Statistics
in Medicine

Research Article

Received 29 July 2015, Accepted 21 April 2016 Published online 18 May 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6987

Probabilistic measures of cost-effectiveness

Ionut Bebu,[a*†] Thomas Mathew[b] and John M. Lachin[a]

Several probability-based measures are introduced in order to assess the cost-effectiveness of a treatment. The basic measure consists of the probability that one treatment is less costly and more effective compared with another. Several variants of this measure are suggested as flexible options for cost-effectiveness analysis. The proposed measures are invariant under monotone transformations of the cost and effectiveness measures. Interval estimation of the proposed measures are investigated under a parametric model, assuming bivariate normality, and also non-parametrically. The delta method and a generalized pivotal quantity approach are both investigated under the bivariate normal model. A non-parametric U-statistics-based approach is also investigated for computing confidence intervals. Numerical results show that under bivariate normality, the solution based on generalized pivotal quantities exhibits accurate performance in terms of maintaining the coverage probability of the confidence interval. The non-parametric U-statistics-based solution is accurate for sample sizes that are at least moderately large. The results are illustrated using data from a clinical trial for prostate cancer therapy.

Keywords: cost-effectiveness; incremental cost effectiveness ratio (ICER); incremental net benefit (INB); generalized pivotal quantity; U-statistics

1. Introduction

Identifying interventions that are cost-effective is clearly an important objective [1, 2]. No single number can capture the multifaceted interplay between costs and benefits, and therefore, various measures of cost-effectiveness that are informative, intuitive, and simple to explain are required to make informed decisions, along with statistical procedures that are accurate and robust with respect to assumptions. Several such measures have been proposed in the literature [3]. The incremental cost-effectiveness ratio (ICER) is defined as the ratio between the difference of expected costs and the difference of expected effectiveness. Although very easy to interpret as the additional cost per unit of effectiveness gained, being a ratio, the ICER is difficult to interpret in certain situations. For example, when the difference in effectiveness is close to zero, the ICER approaches $\pm\infty$ depending on the sign of the difference in cost. To address this issue, another measure of cost-effectiveness has been proposed, the incremental net benefit (INB), that is, the difference between the incremental cost and the incremental effectiveness after multiplying the latter with a willingness-to-pay parameter. Both measures are functions of population means and therefore describe average effects. Various approaches, both parametric (e.g., delta method and Fieller's theorem) and non-parametric (several bootstrap versions) have been proposed to construct confidence intervals for ICER and INB [4–11].

An alternative approach is to conceptualize the relative performance of an intervention at the individual level. In other words, when comparing two potential treatments for a particular subject, it is important to assess the probability that the first treatment will be less costly and more effective compared with the second one. The cost-effectiveness measure proposed herein, named cost-effectiveness probability (CEP), is this probability. Randomized clinical trial data can be used to evaluate this parameter.

A useful property of the proposed CEP is that it is invariant with respect to (possibly different) monotone transformations of cost and effectiveness, a property which is not shared by ICER and INB [3, 12].

[a]*The Biostatistics Center, Department of Epidemiology and Biostatistics, The George Washington University, 6110 Executive Blvd., Rockville, MD 20852, U.S.A.*
[b]*Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A.*
[*]*Correspondence to: Ionut Bebu, The Biostatistics Center, The George Washington University, 6110 Executive Blvd., Rockville, MD 20852, U.S.A.*
[†]*E-mail: ibebu@bsc.gwu.edu*

Moreover, because these measures are defined in terms of expected values of cost and effectiveness, their interpretation and applicability may be more difficult with skewed data [13], which is usually the norm, especially for cost [14]. This is not the case for the proposed CEP. A somewhat similar measure of cost-effectiveness is a variation of the cost-effectiveness acceptability curve (CEAC) proposed by Willan in [15]. Unlike the CEP, the CEAC is the probability that the net benefit is larger for the first treatment than for the second one. However, this measure depends on the willingness-to-pay parameter.

Here, we would like to emphasize that the proposed probabilistic measures are not meant to replace the traditional measures such as the ICER and INB that are functions of population averages. Rather, the proposed measures go beyond averages and complement the traditional approaches by considering proportions corresponding to subsets of the population for which a treatment is cost-effective.

The paper is organized as follows. We start by introducing the CEP parameters in the next section. Then, both parametric and non-parametric statistical methods are used to derive tests and confidence intervals for such measures. The parametric approaches investigated include the delta method and a method based on the generalized pivotal quantity (GPQ), while the non-parametric approach is based on U-statistics. The methods are compared through simulations. The methodology is illustrated using data from a clinical trial for prostate cancer therapy [3]. The paper concludes with a brief discussion.

2. The probabilistic measures

Consider a randomized trial with two arms, and let (C, E) be the bivariate random variable denoting the cost (lower is better) and effectiveness (higher is better). Let (C_1, E_1) and (C_2, E_2) be random variables denoting the cost and effectiveness for two treatments. Furthermore, let μ_{C1} and μ_{C2} be the population mean costs corresponding to the first and second treatments, respectively, and let μ_{E1} and μ_{E2} respectively denote the corresponding population mean effectiveness parameters. Here, we shall consider the case of continuous costs and continuous effectiveness only. The standard approaches in evaluating the cost-effectiveness of the first treatment over the second include the use of the incremental ICER, defined as

$$ICER = \frac{\mu_{C1} - \mu_{C2}}{\mu_{E1} - \mu_{E2}},$$

or the INB

$$INB = \lambda(\mu_{E1} - \mu_{E2}) - \left(\mu_{C1} - \mu_{C2}\right),$$

where the quantity λ is referred to as the willingness-to-pay parameter, defined as the monetary value assigned by the payer to each unit of effectiveness gained [16]. While these are reasonable measures in terms of overall average cost-effectiveness, they do not provide information concerning the proportion of subjects for whom a treatment is cost-effective. Our probabilistic measures are meant to provide such information.

Assume a random distribution of costs and a measure of effectiveness among the subjects within the two treatment groups. For two subjects, one from each group, the first treatment is more cost-effective than the second treatment if

$$C_1 \leqslant C_2 \text{ and } E_1 \geqslant E_2. \tag{1}$$

We note that the region defined by (1) is the southeast quadrant of the cost-effectiveness plane. Define the cost-effectiveness proportion (CEP) as the probability that the intervention is more cost-effective for a random subject in the first group than for a random subject in the second group,

$$CEP = P\left(C_1 \leqslant C_2, \quad E_1 \geqslant E_2\right). \tag{2}$$

Clearly, large values of the *CEP* are desirable. An interesting feature of (2) is that it is invariant with respect to monotone transformations of both cost and effectiveness. Therefore, with continuous cost and effectiveness, one can assume that cost and effectiveness follow a bivariate normal distribution after some monotone transformation. This is of course under the assumption that transformation to normality is possible.

The parameter *CEP* and appropriate variants of it offer considerable flexibility in the assessment of cost-effectiveness. For example, if the first treatment is expected to be not only more effective but also more costly (i.e., the northeast quadrant of the cost-effectiveness plane), *CEP* can be modified as

$$CEP(\delta_C) = P\left(C_1 \leqslant C_2 + \delta_C, \quad E_1 \geqslant E_2\right),$$

for a specified upper limit δ_C of the increase in cost. Cost-effectiveness can be concluded if $CEP(\delta_C)$ is large. The quantity δ_C can be thought of as a willingness-to-pay parameter. Conversely, one can also determine the threshold δ_C after specifying a value for $CEP(\delta_C)$. A further modification could be

$$CEP\left(\delta_C, \delta_E\right) = P\left(C_1 \leqslant C_2 + \delta_C, \quad E_1 \geqslant E_2 + \delta_E\right),$$

which includes another threshold δ_E for the effectiveness. The motivation for including such a threshold is that it may not be enough to have increased effectiveness; we would like to have a 'clinically meaningful improved effectiveness', specified in terms of δ_E.

Figure 1 describes the regions of interest for $CEP(\delta_C, \delta_E)$. The right panel of Figure 1 gives a plot of the region where INB is positive, which is given by $\Delta_C \leqslant \lambda \cdot \Delta_E$, where $\Delta_C = \mu_{C1} - \mu_{C2}$ and $\Delta_E = \mu_{E1} - \mu_{E2}$. It is clear that the INB criterion can select an intervention that is not only less costly but also less effective (shaded area in the south-west quadrant). On the other hand, the plot on the left panel shows that as a criterion, the CEP is defined in terms of a region that is free from the aforementioned drawback. Furthermore, δ_C and δ_E can be subjectively chosen without relying on the data. This certainly does not rule out the possibility of data driven choices of these quantities.

Yet another possibility is to consider conditional probabilities such as $P(C_1 \leqslant C_2 \,|\, E_1 \geqslant E_2)$, which can be used to investigate how likely it is that the first treatment is less costly for a subject for whom it is also more effective.

Regarding the definition of the CEP in (2), a point to remember is that the probability in (2) and the probability $P\left(C_1 \geqslant C_2, \quad E_1 \leqslant E_2\right)$ can both be significantly different from zero. The latter probability corresponds to the northwest quadrant of the cost-effectiveness plane and can be used to assess the proportion of patients for whom the first treatment is more costly and less effective. Therefore, treatment 1 is preferable to treatment 2 if the probability of falling in the southeast quadrant is larger than the probability of the northwest quadrant. In view of this, the following parameter, denoted by ΔCEP, can be used to assess the effectiveness of the first treatment over the second:

$$\Delta CEP = P\left(C_1 \leqslant C_2, E_1 \geqslant E_2\right) - P\left(C_1 \geqslant C_2, E_1 \leqslant E_2\right).$$

A positive value of ΔCEP indicates that the first treatment is to be preferred over the second; a negative value leads to the opposite conclusion. Through algebraic manipulation of joint and marginal probabilities, it can be shown that $\Delta CEP = P(C_1 \leqslant C_2) - P(E_1 \leqslant E_2)$.

In the next section, we shall develop inference concerning the parameter CEP defined in (2); however, the methodologies can be easily adapted for inference concerning the other parameters defined previously.

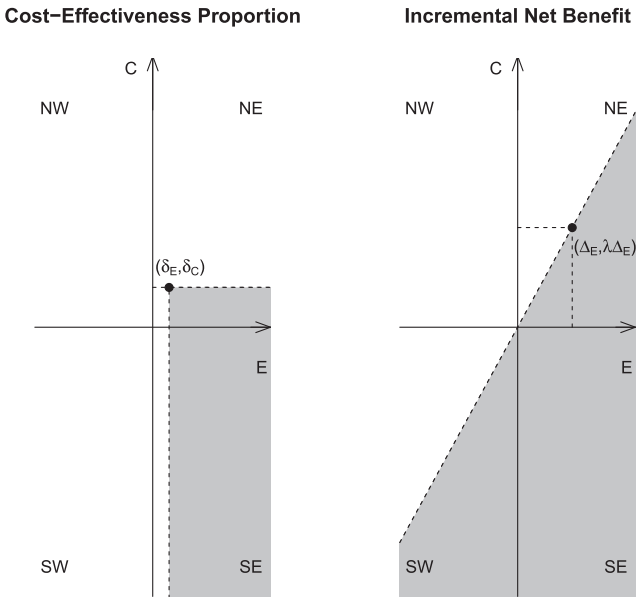

Figure 1. Regions of interest (shaded areas) described by the $CEP(\delta_C, \delta_E)$ (left) and INB (right). CEP, cost-effectiveness probability; INB, incremental net benefit.

3. Statistical inference for *CEP*

Let (C_{1j}, E_{1j}), $j = 1, \ldots n_1$, and (C_{2k}, E_{2k}), $k = 1, \ldots n_2$, be random samples of sizes n_1 and n_2 from the bivariate distributions of (C_1, E_1) and (C_2, E_2), respectively. We shall now investigate both parametric and non-parametric inferences for the parameter *CEP* in (2). We shall also briefly indicate how our methods can be adopted for inference concerning the other probabilistic measures, such as ΔCEP.

3.1. Parametric inference

Parametric inference is developed under the assumption of bivariate normality. As already noted, this is justified whenever a transformation to normality is possible. Thus let

$$\left(C_i, E_i\right) \sim N\left(\mu_i, \Sigma_i\right),$$

$i = 1, 2$, where we write $\mu_i = (\mu_{Ci}, \mu_{Ei})\prime$, $\Sigma_i = \sigma_{ill'}$, and $i, l, l' = 1, 2$.

One has

$$\begin{aligned} CEP &= P(C_1 - C_2 \leqslant 0 \& E_1 - E_2 \geqslant 0) \\ &= P(C_1 - C_2 \leqslant 0) - P(C_1 - C_2 \leqslant 0 \ \& \ E_1 - E_2 < 0) \\ &= \Phi\left(0; \mu_{C1} - \mu_{C2}, \sigma_{111} + \sigma_{211}\right) - \Phi_2\left((0,0)'; (\mu_{C1} - \mu_{C2}, \mu_{E1} - \mu_{E2})', \Sigma_1 + \Sigma_2\right), \end{aligned} \tag{3}$$

where Φ and Φ_2 denote the cumulative distribution functions of the univariate and bivariate normal distributions, respectively.

A point estimate for *CEP* is obtained by replacing the population parameters in (3) with their sample estimates. The delta method [17] can be employed to derive large sample tests and confidence intervals for *CEP* and ΔCEP. Details are provided in Appendix A.

3.1.1. The generalized pivotal quantity approach. An alternative parametric approach is based on the idea of a GPQ introduced by Weerahandi [18]. We first review some notations and basic results and then introduce the GPQ approach. Further details can be found in [19].

Let X_i, $i = 1, 2, \ldots, n$, be independent $p \times 1$ observations following the multivariate normal distribution

$$X_i \sim N(\mu, \Sigma),$$

where the mean vector μ and the covariance matrix Σ are unknown parameters. An unbiased estimator of μ is $\hat{\mu} = \bar{X}$, the sample mean vector. An unbiased estimator of Σ is given by

$$\hat{\Sigma} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)' / (n-1),$$

and we have the distributions

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\Sigma\right), \ (n-1)\hat{\Sigma} \sim W_p(\Sigma, n-1),$$

where $W_p(\Sigma, n-1)$ denotes the p-dimensional Wishart distribution with scale matrix Σ and $n-1$ degrees of freedom.

A GPQ for a parameter is a function of random variables and their observed values that satisfies two conditions: (i) given the observed data, the GPQ has a distribution that is free of the unknown parameters and (ii) the 'observed value' of the GPQ (obtained by replacing the random variables by their respective observed values) is free of any nuisance parameters and is often equal to the parameter of interest [18, 19]. Let $\bar{X}_o$ and $\hat{\Sigma}_o$ denote the respective observed values of $\bar{X}$ and $\hat{\Sigma}$. Then the GPQs for Σ and μ, say T_Σ and T_μ, have the following representations [20]:

$$T_\Sigma = \left[W_p\left(\hat{\Sigma}_o^{-1}, n-1\right)\right]^{-1}$$
$$T_\mu = \bar{X}_o - T_\Sigma^{1/2} \cdot Z/\sqrt{n}, \tag{4}$$

so that T_Σ is the inverse of a random draw from a Wishart distribution with scale matrix $\hat{\Sigma}_o^{-1}$ and $n-1$ degrees of freedom. Here, Z is a random draw from $N(0, I_p)$, and $T_\Sigma^{1/2}$ denotes the positive definite square root of T_Σ. A GPQ for the CEP (say, T_{CEP}) is obtained by replacing the parameters in (3) with the respective GPQ of each in (4). That is,

$$\begin{aligned} T_{CEP} = {} & \Phi\left(0; T_{\mu_{C1}} - T_{\mu_{C2}}, T_{\sigma_{111}} + T_{\sigma_{211}}\right) \\ & - \Phi_2\left((0,0)'; (T_{\mu_{C1}} - T_{\mu_{C2}}, T_{\mu_{E1}} - T_{\mu_{E2}})', T_{\Sigma_1} + T_{\Sigma_2}\right), \end{aligned} \tag{5}$$

where $T_{\mu_i} = (T_{\mu_{Ci}}, T_{\mu_{Ei}})$ and T_{Σ_i} are the GPQs for the mean vector and variance–covariance matrix in group i, $i = 1, 2$. These GPQs are obtained as given in (4).

The percentiles of the distribution of T_{CEP} provide confidence limits for the *CEP*. Because these quantiles do not have a closed form, a Monte Carlo approach can be employed as follows.

(1) Using the given data, compute the observed values $\bar{X}_{oi}$ and $\hat{\Sigma}_{oi}$, $i = 1, 2$.
(2) For $i = 1, 2$, generate $T_{\Sigma_i} \sim \left[W_p\left(\hat{\Sigma}_{oi}^{-1}, n-1\right)\right]^{-1}$ and $Z_i \sim N(0, I_2)$, all independent.
(3) Let $T_{\mu_i} = \bar{X}_{oi} - T_{\Sigma_i}^{1/2} \cdot Z_i/\sqrt{n_i}$, $i = 1, 2$.
(4) Obtain T_{CEP} using (5).
(5) Repeat steps 1–4 M times and obtain M values for T_{CEP}. The 2.5th and 97.5th percentiles of this vector provide a 95% confidence interval for the *CEP*.

In order to obtain a GPQ for the measure ΔCEP defined in Section 2, we note that

$$\Delta CEP = \Phi(0; \mu_{C1} - \mu_{C2}, \sigma_{111} + \sigma_{211}) - \Phi(0; \mu_{E1} - \mu_{E2}, \sigma_{122} + \sigma_{222}). \tag{6}$$

A GPQ for ΔCEP can now be obtained by replacing the parameters in (6) with their GPQs in (4), and confidence intervals can be constructed using the Monte Carlo approach described previously.

3.2. Non-parametric inference

The non-parametric bootstrap can be easily implemented to construct tests and confidence intervals for the *CEP* in (2). Bootstrap samples are drawn from each group, where the selection is at the subject level. This assures that the within-subject correlation between cost and effectiveness is preserved and accounted for. Implementation of the bootstrap requires the evaluation of the quantity defined in Equation (8) for each bootstrap sample. However, notice that as the sample sizes increase, evaluating (8) for each bootstrap sample is computationally very expensive. Thus, an alternative approach, based on U-statistics, is proposed.

3.2.1. A U-statistic approach. For an inference on CEP without any distributional assumptions, let

$$u((C_{1j}, E_{1j}), (C_{2k}, E_{2k})) = \begin{cases} 1, & \text{if } C_{1j} \leqslant C_{2k} \ \& \ E_{1j} \geqslant E_{2k} \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

An unbiased estimator of the CEP can be obtained using the U-statistic

$$U = \frac{1}{n_1 \cdot n_2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} u\left(\left(C_{1j}, E_{1j}\right), \left(C_{2k}, E_{2k}\right)\right). \tag{8}$$

Statistical inference for the CEP can be based on the large sample distribution of U-statistics [17]:

$$\sqrt{N}(U - CEP) \sim N\left(0, \sigma_U^2\right), \tag{9}$$

where

$$\sigma_U^2 = \frac{N}{n_1}\xi_{10} + \frac{N}{n_2}\xi_{01}, \tag{10}$$

with

$$\xi_{10} = Cov\left(u\left(X_1, Y_1\right), u\left(X_1, Y_1'\right)\right)$$
$$\xi_{01} = Cov\left(u\left(X_1, Y_1\right), u\left(X_1', Y_1\right)\right) \ ,$$

where X_1 $(= (C_1, E_1))$ and X_1' refer to values of cost and effectiveness for two different subjects from the first group, and likewise Y_1, Y_1' from the second group, all independent, and $N = n_1 + n_2$.

Let $X \succ Y$ denote the event that subject X has lower cost and higher effectiveness than subject Y, that is, satisfies (1). Then, the terms in (10) can be further simplified, for example,

$$\begin{aligned} \xi_{10} &= P\left(X_1 \succ Y_1 \,\&\, X_1 \succ Y_1'\right) - \left[P\left(X_1 \succ Y_1\right)\right]^2 \\ \xi_{01} &= P\left(X_1 \succ Y_1 \,\&\, X_1' \succ Y_1\right) - \left[P\left(X_1 \succ Y_1\right)\right]^2 . \end{aligned} \tag{11}$$

Let

$$S_{jk} = u\left(X_j, Y_k\right) \ , \quad S_{j\cdot} = \sum_k S_{jk} \, , \quad S_{\cdot k} = \sum_j S_{jk},$$

$j = 1, \ldots, n_1$, $k = 1, \ldots, n_2$. The terms in (11) can be estimated as follows:

$$\begin{aligned} \hat{P}(X_1 \succ Y_1) &= \frac{1}{n_1 \cdot n_2} \sum_{j,k} S_{jk} \\ \hat{P}\left(X_1 \succ Y_1 \,\&\, X_1 \succ Y_1'\right) &= \frac{\sum_j S_{j\cdot}\left(S_{j\cdot} - 1\right)}{n_1 \cdot n_2 \cdot \left(n_2 - 1\right)} \\ \hat{P}\left(X_1 \succ Y_1 \,\&\, X_1' \succ Y_1\right) &= \frac{\sum_k S_{\cdot k}\left(S_{\cdot k} - 1\right)}{n_2 \cdot n_1 \cdot \left(n_1 - 1\right)} . \end{aligned} \tag{12}$$

Then hypothesis tests and confidence intervals for the *CEP* can be obtained using the asymptotic distribution in (9), with σ_U^2 estimated using (10), (11), and (12).

This approach can be easily adapted for inference regarding $CEP(\delta_C, \delta_E)$ and ΔCEP by modifying the kernel function (7) appropriately. For example, inference on ΔCEP can be based on the U-statistic

$$V = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} v\left(\left(C_i, E_i\right), \left(C_j, E_j\right)\right) \ , \tag{13}$$

with

$$v\left(\left(C_i, E_i\right), \left(C_j, E_j\right)\right) = 1_{\{C_i \leqslant C_j\}} - 1_{\{E_i \leqslant E_j\}},$$

where $1_{\{C_i \leqslant C_j\}}$ and $1_{\{E_i \leqslant E_j\}}$ are indicator functions. Further details are provided in Appendix B.

4. Numerical results

In order to assess the performance of the proposed methods, we conducted some simulations assuming bivariate normality for cost and effectiveness. It is worth reiterating that the bivariate normality alone is sufficient after (possibly different) transformations of cost and effectiveness because the measures (probabilities) are invariant to such transformations. The procedures compared in the simulations include the delta method (denoted by delta), the GPQ method (GPQ), and the non-parametric U-statistic-based approach (U-statistic). The non-parametric bootstrap was found to be too computationally intensive to be included in the simulations. The simulation setup is along the lines of a previous simulation study [21]:

$$\begin{aligned} \mu_1 &= (6, 8)' \, , \quad \Sigma_1 = \begin{pmatrix} 1.2 & \sigma_{112} \\ \sigma_{112} & 6 \end{pmatrix} \\ \mu_2 &= (\mu_{21}, 10)' , \, \Sigma_2 = \begin{pmatrix} 1.6 & \sigma_{212} \\ \sigma_{212} & 8 \end{pmatrix} , \end{aligned} \tag{14}$$

Table I. Simulated coverage probabilities of the different confidence intervals for the CEP under a bivariate normal model for cost and effectiveness and for the parameter choices in (14) and a nominal level of 95%.

n	ρ	μ_{21}	CEP	Delta	GPQ	U-statistic
50	−0.5	7	0.27	0.9470	0.9580	0.9410
50	0.1	7	0.20	0.9492	0.9548	0.9352
50	0.5	7	0.15	0.9548	0.9542	0.9252
50	−0.5	8	0.29	0.9426	0.9518	0.9348
50	0.1	8	0.26	0.9502	0.9512	0.9332
50	0.5	8	0.22	0.9412	0.9540	0.9344
100	−0.5	7	0.27	0.9550	0.9558	0.9426
100	0.1	7	0.20	0.9526	0.9526	0.9420
100	0.5	7	0.15	0.9598	0.9422	0.9308
100	−0.5	8	0.29	0.9442	0.9510	0.9398
100	0.1	8	0.26	0.9482	0.9544	0.9420
100	0.5	8	0.22	0.9512	0.9528	0.9416
150	−0.5	7	0.27	0.9478	0.9492	0.9426
150	0.1	7	0.20	0.9550	0.9512	0.9426
150	0.5	7	0.15	0.9606	0.9500	0.9416
150	−0.5	8	0.29	0.9486	0.9492	0.9462
150	0.1	8	0.26	0.9486	0.9494	0.9448
150	0.5	8	0.22	0.9498	0.9466	0.9398
200	−0.5	7	0.27	0.9496	0.9522	0.9472
200	0.1	7	0.20	0.9604	0.9562	0.9514
200	0.5	7	0.15	0.9660	0.9486	0.9440
200	−0.5	8	0.29	0.9506	0.9476	0.9480
200.	0.1	8	0.26	0.9530	0.9514	0.9484
200	0.5	8	0.22	0.9484	0.9514	0.9498

CEP, cost-effectiveness probability; GPQ, generalized pivotal quantity.

where $\mu_{21} = 7, 8$, $\sigma_{112} = \rho \cdot \sqrt{1.2 \times 6}$, $\sigma_{212} = \rho\sqrt{1.6 \times 8}$, with $\rho = -0.5, 0.1, 0.5$, and $n_1 = n_2 = 50, 100, 150, 200$. The proposed methods are compared through simulations (5000 simulations) in terms of coverage probabilities of the 95% confidence intervals, and the results are reported in Table I.

The GPQ approach provides extremely accurate confidence intervals, regardless of the sample size and parameter configurations. The U-statistic procedure is slightly liberal for small sample sizes ($n = 50$ and 100), but it performs very well for $n = 150$ and 200. The confidence intervals obtained using the delta method seem to depend on the sign of the correlation between cost and effectiveness, with over-coverage for positive correlations.

Therefore, if bivariate normality can be assumed (possibly after transformations) for cost and effectiveness, the GPQ approach is recommended regardless of sample size, while the distribution-free U-statistic approach performs well in most cases of practical importance.

5. Illustration

The proposed methods are now illustrated using data from a trial of prostate cancer therapy [3]. Briefly, a total of 114 subjects with symptomatic, hormone-resistant prostate cancer were randomized to either prednisone alone ($n_1 = 61$) or prednisone and mitoxantrone ($n_1 = 53$) and were followed until death. The effectiveness was expressed as quality-adjusted life weeks, while the costs are in Canadian dollars. The reader is referred to [3] and [4] for detailed analyses where the same dataset was used to illustrate the established measures of cost-effectiveness (such as ICER, INB, and CEAC).

After inspection of the normal quantile–quantile plots, both the cost and effectiveness were lognormally distributed, and they were used on the log scale for inference on the ICER. Similar results were obtained using different transformations in the Box–Cox family.

The prednisone and mitoxantrone treatment was less costly ($\hat{\Delta}_C = -1717.07$) and more effective ($\hat{\Delta}_E = 12.78$). Then, $\widehat{ICER} = -134.35$, with a 95% confidence interval (−11922, 569) using Fieller's theorem [3]. Because this interval includes the value zero, the ICER analysis fails to reject the null

Table II. 95% and 90% confidence intervals for ΔCEP using the prostate cancer dataset.

Level	U-statistic	Delta	GPQ
95%	(0.003, 0.234)	(−0.013, 0.212)	(−0.016, 0.209)
90%	(0.021, 0.215)	(0.004, 0.194)	(0.001, 0.191)

CEP, cost-effectiveness probability; GPQ, generalized pivotal quantity.

hypothesis that the combination of prednisone and mitoxantrone is more cost-effective than prednisone alone. However, this parameter does not provide any information regarding the proportion of subjects for which prednisone and mitoxantrone will be less costly and more effective than prednisone alone, information provided by CEP.

The MLE for the CEP obtained using (3) is 0.2366, while the U-statistic estimate (8) is 0.2416. The 95% confidence intervals for the CEP are 0.1602 and 0.3129 using the delta method, 0.1693 and 0.3099 using the GPQ approach, and 0.1699 and 0.3132 using the U-statistic approach. Using the GPQ interval as illustration, prednisone and mitoxantrone will be cost-effective (more effective and less costly) relative to prednisone alone for approximately 23.66% of the population, with a 95% confidence interval of 16.93% and 30.99%. We reiterate that such information cannot be obtained using the ICER parameter.

Let us now consider some of the other parameters mentioned in Section 2. For example, the conditional probability $P(C_1 \leqslant C_2 \mid E_1 \geqslant E_2)$ can be useful to investigate how likely is it that the first treatment is less costly for a subject for which it is also more effective. Using the prostate cancer dataset, we estimate that there is 58.66% chance for prednisone and mitoxantrone to be more effective than prednisone alone (i.e., $P(E_1 \geqslant E_2) = 0.5866$). Furthermore, there is a 23.66% chance for prednisone and mitoxantrone to be less costly and more effective compared with prednisone alone (i.e., $P(C_1 \leqslant C_2 \,\&\, E_1 \geqslant E_2) = 0.2366$). Therefore, for a subject for which the prednisone and mitoxantrone treatment is more effective, there is 40.3% (=0.2366/0.5866) chance that it is also less costly.

The values for δ_C and δ_E can be selected based on clinical relevance or patient preference. For example, assume an improvement of at least four quality-adjusted life weeks is of interest, while an increase in cost of at most 25% relative to the mean cost in the placebo group is considered acceptable, which leads to $\delta_E = 4$ and $\delta_C = 7259.7$. The estimate of $CEP(\delta_C, \delta_E)$ is 0.2923, and 95% confidence intervals using the GPQ approach and the U-statistics approach are 0.2055 and 0.3578 and 0.2145 and 0.3701, respectively.

For the prostate cancer dataset, the parameter ΔCEP is estimated to be 9.91%. Confidence intervals for ΔCEP are reported in Table II. At 5% nominal level, the U-statistic approach rejects the null hypothesis that $\Delta CEP = 0$, while the delta method and GPQ fail to do so. This is likely explained by the fact that the U-statistic approach leads to confidence intervals with coverage lower than the nominal 95% level in small sample sizes (Table I), while the parametric approaches maintain the nominal level. All methods reject the null hypothesis at 10% level.

6. Discussion

Previous measures of cost-effectiveness, such as ICER, INB, and CEAC, are functions only of the mean parameters, and therefore they describe average effects. The proposed CEP parameter and its variants, defined as the probability of higher effectiveness and lower cost for one treatment versus the other treatment, have a very simple and intuitive interpretation. Furthermore, an advantage of the CEP over the other existing measures is that it is invariant with respect to (possibly different) monotone transformations of cost and effectiveness.

Clearly, choosing one treatment over another cannot be based on just a single measure of cost-effectiveness. Thus, the proposed probabilistic measures are not meant to replace the traditional approaches based on the ICER and INB; rather, they provide complementary information that allows both the payer and the patient to make better informed decisions.

The CEP measure is defined in terms of the SE quadrant, while the ΔCEP measure is based on the SE and NW quadrants. As suggested by the Associate Editor and one of the reviewers, depending on the application, all quadrants may be of interest. The probabilities of the four quadrants can be estimated jointly both parametrically and non-parametrically. For example, the non-parametric approach can be employed by using a multivariate U-statistics with kernels corresponding to the SE, NW, and NE quadrants, and inference is then based on the large sample asymptotics for multivariate U-statistics [22].

However, the parameters considered herein can be obtained and studied using univariate U-statistics, and the multivariate case was not included to avoid unnecessary technicalities.

Three inference approaches for CEP have been employed in our work, the delta method, the GPQ approach, and a U-statistic approach. These methods were evaluated through simulations and illustrated using data from a clinical study. With the exception of the delta method, the proposed statistical inference approaches performed well. The U-statistic approach was slightly liberal for small sample sizes ($n = 50$ and 100) but performed very well for $n \geqslant 150$, while the GPQ method was extremely accurate for all sample sizes and parameter configurations considered. Several extensions were considered, including related measures of effectiveness and tests for comparing two treatments based on the CEP parameter.

Both INB and CEAC analyses involve a willingness-to-pay parameter λ. While in most applications it is possible to define λ, in others, it may be more difficult. Consider the case of retinopathy progression detected at periodic examinations in patients with type 1 diabetes. Different schedules of examinations can be compared in terms of cost (proportional to the frequency of visits) and effectiveness, quantified as the elapsed time from the actual onset of progression to the next visit (or the time that progression went undetected). The willingness-to-pay parameter λ is the monetary value assigned to each 1 unit (say 1 month) of time undetected, but it is not clear what the value should be. The parameters δ_C and δ_E considered herein are on the effectiveness and cost scales, respectively, and may be easier to define. In the retinopathy progression example, the δ_C value can be a percentage (say 10%) of the mean cost in the control group, while δ_E may be a clinically meaningful difference that can be elicited from a physician or the patient, say 2 months.

The proposed probabilistic measure of cost-effectiveness was illustrated using continuous cost and continuous effectiveness. Clearly, the CEP measure is valid more generally for ordinal cost and effectiveness. As an example, consider the case of continuous cost with pdf f and binary effectiveness. The joint distribution of (C, E) can be written as the product of the marginal distribution of cost and the conditional distribution of effectiveness given cost,

$$g_{C,E}(c, e) = P(E = e|C = c) \cdot f(c),$$

where $e = 0, 1$, $c > 0$. Then

$$CEP = A_1 + A_2 + A_3,$$

where

$$A(e_1, e_2) = \int_0^\infty \left(\int_{c_1}^\infty P\left(E_2 = e_2|C = c_2\right) \cdot f_2\left(c_2\right)\, dc_2 \right) P\left(E_1 = e_1|C = c_1\right) \cdot f_1\left(c_1\right)\, dc_1\,,$$

and $A_1 = A(0, 0)$, $A_2 = A(1, 0)$, and $A_3 = A(1, 1)$. This does not simplify even in the common case when the cost is assumed normally distributed and the effectiveness follows a logistic model. Notice that the non-parametric inference approaches (e.g., bootstrapping and U-statistics) remain valid in the general case of ordinal cost and effectiveness, while parametric approaches are less appealing because of the lack of closed form formulas in such general cases. The non-parametric methods proposed herein are also particularly appealing when dealing with highly skewed data (especially costs) and with two-part models for excess zeros [3].

The possibility of using a probabilistic measure in order to assess cost-effectiveness was attempted by Willan [15]. That measure is based on the INB and is not invariant under transformations. The measure was criticized by O'Hagan and Stevens [23], who note that the measure could lead to conclusions that contradict what one obtains from the ICER under skewed distributions for the net benefit; see figures 1 and 2 in their paper, and the related discussion. Notice that this criticism does not apply to the probability measures introduced herein, as they are invariant under monotone transformations. These measures also provide considerable flexibility, because they can be tailored to suit the purpose of the cost-effectiveness investigation. For example, the probabilistic measure can be defined so as to infer whether a certain treatment is less costly for the subjects for whom it is more efficient.

The goal of personalized medicine is to identify the best treatment for a particular subject based on his or her individual characteristics. However, the proposed CEP measures do not take into account subject-level characteristics. Rather, they are marginal or 'unadjusted personalized' measures of cost-effectiveness, because they appropriately quantify the proportion of subjects for whom one treatment is cost-effective compared with another treatment. Extending the measures presented herein to account for subject-level characteristics is currently under investigation.

Appendix A: The delta method

If

$$\sqrt{n}(B-\beta) \to N(0, \Sigma^*),$$

then

$$\sqrt{n}\left(h(B)-h(\beta)\right) \to N\left(0, \nabla h(\beta)^T \Sigma^* \nabla h(\beta)\right),$$

where ∇h denotes the gradient of h.

In our case

$$B = \left(\hat{\mu}_{C1}, \hat{\mu}_{E1}, \hat{\sigma}_{111}, \hat{\sigma}_{122}, \hat{\sigma}_{112}, \hat{\mu}_{C2}, \hat{\mu}_{E2}, \hat{\sigma}_{211}, \hat{\sigma}_{222}, \hat{\sigma}_{212}\right),$$
$$\beta = \left(\mu_{C1}, \mu_{E1}, \sigma_{111}, \sigma_{122}, \sigma_{112}, \mu_{C2}, \mu_{E2}, \sigma_{211}, \sigma_{222}, \sigma_{212}\right),$$

and

$$\Sigma^* = diag\left(\Sigma_1^*, \Sigma_2^*\right),$$

where

$$\Sigma_i^* = diag\left(\Sigma_i/n_i, \Omega_i/\left(n_i-1\right)\right),$$
$$\Omega_i = \begin{pmatrix} 2\sigma_{i11}^2 & 2\sigma_{i12}^2 & 2\sigma_{i12}\sigma_{i11} \\ 2\sigma_{i12}^2 & 2\sigma_{i22}^2 & 2\sigma_{i12}\sigma_{i22} \\ 2\sigma_{i12}\sigma_{i11} & 2\sigma_{i12}\sigma_{i22} & \left(1+\rho_i^2\right)\sigma_{i11}\sigma_{i22} \end{pmatrix},$$

with $\rho_i = \sigma_{i12}/\sqrt{\sigma_{i11}\sigma_{i22}}$, $i = 1, 2$.

Appendix B: U-statistics approach for ΔCEP

Asymptotically, one has

$$\sqrt{N}(V - \Delta CEP) \sim N\left(0, \sigma_V^2\right),$$

where

$$\sigma_V^2 = \frac{N}{n_1}\xi_{10}^V + \frac{N}{n_2}\xi_{01}^V,$$

with

$$\xi_{10}^V = Cov\left(v\left(X_1, Y_1\right), v\left(X_1, Y_1'\right)\right)$$
$$\xi_{01}^V = Cov\left(v\left(X_1, Y_1\right), v\left(X_1', Y_1\right)\right),$$

where X_1 $(= (C_1, E_1))$ and X_1' refer to values of cost and effectiveness for two different subjects from the first group, and likewise Y_1, Y_1' from the second group, all independent, and $N = n_1 + n_2$.

Sample estimates can be obtained as follows:

$$\hat{\xi}_{10}^V = \frac{1}{n_1 \cdot n_2 \cdot \left(n_2-1\right)} \cdot \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1,k\neq j}^{n_2} v\left(\left(C_i, E_i\right), \left(C_j, E_j\right)\right) \cdot v\left(\left(C_i, E_i\right), \left(C_k, E_k\right)\right)$$
$$- \left[\frac{1}{n_1 \cdot n_2} \cdot \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} v\left(\left(C_i, E_i\right), \left(C_j, E_j\right)\right)\right]^2,$$

$$\hat{\xi}_{01}^V = \frac{1}{n_1 \cdot \left(n_1-1\right) \cdot n_2} \sum_{i=1}^{n_1}\sum_{j=1,j\neq i}^{n_1}\sum_{k=1}^{n_2} v\left(\left(C_i, E_i\right), \left(C_k, E_k\right)\right) \cdot v\left(\left(C_j, E_j\right), \left(C_k, E_k\right)\right)$$
$$- \left[\frac{1}{n_1 \cdot n_2} \cdot \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} v\left(\left(C_j, E_j\right), \left(C_k, E_k\right)\right)\right]^2.$$

Acknowledgements

We are grateful to the Associate Editor and the two referees for their insightful and helpful suggestions.

References

1. Russel LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. *Journal of the American Medical Association* 1996; **276**:1172–1177.
2. O'Brien BJ, Briggs AH. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. *Statistical Methods in Medical Research* 2002; **11**:455–468.
3. Willan AR, Briggs AH. *Statistical Analysis of Cost-effectiveness Data*. Wiley: Chichester, UK, 2006.
4. Bebu I, Luta G, Mathew T, Kennedy PA, Agan BK. Parametric cost-effectiveness inference with skewed data. *Computational Statistics and Data Analysis* 2016; **94**:210–220.
5. Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in Medicine* 1999; **18**:3245–3262.
6. Cook JR, Heyse JF. Use of an angular transformation for ratio estimation in cost-effectiveness analysis. *Statistics in Medicine* 2000; **29**:2989–3003.
7. Fan MY, Zhou XH. A simulation study to compare methods for constructing confidence intervals for the incremental cost-effectiveness ratio. *Health Services & Outcomes Research Methodology* 2007; **7**:57–77.
8. Mihaylova B, Briggs HA, O'Hagan A, Thompson SG. Review of statistical methods for analyzing health care resources and costs. *Health Economics* 2011; **20**:897–916.
9. Nixon RM, Wonderling D, Grieve RD. Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. *Health Economics* 2010; **19**:316–333.
10. Polsky D, Glick HA, Willke R, Schulman K. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Economics* 1997; **6**:243–252.
11. Wang H, Zhao H. A study on confidence intervals for incremental cost-effectiveness ratios. *Biometrical Journal* 2008; **50**:505–514.
12. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 1998; **17**:283–295.
13. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *Journal of Health Economics* 2001; **20**:461–494.
14. Briggs AH, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Services Research & Policy* 1998; **3**:233–245.
15. Willan AR. On the probability of cost-effectiveness using data from randomized clinical trials. *BMC Medical Research Methodology* 2001; **1**:8.
16. Pauly MV. Valuing health benefits in monetary terms. In *Valuing Health Care: Costs, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies*, Sloan FA (ed.) Cambridge University Press: Cambridge, 1995; 99–124.
17. Lehmann EL. *Elements of Large-sample Theory*. Springer-Verlag: New York, 1999.
18. Weerahandi S. Generalized confidence intervals. *Journal of the American Statistical Association* 1993; **88**:899–905.
19. Weerahandi S. *Exact Statistical Methods for Data Analysis*. Springer-Verlag: New York, 1995.
20. Bebu I, Mathew T. Comparing the means and variances of a bivariate lognormal distribution. *Statistics in Medicine* 2008; **27**:2684–2696.
21. Jiang G, Wu J, Williams GR. Fieller's interval and the bootstrap-Fieller interval for the incremental cost effectiveness ratio. *Health Services & Outcomes Research Methodology* 2000; **1**:291–303.
22. Lehmann EL. Robust estimation in analysis of variance. *The Annals of Mathematical Statistics* 1963; **34**:957–966.
23. O'Hagan A, Stevens JW. The probability of cost-effectiveness. *BMC Medical Research Methodology* 2002; **2**:5–10.