. A commonly used statistic to estimate this trade-off is the incremental cost-effectiveness ratio (ICER). The ICER estimate is a ratio of the extra cost of treatment ($\Delta C$) to the extra effect of treatment ($\Delta E$). Considering an economic analysis where study participants receive "usual care" ($t=0$) or "new treatment" ($t=1$), we denote the population expected values of cost and effect for $t=0$ and $t=1$ as $\mu^C_0$, $\mu^E_0$, $\mu^C_1$ and $\mu^E_1$, respectively. The population ICER statistic is $(\mu^C_1 - \mu^C_0)/(\mu^E_1 - \mu^E_0)$.

From clinical trial data, one can estimate $\Delta C$ and $\Delta E$ in the following way: $\Delta C = \bar{C}_1 - \bar{C}_0$ and $\Delta E = \bar{E}_1 - \bar{E}_0$ where $\bar{C}_i$ and $\bar{E}_i$ are the sample cost and effect averages for subjects receiving treatment option $t$. To express the extra cost associated with the extra gain at a per unit level, analysts calculate an estimate of the ICER, $\hat{R} = \Delta C/\Delta E = (\bar{C}_1 - \bar{C}_0)/(\bar{E}_1 - \bar{E}_0)$. If $\hat{R} < \lambda$, where $\lambda$ represents the willingness to pay for an additional health outcome or effect, then the intervention is said to be cost-effective.

2.1.2. The net benefit approach and regression framework

The net benefit approach is a simple and insightful alternative to the ICER (Stinnett and Mullahy, 1998; Tambour et al., 1998). By assessing cost-effectiveness using the incremental net benefit (i.e., $\Delta E \cdot \lambda - \Delta C$) both estimation and inference are greatly simplified. A key caveat is that one cannot specify $\lambda$, which plays an implicit role in the net benefit approach and an explicit role in the ICER approach.

Recent methodological work has developed a regression framework for the net benefit approach (Hoch et al., 2002). Under this framework, each subject's net benefit is computed from the observed data as

$$NB(\lambda)_i = e_i \cdot \lambda - c_i$$

where $e_i$ and $c_i$ are the data for the $i$th person's effect and cost, respectively, and $\lambda$ is a willingness to pay parameter that must be specified. The notation $NB(\lambda)$ is used to emphasize the fact that the dependent variable (i.e., the net benefit) is a function of $\lambda$ (Lothgren and Zethraeus, 2000). In its simplest form, the net benefit regression framework (NBRF) involves fitting the following regression model

$$NB(\lambda) = \beta_0 + \beta_{TX}X_i + \epsilon_i$$

(1)

where $TX_i$ and $\epsilon_i$ are $i$th person's treatment indicator ($TX_i = 1$ for new treatment and 0 for usual care) and stochastic error term, respectively. Eq. (1) is typically fit several times, each time with a different value of $\lambda$ (in principle, decision makers only consider the results for their own value of $\lambda$). To measure the "marginal" cost-effectiveness of the new treatment for various patient subgroups, Eq. (1) can be enhanced with a vector of subject characteristics ($X_i$) and interaction terms between subject characteristics and the treatment indicator ($X_i \cdot TX$); for more details see (Briggs et al., 2002; Hoch et al., 2002).

The Ordinary Least Squares (OLS) estimate of $\beta_{TX}$ in Eq. (1) is simply

$$\hat{\beta}_{TX} = \hat{NB}_1(\lambda) - \hat{NB}_0(\lambda) = (\hat{E}_1 \cdot \lambda - \hat{C}_1) - (\hat{E}_0 \cdot \lambda - \hat{C}_0) = \Delta E \cdot \lambda - \Delta C$$

(2)

where $\hat{NB}(_{\lambda})$ is the mean net benefit for treatment $t$ (for a sketch of the proof, see the technical appendix of Hoch and Dewa, 2007). Thus, $\hat{\beta}_{TX}$, the estimate of $\beta_{TX}$, is the incremental net benefit (INB). When the assumptions for OLS hold, one can use the probability that $\hat{\beta}_{TX} > 0$ for the $\gamma$-axis and $\lambda$ for the $\lambda$-axis to generate the CEAC (Hoch et al., 2006; Lothgren and Zethraeus, 2000).
2.1.3. The cost-effectiveness acceptability curve (CEAC)

The cost-effectiveness acceptability curve (CEAC) has been advocated for summarizing the results of a CEA because it highlights the relationship between cost-effectiveness and the unknown constant \( \lambda \) (Briggs and Fenn, 1998; Fenwick et al., 2001, 2004; Lothgren and Zethraeus, 2000; van Hout et al., 1994). The CEAC is equal to one minus the one-sided significance level for testing the null hypothesis of non-positive incremental net benefits (i.e., the new intervention is not cost-effective; additional benefits are outweighed by additional costs) (Lothgren and Zethraeus, 2000). Under this frequentist framework, the CEAC can be viewed as illustrating a decision rule for rejecting the null hypothesis that the intervention is not cost-effective. As a direct function of the \( p \)-value, the CEAC is subject to the same flaws and problems of interpretation known to be associated with \( p \)-values (Blume and Peipert, 2003; Goodman, 1998; Goodman and Royall, 1988; Royall, 1986, 1997).

Alternatively, the \( p \)-value based CEAC can be interpreted in a ‘Bayesian’ fashion (Briggs and Fenn, 1998; van Hout et al., 1994) as the probability that an individual, who initially believed that every real value for the intervention’s incremental net benefit was equally likely \( \text{a priori} \) (i.e., the ‘non-informative’ prior assumption), now believes the intervention to be cost-effective (i.e., \( \text{INB} > 0 \)). While a Bayesian approach provides a general well-justified interpretation for a CEAC, it presents other challenges. For example, there exist many ‘Bayesian’ CEACs – namely one for every unique prior – with no criteria for choosing between them. This is an important point because every CEAC is ‘correct’ for its given prior. Thus, the calculation of a Bayesian CEAC requires the specification of the prior distribution of incremental net benefit (i.e., the distribution of \text{INB} before the data were collected). Using a ‘non-informative’ prior because it is convenient or because it has the appearance of impartiality creates potential areas of contention in the analysis. This may not be desirable given the huge financial implications of the conclusions from economic evaluations of healthcare products. Some may see an advantage in being able to distinguish the data’s evidence from the analyst’s prior; this cannot be accomplished with a CEAC, as it combines both the data and the prior. Other limitations associated with CEACs have been debated (Fenwick and Briggs, 2007; Groot Koerkamp et al., 2007).

2.1.4. A role for the likelihood paradigm

Given the varied interpretations, it should be clear that the CEAC does not characterize only the evidence of cost-effectiveness. Each interpretation assumes some combination of data and external inputs (e.g., prior beliefs, expected gains or losses). What is missing is a well-defined characterization of what the data themselves say about the evidence of an intervention’s cost-effectiveness. Under the net benefit regression approach this translates into characterizing the statistical evidence about the parameter \( \beta_{\text{TX}} \), the incremental net benefit of the intervention, without regard to prior beliefs or probable gains or losses. Many see the results of economic evaluation as descriptive instead of prescriptive (Drummond et al., 2005), and this view is congruent with producing analyses...
that characterize evidence about a novel technology or intervention’s value for money in relation to an appropriate alternative. Given that the objective measurement of statistical evidence is the raison d’être of the likelihood paradigm, it seems natural to characterize the statistical evidence in a cost-effectiveness analysis using the likelihood paradigm.

2.2. The likelihood paradigm

2.2.1. Statistical evidence and the Law of Likelihood

The Law of Likelihood explains how to interpret observations, in the context of a probability model, as statistical evidence for one hypothesis vis-à-vis another. More formally, the Law of Likelihood (Hacking, 1965; Royall, 1997) states that

If the first hypothesis ($H_1$) implies that the probability that a random variable $X$ takes the value $x$ is $P_1(x)$, while the second hypothesis ($H_2$) implies that the probability is $P_2(x)$, then the observation that $X=x$ is evidence supporting $H_1$ over $H_2$ if and only if $P_1(x) > P_2(x)$, and the likelihood ratio, $P_1(x)/P_2(x)$, measures the strength of that evidence.

Here $P_1(x) = P(x|H_1)$ is the probability of observing $x$ given that $H_1$ is true, and $P_2(x) = P(x|H_2)$ is the probability of observing $x$ given that $H_2$ is true. The likelihood ratio is the ratio of these conditional probabilities; i.e., $P(x|H_1)/P(x|H_2)$. It measures the strength of the evidence for $H_1$ versus $H_2$. The Law simply formalizes the intuitive notion that the hypothesis that assigns a higher probability to the observed event ($X=x$) is better supported by the data. If the likelihood ratio is greater than one, the evidence favours $H_1$ over $H_2$ and if the likelihood ratio is less than one, the evidence favours $H_2$ over $H_1$. If both hypotheses place the same probability on the observed events then the likelihood ratio is unity, and the data do not support one hypothesis over the other.

Likelihood ratios may take any non-negative value, from zero (indicating overwhelming evidence for $H_2$ over $H_1$) to infinity (indicating the reverse). It has been suggested that likelihood ratios between 1 and 8 suggest weak evidence for $H_1$ over $H_2$, while likelihood ratios between 8 and 32 suggest moderate evidence, and likelihood ratios greater than 32 suggest strong evidence for $H_1$ over $H_2$ (Edwards, 1992; Jeffreys, 1961; Kass and Raftery, 1995). Note that these suggested interpretations are not strict cutoff values; a likelihood ratio of 32 represents fairly strong evidence, but so does a likelihood ratio of 31 or 33 (albeit to a lesser or greater degree).

In situations with many hypotheses of interest, multiple likelihood ratios should, in principle, be reported (namely one for each pair of hypotheses), but simply listing all of them becomes quite cumbersome. A more concise way of reporting the evidence is to show all the likelihood ratios by plotting the likelihood function.

2.2.2. The likelihood function

The likelihood function is the mathematical representation of the statistical evidence in the data (Berger, 1984; Birnbaum, 1962). In essence, a likelihood function expresses the conditional probabilities (i.e., $P(x|H_1)/P(x|H_2)$). It measures the strength of the evidence for $H_1$ versus $H_2$. The hypothesis that is best supported by the data reveals which hypotheses are better supported by the data because these hypotheses will have a larger $P(x|H_h)$ relative to other hypotheses. The hypothesis that is best supported by the data will have the highest likelihood.

Hence, looking at a graph of the likelihood function literally shows what the data say. To illustrate, consider an economic evaluation that generates evidence about the incremental net benefit (INB) of a new treatment compared to usual care. In the regression framework outlined previously, $\beta_{TX}$ is the parameter of interest because it represents the INB of the new intervention. If we assume that the errors are normally distributed with constant variance (we address the robustness of this assumption later) then standard regression theory indicates that the statistic $(\hat{\beta}_{TX} - \beta_{TX})/se(\hat{\beta}_{TX})$ has a $t$-distribution with $v = n - 2$ degrees of freedom yielding a likelihood function for $\beta_{TX}$ that is proportional to the kernel of the $t$-distribution. That is,

$$L(\beta) \propto \left[1 + \frac{t^2}{v}\right]^{-\frac{v+1}{2}} \text{ where } t = \frac{\hat{\beta} - \beta}{se(\hat{\beta})}$$
where the treatment subscript (TX) is removed for clarity. Note that $\hat{\beta}$ and $\text{se}(\hat{\beta})$ are the parameter and standard error estimates obtained directly from the data under the NBRF. Thus, for a given $\lambda$, the evidence supporting $H_1: \beta_{TX} = 0$ over $H_2: \beta_{TX} = 1000$ is measured by the likelihood ratio $L(H_1)/L(H_2) = L(0)/L(1000)$, and likewise for $H_3: \beta_{TX} = -20.75$ over $H_1$ with $L(-20.75)/L(0)$. The hypothesis that is best supported by the data is the value of $\beta_{TX}$ that maximizes the likelihood function (i.e., the maximum likelihood estimator or MLE). When the regression errors are normally distributed, $\hat{\beta}$, the OLS estimate of $\beta_{TX}$, is also the MLE and therefore the best supported hypothesis.

Likelihood functions are graphed to provide a visual impression of the evidence over the parameter space. For presentation purposes, likelihood functions are scaled by their maximum value (a constant). A standardized likelihood function is defined as

$$
\frac{L(\beta)}{\max L(\beta)} = \frac{L(\beta)}{\hat{L}(\hat{\beta})} = \left[1 + \frac{t^2}{v}\right]^{-\frac{v+1}{2}} \quad \text{where} \quad t = \frac{\hat{\beta} - \beta}{\text{se}(\hat{\beta})} \quad \text{for all real } \beta
$$

Fig. 1 displays a standardized likelihood function when $\hat{\beta} = 55,000$, $\text{se}(\hat{\beta}) = 20,600$, and $v = 143$ degrees of freedom. The x-axis represents the parameter space (i.e., all possible hypotheses for $\beta_{TX}$) while the y-axis gives the re-scaled (standardized) likelihood value. The best supported hypothesis, $\beta_{TX} = 55,000$, is at the crest of the likelihood function. Nevertheless, other hypotheses are (essentially) equally well supported by the data. That is, there is only very weak evidence to support $\beta_{TX} = 55,000$ over $\beta_{TX} = 60,000$ because $L(60,000)/L(55,000) = 0.97$.

2.2.3. Likelihood support intervals

Instead of reporting numerous likelihood ratios or graphing the likelihood function, a collection or set of hypotheses that are consistent with the data can be used to summarize the evidence about an intervention’s cost-effectiveness. That set is called a $1/k$ likelihood support interval (SI), where $k$ indicates the strength of support. Precisely, a $1/k$ likelihood support interval is defined as the set of hypotheses that are supported over the best hypothesis by a factor of $1/k$ or more (i.e., when the standardized likelihood function is greater than $1/k$). In set notation, the definition is \{all $\beta_{TX}: L(\beta_{TX})/\max L(\beta_{TX}) \geq 1/k$\} = \{all $\beta_{TX}: L(\hat{\beta}_{TX})/L(\hat{\beta}_{TX}) \leq k$\}. If $k = 8$ (32), at most there is only weak (moderate) evidence to support the best supported hypothesis over the other hypotheses in the support interval. In passing, we note that support intervals and confidence intervals are mathematically

---

11 The scaling constant for the likelihood function can be chosen arbitrarily because only ratios of likelihood functions measure the statistical evidence. When standardizing to the maximum value, the resulting ratio measures the relative support for each hypothesis compared to the best supported hypothesis.

12 $L(\beta_{TX})/L(\hat{\beta}_{TX}) = [1 + 1.000412]^{-72} = 0.97$. For the best supported hypothesis $\hat{\beta}_{TX}$, $L(\beta_{TX})/L(\hat{\beta}_{TX}) \leq 1$ for all $\beta_{TX}$.

13 Note that hypotheses within a $1/k$ SI may be better supported over others within the same interval, but the additional support is always weak (less than a factor of 8 for a 1/8 SI). For those hypothesized values outside of the 1/8 SI, there always exists another hypothesized value for $\beta_{TX}$. 

---

Fig. 1. The standardized likelihood function.
related, but the interpretation of these intervals is quite different. For an in-depth discussion see Blume (2002) and Royall (1997). Another difference is that likelihood support intervals depend only on the data and keep the same form regardless of the number of looks at the data and regardless of the number of multiple comparisons performed.

More often than not, likelihood support intervals must be determined by numerical methods, but under the \( t \)-distribution an analytical form does exist. Specifically, the endpoints of a \( 1/k \) SI for \( \beta_{TX} \) are given by\(^{14}\)

\[
\hat{\beta} \pm \text{se}(\hat{\beta}) \sqrt{v \exp \left( \frac{2 \ln k}{v+1} \right) - v}
\]

Continuing with the previous example, the \( 1/8 \) SI for \( \beta_{TX} \) is \$12,831.75 to \$97,168.25 and the \( 1/32 \) SI is \$296.65 to \$109,703.30 indicating that the data provide strong evidence that the benefits outweigh the costs of the new treatment by at least \$296.65. These support intervals are displayed in Fig. 1 as the horizontal lines drawn parallel to the \( x \)-axis within the likelihood function (their ‘height’ is equal to the level of support, \( 1/8 \) or \( 1/32 \) as the case may be). The \( 1/32 \) SI is considered a “stronger” support interval compared to the \( 1/8 \) SI because a wider range of hypotheses are said to be consistent with the data. In general, if the SI contains negative values then there is only weak evidence in the data to support claims of cost-effectiveness. Note that as \( \lambda \) varies, so do the support intervals. Hence it is useful to plot the support intervals as a function of \( \lambda \) in order to assess evidence of cost-effectiveness as \( \lambda \) is varied. We provide an example of this graphic (which we call the “evidence of cost-effectiveness” plot) in the context of the empirical example discussed in Section 3.4.

2.2.4. Right likelihood; wrong answer?

It is possible to observe strong evidence for \( H_2 \) over \( H_1 \) when, in actuality, \( H_1 \) is correct (i.e., \( L(H_2)/L(H_1) \geq k \) when \( H_1 \) is correct). This is an example of misleading evidence defined as strong (\( k \)-strength) evidence in favour of the incorrect hypothesis over the correct hypothesis. When the data are collected, the strength of the evidence will be determined by the likelihood ratio. Whether the evidence is weak or strong will be clear from the numerical value of the likelihood ratio. However, it remains unknown if the evidence is misleading or not. Fortunately, the probability of observing misleading evidence is, in general, low (Blume, 1999, 2002; Pratt, 1977; Royall, 1997, 2000a,b). The probability of observing misleading evidence of \( k \)-strength or greater is always less than or equal to \( 1/k \) (Birnbaum, 1962; Royall, 1997).\(^{15}\) For example, the probability of observing strong misleading evidence (i.e., evidence supporting an incorrect hypothesis over the correct hypothesis by a factor of 32 or more) cannot exceed \( 1/32 = 0.031 \). Any viable theory of statistical evidence must provide the ‘correct’ answer more often than not and Likelihood ratios satisfy this criterion.\(^{16}\)

The measure of the strength of evidence (i.e., the likelihood ratio) is a distinct mathematical quantity from the frequency with which misleading evidence is observed. If this distinction is overlooked problems arise. For example, frequentist significance testing uses the \( p \)-value to represent both the strength of the evidence and the probability of observing misleading evidence (Blume and Peipert, 2003; Goodman, 1993; Goodman and Royall, 1988; Royall, 1997). Controversies ensuing over adjustments for multiple looks at the data or multiple comparisons are irresolvable because these activities increase the probability of observing misleading evidence but do not affect the measure of the strength of observed evidence. It is impossible for the \( p \)-value to reflect both activities because the \( p \)-value would need to be adjusted (in its role as the probability of observing misleading evidence) but, at the same time, remain unadjusted (as the measure of the strength of statistical evidence).

\(^{14}\) Direct calculation gives the result; see also Royall (1997) and Blume (2002).
\(^{15}\) Mathematically, if \( H_f: X \sim f(X) \) while \( H_g: X \sim g(X) \) then \( P_g(f(X)/g(X) \geq k) \leq 1/k \). This “universal bound” indicates that for moderately large \( k \), misleading evidence will not be observed very often.
\(^{16}\) In fact, the probability of observing misleading evidence rarely, if ever, achieves the universal bound in a fixed sample size study; however, the probability of observing misleading evidence does increase with each planned look at the data, but also remains bounded by the universal bound (Robbins, 1970; Blume, 1999, 2002). A comprehensive discussion of the probability of observing misleading evidence is given by Royall (2000a,b) and Blume (2002).
2.2.5. Wrong likelihood; right answer?

Robust likelihood functions can be made insensitive to model misspecification (Royall and Tsou, 2003) in a regression setting (Blume et al., 2007). This is accomplished by adjusting the likelihood function, so that the adjusted likelihood behaves as if the model were correctly specified. The practical implication is that the probability of observing misleading evidence under the adjusted likelihood function is low in large samples. When the object of inference is a mean, the t-distribution is naturally robust to model misspecification and does not require any adjustment (Royall and Tsou, 2003). More generally, however, are cases in which we are interested in a single parameter from a regression model (e.g., $\beta_{TX}$ in the NBRF). In Appendix A, we derive the robust adjusted profile likelihood for a regression parameter from a normal linear regression model.

Robust likelihood functions play a valuable part when using the likelihood paradigm in the NBRF. For small values of $\lambda$, the regression errors in Eq. (1) will most likely not be normally distributed due to the high skewness of cost data. In order to learn reliably about $\beta_{TX}$, which is itself a mean, we could use its marginal t-distribution or, alternatively, use its robust adjusted likelihood function under the working regression model (given in Appendix A). Hence, even though the error distribution from our regression model may not be correctly specified, the probability of observing misleading evidence about the INB ($\beta_{TX}$) will, with a large enough sample, be no more than if we had initially specified the correct error distribution. Sometimes, the price we pay for this robustness is a loss of efficiency. More details are available elsewhere (Royall and Tsou, 2003; Tsou and Royall, 1995).

2.2.6. Nuisance parameters

Nuisance parameters are parameters in the likelihood function that are not of interest at the moment (e.g., the variance would be a nuisance parameter when the mean of a normal model is the object of interest) (Blume, 2002; Royall, 1997). They represent a difficulty for the application of likelihood methods because they are typically unknown and the likelihood function cannot be computed without them. To this problem there is no general solution. In some situations, a marginal or conditional likelihood that is free of the nuisance parameter will be available. In others, when no ‘true’ likelihood is available, a profile likelihood or estimated likelihood could be used. Here the different options are comparable by how often they would yield misleading evidence (this is another reason why the properties discussed in Sections 2.2.4 and 2.2.5 are vital). Because marginal or conditional likelihoods are legitimate likelihood functions, they are no more likely to be misleading than the true likelihood function. Profile likelihoods share this property in large samples, while estimated likelihoods are, in general, more likely to be misleading (Blume, 2002; Royall, 1997, 2000a,b; Royall and Tsou, 2003). Hence there is a clear and objective criterion for choosing between approaches for dealing with nuisance parameters; one should choose the approach that is least likely to be misleading.18

2.3. Bayes factors

A Bayes factor is a likelihood ratio derived from the posterior probability model (i.e., the probability model that results from combining the likelihood function and the prior distribution) (Blume, 2002; Royall, 2000b). As such, Bayes factors can be interpreted, in accordance with the Law of Likelihood, as measuring the strength of the evidence under the posterior model (Goodman, 1999; Kass and Raftery, 1995). When the two hypotheses under consideration are simple hypotheses (e.g., $H_1$: $\beta = 0$ vs. $H_2$: $\beta = 10$), the Bayes factor reduces to the likelihood ratio that is derived directly from the likelihood function. In this sense, the measurement of evidence is robust; it is valid regardless of the prior distribution. However, when one or both of the two hypotheses are composite (e.g., $H_1$: $\beta > 0$ vs. $H_2$: $\beta \leq 0$), only the Bayes factor can be computed. This is because the likelihood function $L(\beta) = f(x|\beta)$ depends on

---

17 This is a special case. Typically likelihoods require some form of adjustment. Also, it should be noted that, technically, Royall and Tsou (2003) show that the normal likelihood function with the variance parameter profiled out is robust for measuring statistical evidence about a mean. But the normal profile likelihood for a mean is proportion to a t-distribution (Blume, 2002), so the robustness carries through.

18 Some might argue that nuisance parameters threaten the objectivity of the likelihood approach because the choice of a substitute likelihood could be considered subjective. We think this argument is faulty given the choice of a substitute likelihood is made according to a clear, objective criterion: choose the substitute likelihood that is least likely to be misleading. We note that no such analogous criterion exists for choosing a prior distribution or a sample space.
a single value of $\beta$ (not a range of $\beta$’s) and cannot be computed without additional constraints on the model. The Bayes factor solves this problem by taking a weighted average of the likelihood function over the range in question, where the weights are given by the prior distribution, $g(\beta)$. Now, however, the measurement of evidence is dependent on the prior distribution. Hence the cost of summarizing the strength of evidence over a composite hypothesis is that one must specify a prior distribution. For more on this topic, see Blume (2002) and Royall (1997).

3. Empirical example

3.1. Background

The Program in Assertive Community Treatment (PACT) is one of the most studied models of care for persons with severe and persistent mental illnesses (SPMI) (Burns and Santos, 1995; Olsson, 1990; Scott and Dixon, 1995; Stein and Test, 1980; Test and Stein, 1980; Weisbrod et al., 1980). A study in Baltimore, Maryland (USA), found that an assertive community treatment (ACT) program, relative to usual community services, reduced psychiatric inpatient days, emergency room visits, days homeless, and days in jail for homeless persons with SPMI (Lehman et al., 1997). The study’s rationale was that by providing potentially more expensive but coordinated, community-based care through the ACT program, homeless persons with severe mental illnesses would spend more days in stable community housing with savings realized by shifting the patterns of care from higher cost crisis-oriented inpatient and emergency services to lower cost, ongoing ambulatory services. The results suggested that in the city of Baltimore, ACT was effective in achieving significant outcomes warranting an economic evaluation. The initial cost-effectiveness analysis of the ACT program used ICERS (Lehman et al., 1999); a subsequent cost-effectiveness analysis introduced net benefit regression using these data (Hoch et al., 2002).

3.2. Study design and data

Direct treatment costs across the 1-year intervention period were examined from the perspective of the state mental health authority. Housing status was chosen as the main effectiveness measure because of its established validity as a primary outcome for homeless persons with SPMI (Newman, 1992). A day of stable housing was defined as living in a non-institutionalized setting not intended to serve the homeless (e.g., independent housing, living with family, etc.). Subjects randomized to the comparison usual care condition had access to services usually available to homeless persons in the city of Baltimore. Additional details about the study’s methodology are available (Lehman et al., 1999).

One hundred and forty-eight persons who were homeless with SPMI were randomized to either the experimental ACT program or to usual community services. Subjects were recruited during a 19-month period in 1991 and 1992 from inner-city psychiatric hospitals, primary health care agencies, shelters, missions and soup kitchens. Baseline data collection included age and race variables as well as an assessment of overall mental health functioning using the Global Assessment of Functioning (GAF) Scale (American Psychiatric Association, 1987). For this paper, we obtained complete data on 73 participants randomly assigned to the ACT program and 72 randomly assigned to usual care (comparison) services.

3.3. Standard results

Baseline group comparisons examined differences between the two intervention groups on demographics, diagnoses and histories of homelessness at baseline (Lehman et al., 1999). The two groups were well balanced except in regard to race; there was a greater than expected percentage of African Americans randomized to the comparison condition (83% vs. 62%). Table 1 provides a brief statistical summary of the cost and effect data and provides a conventional cost-effectiveness analysis of the data by looking at the incremental costs and effects between the two treatment groups. On average, ACT subjects had more days of stable housing and lower costs. Due to the imbalance in the race distribution between treatment arms, Table 1 also reports the results of a stratified analysis.

\[ \int_{-\infty}^{0} f(x|\beta)g(\beta) \, d\beta / \int_{-\infty}^{0} f(x|\beta)g(\beta) \, d\beta \].
Table 1
Sample statistics and net benefit regression results from the economic evaluation data

<table>
<thead>
<tr>
<th>Group variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall analysis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison arm (N=72)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>67,400</td>
<td>76,500</td>
<td>9,020</td>
</tr>
<tr>
<td>Effect</td>
<td>159</td>
<td>105</td>
<td>12.4</td>
</tr>
<tr>
<td>Correlation = −0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT arm (N=73)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>51,900</td>
<td>61,100</td>
<td>7160</td>
</tr>
<tr>
<td>Effect</td>
<td>212</td>
<td>104</td>
<td>12.2</td>
</tr>
<tr>
<td>Correlation = −0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost difference</td>
<td>−15,500</td>
<td>–</td>
<td>11,500</td>
</tr>
<tr>
<td>Effect difference</td>
<td>52.7</td>
<td>–</td>
<td>17.4</td>
</tr>
<tr>
<td>Correlation = −0.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stratified analysis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost difference</td>
<td>−5,070</td>
<td>–</td>
<td>13,200</td>
</tr>
<tr>
<td>Effect difference</td>
<td>35.6</td>
<td>–</td>
<td>21.8</td>
</tr>
<tr>
<td>White subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost difference</td>
<td>−62,700</td>
<td>–</td>
<td>33,000</td>
</tr>
<tr>
<td>Effect difference</td>
<td>98.1</td>
<td>–</td>
<td>39.0</td>
</tr>
</tbody>
</table>

Incremental net benefits

<table>
<thead>
<tr>
<th>Value of ceiling ratio</th>
<th>Net benefit (S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>( \lambda = 0^d )</td>
<td>15,500 (11,500)</td>
</tr>
<tr>
<td>( \lambda = 100 )</td>
<td>20,800 (12,300)</td>
</tr>
<tr>
<td>( \lambda = 500 )</td>
<td>41,900 (17,000)</td>
</tr>
<tr>
<td>( \lambda = 1000 )</td>
<td>68,200 (24,500)</td>
</tr>
</tbody>
</table>

\[ \text{NB}_i(\lambda) = \gamma_0 + \gamma_{TX}TX_i + \gamma_{\text{Black}_{\text{Black}}}\text{Black}_i + \gamma_{TX\text{Black}}TX_i\cdot\text{Black}_i + \varepsilon_i \]

where \( \text{Black}_i \) is an indicator variable for race are presented at the bottom of Table 1. Net benefits were calculated for \( \lambda = 0, $100, $500, \) and $1000. Complete regression results are reported in Table 2.

The results suggest an interaction between race and treatment with black subjects achieving lower incremental net benefits from ACT in comparison to their white counterparts. Fig. 2 displays the CEACs separately for white and black participants as well as all participants combined. The CEAC for white participants never dips below 95%, and the CEAC for black participants never reaches 95%.\(^{20}\) A potential frequentist interpretation of the CEACs (Fig. 2) is that, for a given \( \lambda \), if the CEAC is above 95% then we should act as if the intervention is cost-effective. In contrast, the CEACs (Fig. 2) can be interpreted as representing a Bayesian’s belief (i.e., the probability) that the intervention is cost-effective, based on the evidence and the \textit{a priori} belief that every possible hypothesis for the INB was equally likely.

\(^{20}\) Note, however, that these are not necessarily the same CEACs that other analysts of this data would produce. Activities like performing other statistical comparisons and monitoring study data as it accumulates would require some form of adjustment to the CEAC (in theory).
Table 2
Net benefit regression estimates with treatment interaction (N = 145)\textsuperscript{a}

<table>
<thead>
<tr>
<th>OLS regression equation</th>
<th>NB with ( \lambda = $0 ) [S.E.] \textsuperscript{c} (p-value)</th>
<th>NB with ( \lambda = $100 ) [S.E.] \textsuperscript{c} (p-value)</th>
<th>NB with ( \lambda = $500 ) [S.E.] \textsuperscript{c} (p-value)</th>
<th>NB with ( \lambda = $1000 ) [S.E.] \textsuperscript{c} (p-value)</th>
<th>Effect\textsuperscript{d} [S.E.] \textsuperscript{c} (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant term</td>
<td>(-67,400 \ [8,160] \ (&lt;0.001))</td>
<td>(-112,200 \ [30,700] \ (&lt;0.001))</td>
<td>(-51,500 \ [8,740] \ (&lt;0.001))</td>
<td>(-99,000 \ [32,900] \ (0.003))</td>
<td>(-45,900 \ [43,400] \ (0.292))</td>
</tr>
<tr>
<td>Treatment dummy</td>
<td>(15,500 \ [11,500] \ (0.179))</td>
<td>(62,700 \ [32,100] \ (0.053))</td>
<td>(20,800 \ [12,300] \ (0.093))</td>
<td>(72,600 \ [34,400] \ (0.037))</td>
<td>(41,900 \ [17,000] \ (0.015))</td>
</tr>
<tr>
<td>Covariate</td>
<td>(53,400 \ [31,900] \ (0.094))</td>
<td>(57,000 \ [34,100] \ (0.097))</td>
<td>(69,800 \ [45,100] \ (0.125))</td>
<td>(111,800 \ [45,600] \ (0.015))</td>
<td>(68,200 \ [24,500] \ (0.006))</td>
</tr>
<tr>
<td>Treatment–covariate interaction</td>
<td>(57,700 \ [34,800] \ (0.099))</td>
<td>(63,900 \ [37,200] \ (0.088))</td>
<td>(88,900 \ [49,800] \ (0.076))</td>
<td>(120,200 \ [68,700] \ (0.082))</td>
<td>(120,200 \ [68,700] \ (0.082))</td>
</tr>
<tr>
<td>(R)-squared (adjusted)</td>
<td>0.006</td>
<td>0.018</td>
<td>0.041</td>
<td>0.076</td>
<td>0.080</td>
</tr>
<tr>
<td>Prob &gt; F</td>
<td>0.179</td>
<td>0.064</td>
<td>0.015</td>
<td>0.030</td>
<td>0.006</td>
</tr>
</tbody>
</table>

\textsuperscript{a} All monetary measures in U.S. dollars, all results to three significant figures.

\textsuperscript{b} When \( \lambda = 0 \), NB = \(-\)cost.

\textsuperscript{c} Huber-White robust standard errors and p-values corrected for heteroskedasticity.

\textsuperscript{d} The coefficients from the effect regression are reported since as \( \lambda \) approaches infinite, the p-values for the regression estimates are equivalent to those obtained when ‘days stable housing’ is the dependent variable.
3.4. Results from the likelihood approach

For a given $\lambda$, the likelihood function characterizes the statistical evidence about the treatment coefficient in the net benefit regression (i.e., the INB). This evidence is summarized in the 1/8 and 1/32 likelihood support intervals for $\beta_{TX}$ which were derived from the robust profile likelihood function. Unlike the CEACs, these intervals depend only on the data and summarize the statistical evidence in the data. To obtain these support intervals for any $\lambda$ of interest, one can refit Eq. (1). Plotting these support intervals as a function of $\lambda$, the end result is an “evidence of cost-effectiveness” graph. The figure illustrates the evidence of cost-effectiveness in the data for the INB parameter, $\beta_{TX}$, over a range of $\lambda$. We constructed such plots for the entire sample and the two racial subgroups. Note that the assumption of normal errors in the regression is most likely not met, but our sample size should be sufficiently large to rely on the robustness properties previously described. The price we pay for this robustness is an increase in the width of the support intervals.

Four different “evidence of cost-effectiveness” plots were constructed for this analysis. Figs. 3A and 4A and B display the evidence about the INB parameter, $\beta_{TX}$, for different groupings of participants. Fig. 3A represents the evidence from all participants; Fig. 4A and B shows the evidence for the two racial subgroups. The support intervals from the robust profile likelihood function were nearly identical to those derived from the marginal $t$-distribution when the evidence from the whole sample was considered (Fig. 3A). For the black subgroup, the support intervals from the robust likelihood were, on average, 1.4% wider. However for the white subgroup, the robust intervals were, on average, 30% wider (range 27–34% over $\lambda$). The increase in inefficiency for the white subgroup is partially due to its small sample size (40 out of the 145 participants $\approx 28\%$).

Support intervals including non-positive (or near-zero) values indicate that the intervention may not be cost-effective (or that there is only weak evidence in the data to support claims of cost-effectiveness). Fig. 3A shows that, overall, there is strong evidence that the intervention is cost-effective for $\lambda>$ $800$ because the 1/32 SI does not include near-zero values. For $250 < \lambda < 800$, there is moderate evidence that the intervention is cost-effective because near-zero values, while not in the 1/8 SI, are a member of the 1/32 SI. However for $\lambda < 250$, there is only weak evidence that the intervention is cost-effective because the 1/8 SI includes near-zero values.

Fig. 3B displays the evidence about the interaction coefficient, $\gamma_{TX\cdot Black}$, in regression equation (3). It shows that there is inconclusive or weak evidence that the cost-effectiveness of treatment varies by race. The 1/8 and 1/32 SIs contain very small values for the coefficient, including zero, but they also contain very large values, including the best supported hypotheses, suggesting an interaction may be present. Shorter support intervals are needed, but for that more data are required. Importantly, we noted that the support intervals from the robust profile likelihood function were, on average, 22% wider (range 15–29% over $\lambda$) than those derived from the $t$-distribution.

To explore this issue, we conducted an analysis stratified by race, which produced moderate evidence that the intervention was cost-effective for whites at almost any $\lambda$, while just the opposite was true for blacks (there was only very weak evidence that the intervention was cost-effective for blacks). This is seen in Fig. 4A and B, where the
evidence supports a larger effect of the intervention for whites than for blacks. Note that, at each \( \lambda \), the best supported hypothesis for the interaction coefficient (solid line in Fig. 3B) is essentially the difference between the best supported hypotheses for blacks and whites (solid line in Fig. 4B minus the solid line in Fig. 4A). Further exploration of the data shows that the primary difference by race is seen in subjects receiving “usual care”; black subjects receiving standard care enjoy much lower net benefits compared to their white counterparts (Hoch, 2005). The wide support intervals for the interaction coefficient in the regression model result from a violation of the regression model’s equal variance assumption. Here a case can be made that the more flexible stratified analysis is to be preferred to the classic approach of ‘borrowing strength across groups’ with a regression model that (incorrectly) assumes the variance is the same for all race-by-intervention subgroups.

4. Discussion

The original economic evaluation of these data (Lehman et al., 1999) recognized the discrepancy in the intervention’s cost-effectiveness for black and white subjects. The authors suggested that the pattern of usual care for homeless persons with SPMI varies according to race in Baltimore and concluded that the overall lower efficiency of ACT for black subjects in producing stable housing suggested that more attention should be given in the program to differences between races on patterns of homelessness and service utilization. The likelihood analysis echoes this assessment, presenting these findings in a vivid fashion. Note here that in interpreting these results, there is no adjustment for prior distributions, the number of looks at the data, or the number of statistical comparisons. This is why these evidential
descriptions do not change from investigator to investigator and why the CEAC and its implications do. The likelihood analysis simply describes the strength of the evidence in the data.

4.1. Do we all get the same answer?

To some, discussions about statistical methodologies (e.g., the frequentist, Bayesian, Likelihood debate) are only a matter of semantics. Fig. 5A and B address this viewpoint. Fig. 5A plots the CEAC curve as a frequentist should for the data analyzed in Section 3 if the analysis were part of a cost-effectiveness analysis calculating ICERs or INBs for 2, 3, 4 or 5 other patient outcomes. In each case, we have adjusted the CEAC for the multiple comparisons using several different adjustment techniques (Wright, 1992). The variety of CEAC curves for the same data set suggests that frequentist inference about cost-effectiveness depends on more than just the data (in this case, it clearly depends on how many additional comparisons the investigators conduct and the method used to make the adjustment). Fig. 5A illustrates that such adjustments can make a difference.

---

21 The US Food and Drug Administration (FDA), for example, would require such adjustments if the cost-consequence analysis were to be used in a new drug application.
Fig. 5. (A and B) Multiple, statistically valid CEAC curves for the cost-effectiveness of assertive community treatment (ACT). These curves are derived from the same data and same working model; only the specification of the sample space (A) and prior distribution (B) varies. (A) Plots the CEAC curve as a frequentist should for the data analyzed in Section 3 if the analysis were part of a cost-effectiveness analysis with 2, 3, 4 or 5 other patient outcomes. In addition, for each case, we have adjusted the CEAC for the multiple comparisons using different adjustment techniques. (A) Illustrates that the method of adjustment can produce different findings. (B) Plots the CEAC curves for the data from Section 3 under a variety of different assumptions about the prior distribution. Each CEAC in (B) is valid for the data set used. Under some assumptions, the CEAC indicates cost-effectiveness for almost any $\lambda$ value; while in other cases, this is only true for large values of $\lambda$. The same strength of evidence in the data yields different findings depending on one’s prior beliefs. Note: $\Delta_1 E$ and $\Delta_1 C$ represent the variances of the difference in effect and cost means, respectively; $\phi^2_e$ and $\phi^2_c$ (the variances of the difference in effect and cost means, respectively), respectively.

Likewise, Fig. 5B displays the CEAC curves for the data from Section 3 under a variety of different assumptions about the prior distribution (see Appendix B). That is, each CEAC in Fig. 5B is valid for the data set used in the empirical example. And, as in the frequentist analysis, each curve tells a different story about the cost-effectiveness of ACT. Under some assumptions, the CEAC indicates that ACT is cost-effective for almost any value of $\lambda$; while others indicate that this is only true for large values of $\lambda$. The CEACs diverge less when there is stronger evidence. For weak evidence regions (e.g., $\lambda < 250$), it is the prior driving the conclusions (i.e., one’s conclusions strongly rely on one’s prior beliefs). As the strength of evidence increases (with larger values of $\lambda$), the variability in the probability of cost-effectiveness decreases because the effect of the differing priors is being washed out by the evidence. Interpreting each CEAC in Fig. 5B as representing solely the evidence in the data about cost-effectiveness is incorrect. The same data and the same model yield different answers depending on one’s prior beliefs. In a larger sense, the problem is not
with the different prior distributions\textsuperscript{22}; rather it is with how we interpret CEACs in the first place. These curves show exactly what they are designed to show: an individual’s new belief about the cost-effectiveness of an intervention after his/her old belief has been updated by the data. The CEAC communicates something in additional to the strength of evidence in the data.

Using the same data and the same model, we see that frequentist and Bayesian analyses provide a range of different CEAC curves depending on the specified sample space or prior distribution. With different expected losses or gains or with different prior beliefs, analysts can recommend different actions or beliefs with exactly the same evidence in the data. Hence there are real differences and real consequences to choosing one statistical methodology over the other. However, it is our observation that, in practice, most economic evaluations intuitively minimize these differences by ignoring the sample space or by using a non-informative prior. As a result, it is the likelihood function that is being indirectly interpreted when the sample space is ignored or the prior is vague, suggesting implicit value in examining the likelihood function directly.

4.2. Decision making and the use of evidence from a trial-based economic evaluation

As newer and more expensive treatments are developed and tested, clinical enthusiasm (based on clinical trials) must be coupled with economic discipline (based on economic evaluation of those trials). While conducting economic evaluations with patient level data is an essential first step, it is clear that decisions about whether a new treatment should be covered by a third party payer (e.g., the government) are almost always based on more than just the results from a clinical trial. For this reason, we think it is essential for cost-effectiveness analysis of clinical trial data to report the evidence in the data as a stand alone quantity. Reporting evidence as a distinct entity allows for statements like “The clinical trial provided evidence that the new intervention is cost-effective; however, the intervention should not be recommended for coverage in this context because other relevant factors, external to study, suggest the intervention is not economically attractive.” The likelihood paradigm provides a way of separating recommendations and convictions from the evidence at hand. Action and belief change by culture and according to alliances; evidence does not.

Others, in different contexts, who encounter the findings will then be able to use the reported evidence more readily. Given the importance of context, it would seem that the utility of an economic evaluation from clinical trial data is not the researcher’s new set of beliefs or recommend set of actions; the benefit is in knowing the answer to the question, “Based on the data alone, what is the evidence that the new treatment is cost-effective?” An economic evaluation using likelihood methods provides an objective answer to this question by measuring and reporting the evidence of cost-effectiveness in the data.

If decision makers wish to combine the cost-effectiveness evidence from a variety of clinical trials (and perhaps weight them as well), this is easily accomplished if the selected analyses have reported the evidence of cost-effectiveness via the likelihood function\textsuperscript{23}. To be clear, we are not suggesting that evidential analysis is the only relevant statistical tool for analyzing data, particularly as most situations involve other evidence from outside of the trial. However, researchers should provide a summary of the evidence of cost-effectiveness in the data they analyze. Readers can then apply their own priors, loss functions or other contextual factors they deem necessary.

Economic evaluations are used in a practical sense to provide evidence about the cost-effectiveness of a new treatment or intervention. Because likelihood methods measure the strength of statistical evidence, the benefit of applying the likelihood paradigm to CEA appears to be not only theoretically promising but also strongly supported conceptually. It is our opinion that analysts should consider reporting and illustrating statistical evidence in this fashion. This does not preclude using frequentist or Bayesian methods to answer other central questions as well\textsuperscript{24}. Moreover, it is not

\textsuperscript{22} After looking at Fig. 5B it is tempting to label some priors as acceptable or unacceptable. Such labels illustrate our reservations. A likelihood analysis avoids such distractions and focuses on what the data say and not on what is (or is not) an acceptable or reasonable prior for these data.

\textsuperscript{23} In fact, in the likelihood context this amounts to nothing more than multiplying the study specific likelihood functions to get a single likelihood function that represents the combined evidence.

\textsuperscript{24} For example, it is possible that the data, while representing strong statistical evidence for $H_1$ over $H_2$, do not represent strong enough evidence to make $H_1$ appear more probable than $H_2$ (Blume, 2002). Yet a Bayesian analysis could show that $H_2$ is indeed more probable if we were willing to use a strong enough prior. Likewise, a frequentist analysis could show that we should act as if $H_2$ were true if we were willing to adjust for multiple comparisons that have yet to be performed. With the likelihood method, statistical evidence is measured by the data dependent factor that uniformly modifies all beliefs (and subsequently action), no matter what their initial magnitude. That factor is the likelihood ratio. The beauty of the likelihood
difficult to illustrate “evidence of cost-effectiveness” using the likelihood paradigm and the NBRF. One simply plots the 1/8 and 1/32 support interval endpoints as a function of willingness to pay (S-plus functions for this purpose are available from the authors).

5. Concluding comments

Detailed and technical analyses of the likelihood paradigm’s virtues currently exist (Blume, 1999, 2002, 2005; Blume et al., 2007; Royall, 1997, 2000a,b). This paper reflects on the nature of economic evaluation and what makes its questions amenable to answers provided by likelihood methods. Appreciation of the link between seeking evidence of cost-effectiveness and measuring evidence with likelihood ratios is enhanced within the net benefit regression framework. This paper’s empirical example demonstrates how the application of likelihood methods to economic evaluation can provide a non-controversial way to represent, quantify and illustrate statistical evidence.

Lastly, we wish to emphasize that action, belief, and evidence are distinct concepts. It is not enough to answer the question “What do these data say?” by explaining what one believes or how one should act. Likewise for statistical methodology, the techniques for answering questions such as “What should I do?” and “What should I believe?” cannot be used to answer the question “What do the data say?”. When speaking of statistical inference, it is necessary to distinguish between these three questions, not only because the methodologies required to address them are different, but also because each question is important in its own right.

Acknowledgements

The authors are grateful to Anthony Lehman at the University of Maryland, for the use of the data. Valuable comments on an earlier draft of this paper were received from Andrew Briggs, Carolyn Dewa, Tony Lancaster and two anonymous reviewers. The views expressed and any omissions are, however, those of the authors alone. During the development of this paper, Dr. Hoch was a recipient of the Ontario Ministry of Health and Long Term Care Career Scientist Award and also received funding from the Natural Sciences and Engineering Research Council of Canada. The Centre for Research on Inner City Health is sponsored by the Ontario Ministry of Health and Long-Term Care. The opinions, results, and conclusions are those of the authors and no endorsement by the Ontario Ministry of Health and Long Term Care is intended or should be inferred.

Appendix A. The robust profile likelihood under a regression model

Let our working regression model be \( Y \sim N(X\beta + Z\gamma, \sigma^2 I_n) \) where \( Y=(Y_1, \ldots, Y_n) \) is an \( n \times 1 \) vector of independent responses, \( X \) is \( n \times 1 \) vector of the covariate of interest, \( Z \) is an \( n \times r \) matrix of additional covariates, \( I_n \) is the identity matrix, \( \sigma^2 \) is a constant, \( \gamma \) is a \( r \times 1 \) vector or covariate parameters, and \( \beta \) is the scalar parameter of interest (Blume, 2005). We wish to measure the evidence about \( \beta \) and hence the one-dimensional likelihood for \( \beta \) is desired. Rather than relying on the marginal distribution from our working model, we can construct a profile likelihood for \( \beta \) (by maximizing over both \( \gamma \) and \( \sigma^2 \) at every \( \beta \)). This yields

\[
L_p(\beta) = \max_{\gamma, \sigma^2} L(\beta, \gamma, \sigma^2) \propto [\hat{\sigma}^2(\beta)]^{-n/2} \exp \left[ -\frac{(Y - X\hat{\beta} - Z\hat{\gamma}(\beta))^T (Y - X\hat{\beta} - Z\hat{\gamma}(\beta))}{2\hat{\sigma}^2(\beta)} \right]
\]

\[
\propto \left[ (Y - X\hat{\beta} - Z\hat{\gamma}(\beta))^T (Y - X\hat{\beta} - Z\hat{\gamma}(\beta)) \right]^{-n/2}
\]

where \( \hat{\gamma}(\beta) = (Z^T Z)^{-1} Z^T (Y - X\hat{\beta}) \) and \( \hat{\sigma}^2(\beta) = (Y - X\hat{\beta} - Z\hat{\gamma}(\beta))^T (Y - X\hat{\beta} - Z\hat{\gamma}(\beta)) / n. \) Notice that the residual vector \( R = Y - X\hat{\beta} - Z\hat{\gamma}(\beta) \) can be re-expressed as \( R = (Y - Z(Z^T Z)^{-1} Z^T Y) - (X - Z(Z^T Z)^{-1} Z^T X)\hat{\beta} \) to show that the profile likelihood is just the likelihood obtained from regressing the ‘corrected’ covariates of interest \( X^* = X - Z(Z^T Z)^{-1} Z^T X \) on the ‘corrected’ response \( Y^* = Y - Z(Z^T Z)^{-1} Z^T Y. \)
It follows that the robust profile likelihood is given by $[L_P(\beta)]^E$ where

$$E = \frac{X^s^T X^s R^T R/n}{X^s^T \text{diag}(R^T R)X^s} = \frac{(X^s^T X^s)^{-1} R^T R/n}{(X^s^T X^s)^{-1} X^s \text{diag}(R^T R)X^s(X^s^T X^s)^{-1}}$$

Notice that the exponential adjustment factor $E$ is simply the ratio of the model based variance estimate to the robust variance estimate (i.e., the sandwich estimator) but using the ‘corrected’ data. The associated support intervals must be obtained numerically.

This robust adjusted profile likelihood will, in large samples, have approximately the same probability of observing misleading evidence as if we knew the true model. This holds even if the underlying distribution is not normal or if the errors are in fact correlated. However a necessary condition for this robustness property is that the MLE for $\beta$ under the working model is consistent. A sufficient condition for that consistency is that the mean structure be correctly specified or omitted covariates be uncorrelated with the regressor of interest $X$. The reader is referred to a technical report from the second author examining sensitivity to this assumption.

**Appendix B. Derivation of CEAC curves under Bayesian framework**

Assume that for each of the $i = 1, \ldots, n_0$ participants of the control arm we measure their effects, $e_{0i} \sim N(\mu_{0}, \sigma^2_{e0})$, and costs, $c_{0i} \sim N(\mu_{0}, \sigma^2_{c0})$, where corr($e_{0i}, c_{0i}$) $= \rho$. Likewise for the $j = 1, \ldots, n_1$ participants on the intervention arm with $e_{1j} \sim N(\mu_{1}, \sigma^2_{e1})$ and $c_{1j} \sim N(\mu_{1}, \sigma^2_{c1})$ where corr($e_{1j}, c_{1j}$) $= \rho$. For simplicity we assume that these variances and correlations are known and equal to their sample estimates (in large samples this assumption has little, if any, impact).

The average net benefit for each arm is $NB_k(\lambda) = \bar{e}_k \lambda - \bar{c}_k \sim N(\mu_k(\lambda), \sigma^2_k(\lambda))$ where $\mu_k(\lambda) = \mu_k \lambda - \mu_c$ and $\sigma^2_k(\lambda) = (\sigma^2_{e_k} \lambda^2 + \sigma^2_{c_k} - 2\rho\lambda\sigma_{e_k}\sigma_{c_k})/n_k$ with $k = 0, 1$. Hence, the incremental net benefit, $INB(\lambda) = NB_1(\lambda) - NB_0(\lambda)$, is also normally distributed with mean $\bar{\beta}(\lambda) = \mu_1(\lambda) - \mu_0(\lambda)$ and variance $\sigma^2(\lambda) = \sigma^2_1(\lambda) + \sigma^2_0(\lambda)$. The standard frequentist CEAC curve, which is a function of the $p$-value, is simply $P(INB(\lambda) > 0) = \Phi\left[\bar{\beta}(\lambda)/\sigma(\lambda)\right]$.

Because $\sigma^2(\lambda)$ is known, we only need a prior for the true mean incremental net benefit. That is, we specify $\beta(\lambda) \sim N(\beta_0, \tau^2_0)$, with $\beta_0 = \Delta_c \lambda - \Delta_e$ and $\tau^2_0 = \lambda^2 \phi^2_e + \phi^2_c - 2\rho\lambda\phi_e\phi_c$ where the hyperparameters $\Delta_e, \Delta_c$ (the average differences in effect and cost means, respectively), $\phi^2_e$, and $\phi^2_c$ (the variances of the difference in effect and cost means, respectively) must be specified. We parameterized the prior distribution in this way to assure that the mean and variance would vary properly with $\lambda$. Given these hyperparameters, the posterior distribution is $\beta(\lambda)INB(\lambda) \sim N(\beta_1, \tau^2_1)$ where $\beta_1 = (\beta_0/\tau^2_0 + \beta(\lambda)/\sigma^2(\lambda))/(1/\tau^2_0 + 1/\sigma^2(\lambda))$ and $\tau^2_1 = \tau^2_0 \sigma^2(\lambda)/(\tau^2_0 + \sigma^2(\lambda))$ (Gelman, 1995). From the posterior distribution, it is easy to calculate the probability that the intervention is cost-effective as $P(\beta(\lambda) > 0) = \Phi\left[\beta_1/\sqrt{\tau^2_1}\right]$.

We then estimate this probability by using our data to estimate $\mu_{ek}, \sigma^2_{ek}, \mu_{ck}, \sigma^2_{ck}, \rho$ for $k = 0, 1$ and by choosing hyperparameters $\Delta_e, \Delta_c, \phi^2_e$, and $\phi^2_c$. The resulting Bayesian CEAC curve is $P(\beta(\lambda) > 0)$ as a function of $\lambda$. Note that, as expected, $P(\beta(\lambda) > 0) \rightarrow P(INB(\lambda) > 0)$ as $\phi^2_e$ and $\phi^2_c$ converge to infinity at appropriate rates.

**References**


