ACCOUNTING FOR UNCERTAINTY IN COST-EFFECTIVENESS STUDIES

ACCOUNTING FOR UNCERTAINTY IN COST-EFFECTIVENESS STUDIES

By

JOANNA M. BIERNACKA, B.Sc.

A Project

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (1998) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Accounting for Uncertainty in Cost-Effectiveness Studies

AUTHOR: Joanna M. Biernacka, B.Sc. (McMaster University)

SUPERVISOR: Dr. A. R. Willan

NUMBER OF PAGES: x, 95

ABSTRACT

Due to an increasing demand from decision makers for proper economic evaluations of health care services, cost-effectiveness analyses are becoming increasingly frequent. The statistic of interest in cost-effectiveness analysis is the incremental cost effectiveness ratio (ICER). When patient-specific data on costs and effects of alternative interventions is available, it can be used to quantify the uncertainty in the estimate of the ICER. Expressing this uncertainty by using confidence intervals has been recommended . However, because the statistic of interest is a ratio of two correlated random variables, its variance cannot be estimated exactly. Furthermore, the distribution of the ratio is unknown.

Recently, several approximate methods have been proposed for calculating confidence intervals for the incremental cost-effectiveness ratio. These include two parametric methods: one which relies on a Taylor's Series approximation of the variance, and one based on Fieller's theorem; as well as a number of methods which rely on bootstrapping methodology. In this manuscript, these methods were applied to data obtained from a randomized clinical trial in which both health resources consumed and health outcomes were observed. Furthermore, several variations of the bootstrapping methods were proposed and applied to the same data set. Probabilities of the true ICER being in given ranges were also estimated using a bootstrapping approach. Finally, issues

of sample size and power were briefly considered. The relative advantages and disadvantages of the different approaches were discussed.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. A. R. Willan, for his time and inspiring guidance, throughout the course of my project. I thank him for sharing his expertise and knowledge with me, allowing me to make to most of this opportunity to explore a new area of research.

I would also like to thank my examiners, Dr. P. Macdonald and Dr. H. Shannon, for their time and helpful remarks.

I express my appreciation to the members of the Department of Mathematics and Statistics at McMaster University, who made my years spent at McMaster both educational and very enjoyable. I sincerely thank the many professors at McMaster who shared their advice with me and encouraged me in so many ways.

I am grateful to my parents, my sisters, and my friends, for their encouragement and support throughout the course of my studies.

Finally, I would like to especially thank Steve Carino, for his constant encouragement, support, and belief in my ability to achieve anything I set out to do.

Financial support throughout the course of this project was provided by the Natural Sciences and Engineering Research Council of Canada.

Contents:

List of Figures:

List of Tables:

Chapter 1: Introduction

Randomized controlled trials, and subsequent statistical analyses, have been used for many years to compare the effectiveness of new therapies in the provision of healthcare services. Recently, economic evaluations of health care services (Laupacis *et al.*, 1992), and specifically cost-effectiveness analyses of new therapies have become increasingly more frequent (O'Brien *et al.*, 1994). Gold *et al.* (1996) define cost-effectiveness analysis (CEA) as:

"An analytic tool in which costs and effects of a program and at least one alternative are calculated and presented in a ratio of incremental cost to incremental effects. Effects are health outcomes, such as cases of a disease prevented, years of life gained, or quality-adjusted life years..."

Examining data on the costs as well as the effectiveness is important for policy recommendation and decision making regarding the adoption of new therapies (Adams *et al.*, 1992). Thus, there is an increasing demand from health care policy makers for properly analyzed data regarding the cost-effectiveness of new therapies.

Traditionally cost effectiveness analyses made use of sampled effectiveness data and non-sampled cost data. Such data were analyzed using sensitivity analysis to determine the findings' robustness (O'Brien *et al.*, 1994). Sensitivity analysis involves varying uncertain features and assumptions of the model, one at a time, over a range of possible values, to see if the basic conclusions change when a particular feature or

assumption is varied (Sacristán *et al.*, 1995). Although sensitivity analysis has been recommended for assessing problems of data uncertainty in economic appraisal of health care programmes, it has some major limitations (O'Brien *et al.*, 1994).

Recently, an increasing number of randomized controlled trials have included prospective collection of patient-level health care resource utilization. Whereas deterministic models based upon secondary analysis of retrospective data rely on sensitivity analysis for dealing with and presenting uncertainty, sampling error in the results from prospectively collected data can be handled by using conventional statistical methods (O'Brien *et al.*, 1994).

In randomized clinical trials designed for prospective cost-effectiveness evaluation, a bivariate random vector is observed on each patient from two therapy groups: one referred to as the standard arm (S) and the other as the treatment arm (T). These vectors contain a measure of the effectiveness as well as the total cost of treating the patient. The total cost is based on the sum of the total resources used times the "price" of each resource. Thus, for each patient an observation of the form (E_{ij}, C_{ij}) is obtained, where $i=S,T, j=1,...,n_i$. n_S and n_T are the numbers of patients in the standard and treatment arms, respectively.

The following notation will be used throughout the manuscript:

Population parameters:

μ_{CT} - true mean cost in treatment arm;

μ_{CS} - true mean cost in standard arm;

μ_{ET} - true mean effectiveness in treatment arm;

μ_{ES} - true mean effectiveness in standard arm.

σ^2_{CT}, σ^2_{CS}, σ^2_{ET}, σ^2_{ES} - true (population) between patient variances for the costs in treatment arm, costs in standard arm, effectiveness in treatment arm, and effectiveness in standard arm, respectively.

ρ_T, ρ_S - population correlation coefficients between the costs and effects of the treatment and standard groups, respectively (i.e., $\rho_i = Cor(E_{ij},C_{ij})$; $i = T,S$).

Sample data:

E_{ij} - effectiveness for patient j in group i (i=S or T), [note: $E(E_{ij}) = \mu_{Ei}$];

C_{ij} - total cost for patient j in group i, [note: $E(C_{ij}) = \mu_{Ci}$];

n_S - number of patients in standard arm;

n_T - number of patients in treatment arm.

Statistics calculated from the data:

$\overline{E}_S = \dfrac{\sum_{j=1}^{n_S} E_{Sj}}{n_S}$ - sample average effectiveness for standard arm, $E(\overline{E}_S) = \mu_{ES}$;

$\overline{E}_T = \dfrac{\sum_{j=1}^{n_T} E_{Tj}}{n_T}$ - sample average effectiveness for treatment arm, $E(\overline{E}_T) = \mu_{ET}$;

$\overline{C}_S = \dfrac{\sum_{j=1}^{n_S} C_{Sj}}{n_S}$ - average sample cost for standard arm, $E(\overline{C}_S) = \mu_{CS}$;

$\overline{C}_T = \dfrac{\sum_{j=1}^{n_T} C_{Tj}}{n_T}$ - average sample cost for treatment arm, $E(\overline{C}_T) = \mu_{CT}$.

S^2_{CT}, S^2_{CS}, S^2_{ET}, S^2_{ES} - sample variances for the costs in the treatment arm, costs in the standard arm, effects in the treatment arm, and effects in the standard arm, respectively (unbiased estimators for the corresponding population variances).

r_T, r_S - sample correlation coefficients for costs and effects of the treatment and standard arms, respectively.

To make inferences from prospective cost-effectiveness evaluations, several methods have been developed using traditional inferential statistical methods which allow for the calculation of point estimates of the cost-effectiveness measures, hypothesis testing, and calculation of confidence intervals.

Results from a cost-effectiveness analysis can be graphically displayed on the cost-effectiveness plane, proposed by Black (1990). In this plane (Figure 1) the vertical axis is the difference in mean costs for the two groups, and the horizontal axis is the difference in mean effectiveness. If we let $\Delta C = \mu_{CT} - \mu_{CS}$ and $\Delta E = \mu_{ET} - \mu_{ES}$, and consider the point (ΔE,ΔC) in the cost-effectiveness (c/e) plane, the four quadrants represent the following situations which may arise:

1) Quadrant 1: $\Delta C>0$, $\Delta E>0$, which means that although treatment is more effective, it leads to greater costs than standard;

2) Quadrant 2: $\Delta C>0$, $\Delta E<0$, which means that treatment is less effective and more costly, and is said to be dominated by standard;

3) Quadrant 3: $\Delta C<0$, $\Delta E<0$, which means that treatment is less effective, but also less costly;

4) Quadrant 4: $\Delta C<0$, $\Delta E>0$, which means that treatment is more effective and less costly, and is said to dominate standard.

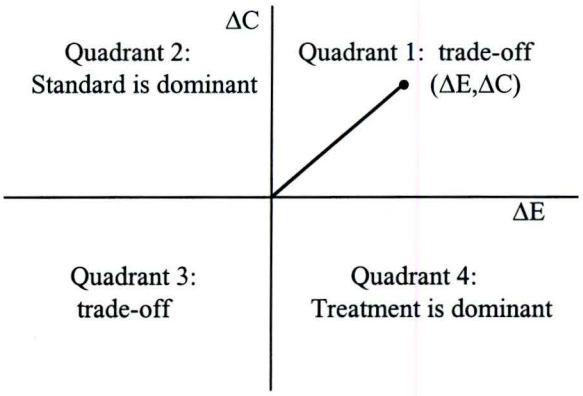


Figure 1: The cost-effectiveness plane (c/e plane)

Often a treatment is more effective but also more costly and the point (ΔE,ΔC) falls in quadrant 1. In such cases a useful measure of the cost-effectiveness of treatment relative to standard is the incremental cost-effectiveness ratio (ICER) proposed by Weinstein and Stason (1977), which is the ratio of the difference in costs to the difference in effectiveness:

$$R = \frac{\Delta C}{\Delta E} = \frac{\mu_{CT} - \mu_{CS}}{\mu_{ET} - \mu_{ES}}.$$

R is interpreted as the additional cost of achieving an extra unit of effectiveness from using treatment rather than standard. Using the collected data, this unknown population parameter (R) is estimated by:

$$\hat{R} = \frac{\overline{C}_T - \overline{C}_S}{\overline{E}_T - \overline{E}_S} = \frac{\Delta\overline{C}}{\Delta\overline{E}}.$$

On the c/e plane, the ICER point estimate is the slope of the line joining the point $(\overline{E}_T - \overline{E}_S, \overline{C}_T - \overline{C}_S)$ with the origin. The true ICER can fall in any one of the four quadrants, and the appropriate conclusion depends on which quadrant. Where the point falls within the quadrant is also important. For example, in quadrant 1, if a point, when joined with the origin, produces a large slope (i.e., large ICER) it means that although treatment is more effective, the added costs associated with it may be too high. If the slope is low (small ICER) the additional costs of treatment may be acceptable, considering its greater effectiveness. Of course what is considered too costly is a matter of judgment, and ultimately depends on the society's willingness to pay for added health benefits. Laupacis *et al.* (1992) suggested thresholds meant to distinguish whether a new therapy can be considered to be an attractive economic proposition or not, when effectiveness is measured in quality adjusted life years (QALYs). These thresholds are shown in Figure 2. Laupacis *et al.* (1992) used the thresholds to divide the c/e plane into regions that classify therapies into five "grades of recommendation". In quadrant 1 of Figure 2, the region between the line with slope 20,000 and the horizontal axis represents a situation where treatment would be considered a good investment, because the additional cost per QALY is low. The region between the lines with slopes 20,000 and 100,000 depicts a situation where the added cost per QALY is acceptable. In the region between the line with slope 100,000 and the vertical axis, the added cost per QALY

resulting from using treatment rather than standard is very high, and may be unacceptable for policy makers.

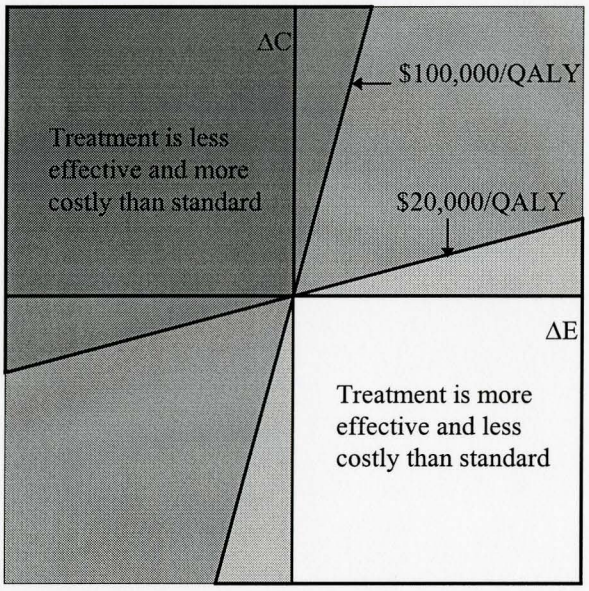


Figure 2: Thresholds which divide the c/e plane into "grades of recommendation" (Laupacis *et al.*, 1992)

The point estimate alone has limited value if the level of uncertainty is not quantified in some way. A common way of expressing the uncertainty of an estimate is by attaching a confidence interval (CI) to the point estimate. The use of confidence intervals for presenting clinical trial results is highly recommended since they not only provide information on statistical significance, but also provide information about the values for the parameter of interest that are consistent with the data (Guyatt *et al.*, 1995; O'Brien *et al.*, 1994). If sampled data is available for both the costs and the effectiveness, then estimates of the mean costs and effectiveness for both groups, as well as the variances of these estimates and correlations of cost and effectiveness in each group can be calculated. However, because in the estimator of the ICER the numerator

and the denominator are both stochastic, the variance of this estimator cannot be calculated exactly (Chaudhary and Stearns, 1996). Furthermore, the distribution of the ICER estimator is unknown.

Although the ratio estimator is biased, it is consistent, and thus the bias can be neglected for large sample sizes (Cochran, 1977; Chaudhary and Stearns, 1996). Chaudhary and Stearns (1996) found that in most cases, the distribution of $\hat{R}$ has been found to be positively skewed. However, the limiting distribution of $\hat{R}$ is normal as the sample sizes become very large, subject to some mild restrictions. They proposed that, as a rule, the large sample results can be used if the sample sizes are greater than 30 and the coefficients of variation of both the numerator and the denominator are less than 0.1 (Cochran, 1977).

Early attempts at estimating CIs included a method referred to as the box method (O'Brien *et al.*, 1994). According to this method, confidence intervals for the difference in costs, and difference in effects are calculated individually, and are used to estimate best and worst-case scenarios for the ICER as illustrated in Figure 3. The fundamental problems of this method were discussed by O'Brien *et al.* (1994). Particularly the assumption that costs and effects vary independently is an unjustified oversimplification. Furthermore, the bivariate probability density function of $(\Delta\overline{E}, \Delta\overline{C})$ is expected to have an elliptical shape as shown in Figure 4, rather than the box-shape depicted in Figure 3. The upper and lower confidence limits of the ICER are depicted in Figure 4 using rays from the origin that are tangent to the density function. The goal of the analysis

described in this manuscript is to find the slopes of these rays. Polsky *et al.* (1997) proposed improving the box-method by using narrower independent CIs for costs and effects, such that they jointly yielded a 95% CI for the ratio. Laska, Meisner, and Siegel (1997) used a similar procedure to calculate "Bonferroni CIs". A method proposed by van Hout *et al.* (1994) involved the calculation of a 95% probability ellipse for assessing the uncertainty around the cost effectiveness ratio.

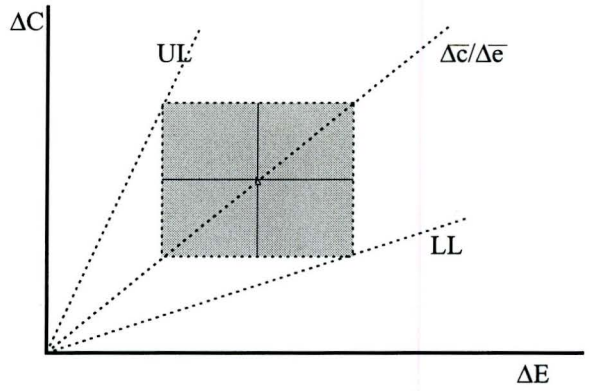


Figure 3: Box-method CI for the ICER (slope of UL line = upper confidence limit, slope of LL line = lower confidence limit)

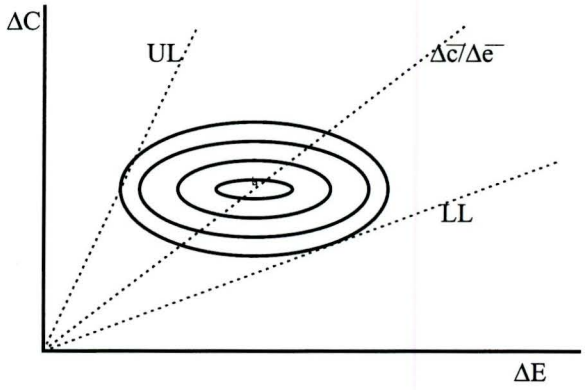


Figure 4: Contour view of the joint density function of difference in costs and difference in effectiveness (slope of LL line = lower confidence limit, slope of UL line = upper confidence limit)

Recently, several approximate yet improved methods have been proposed for obtaining confidence intervals for cost-effectiveness ratios. These include a method which uses a Taylor series approximation for calculating the variance of $\hat{R}$; a method based on Fieller's Theorem; and methods which use bootstrap methodology. All these methods are considered in this manuscript and are illustrated using data from a clinical trial which collected both resource use and health outcome data. Furthermore, several 'variations' of bootstrap intervals are proposed.

The data used to illustrate the different approaches comes from a clinical trial which compared chemotherapy with mitoxantrone plus prednisone for symptomatic hormone resistant prostate cancer to treatment with prednisone alone (Tannock *et al.*, 1996). An analysis of cost-effectiveness using the Fieller's Theorem method has been submitted for publication (Bloomfield *et al.*, 1998).

Finally, the issue of power and sample size in trial-based cost-effectiveness analysis is considered briefly. Methods which address these issues derived by Willan and O'Brien (1998) are applied to the same data set.

Chapter 2: Confidence Intervals for the ICER

2.1 Taylor's Series Approximation Method

O'Brien *et al.* (1994) proposed calculating a confidence interval for the ICER by using a Taylor's series approximation for estimating the variance. This method is based on the fact that it is possible to derive an approximation of the variance for any function of a random variable, or several random variables, by using Taylor's approximation. Casella and Berger (1990) described this method. In their derivation, they let $X_1, \ldots, X_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $X = (X_1, \ldots, X_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$, and suppose there is a differentiable function $g(X)$ (an estimator of some parameter) for which an approximate estimate of variance is required. The first-order Taylor series expansion of g about θ is

$$g(x) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(x_i - \theta_i) + \text{remainder}.$$

Thus, $$g(x) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(x_i - \theta_i).$$

Using the above approximation

$$\mathrm{E}_\theta[g(X)] \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)\mathrm{E}_\theta(x_i - \theta_i) = g(\theta)$$

and thus

$$Var_\theta g(X) \approx \mathrm{E}_\theta[(g(X) - g(\theta))^2]$$

$$\approx \mathrm{E}_\theta[(\sum_{i=1}^{k} g_i{}'(\theta)(X_i - \theta_i))^2]$$

$$= \sum_{i=1}^{k}[g_i{}'(\theta)]^2 Var_\theta X_i + 2\sum_{i<j} g_i{}'(\theta) g_j{}'(\theta) Cov_\theta(X_i, X_j).$$

For the purpose of finding an approximate variance of the ICER estimator, an approximation of the variance of the ratio of two random variables is required. Casella and Berger (1990) provide an example in which they suppose that X and Y are random variables with means μ_X and μ_Y respectively and the parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X / \mu_Y$ and show that

$$Var\left(\frac{X}{Y}\right) \approx \left(\frac{1}{\mu_Y}\right)^2 Var(X) + \left(\frac{\mu_X}{\mu_Y^2}\right)^2 Var(Y) - 2\left(\frac{1}{\mu_y}\right)\left(\frac{\mu_X}{\mu_Y^2}\right) Cov(X,Y)$$

$$= \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{Var(X)}{\mu_X^2} + \frac{Var(Y)}{\mu_Y^2} - \frac{2Cov(X,Y)}{\mu_X \mu_Y}\right) \qquad (2.1)$$

Thus, to find an approximate formula for $Var(\hat{R})$ we let $\overline{X} = \overline{C}_T - \overline{C}_S$, $\overline{Y} = \overline{E}_T - \overline{E}_S$, and then find $Var(\hat{R}) = Var(\overline{X} / \overline{Y})$ using the above formula.

In this case,

$Var(\overline{X}) = Var(\overline{C}_T) + Var(\overline{C}_S)$ since the costs for the two groups are independent,

$$= \frac{\sigma_{CT}^2}{n_T} + \frac{\sigma_{CS}^2}{n_S}.$$

The variances of C_{Ti} and C_{Sj} (σ_{CT}^2 and σ_{CS}^2, respectively) can be approximated by their unbiased estimators, the sample variances S_{CT}^2 and S_{CS}^2, giving:

$$Var(\overline{X}) \approx \frac{S_{CT}^2}{n_T} + \frac{S_{CS}^2}{n_S}.$$

Similarly,

$Var(\overline{Y}) = Var(\overline{E}_T) + Var(\overline{E}_S)$ can be approximated by

$$Var(\overline{Y}) \approx \frac{S_{ET}^2}{n_T} + \frac{S_{ES}^2}{n_S}.$$

Also,

$$Cov(\overline{X},\overline{Y}) = Cov(\overline{C}_T - \overline{C}_S, \overline{E}_T - \overline{E}_S)$$

$$= \frac{\rho_T \sigma_{CT} \sigma_{ET}}{n_T} + \frac{\rho_S \sigma_{CS} \sigma_{ES}}{n_S}$$ can be approximated by

$$Cov(\overline{X},\overline{Y}) \approx \frac{r_T S_{CT} S_{ET}}{n_T} + \frac{r_S S_{CS} S_{ES}}{n_S}.$$

Finally,

$\mu_{\overline{X}} = E[\overline{X}] = E[\overline{C}_T - \overline{C}_S] = \mu_{CT} - \mu_{CS}$, and

$\mu_{\overline{Y}} = E[\overline{Y}] = E[\overline{E}_T - \overline{E}_S] = \mu_{ET} - \mu_{ES}$.

Substituting the expressions for all the components into equation 2.1, the following approximation of $Var(\hat{R})$ is obtained:

$$Var(\hat{R}) \approx R^2 \left[\frac{\sigma_{CT}^2 / n_T + \sigma_{CS}^2 / n_S}{(\mu_{CT} - \mu_{CS})^2} + \frac{\sigma_{ET}^2 / n_T + \sigma_{ES}^2 / n_S}{(\mu_{ET} - \mu_{ES})^2} - \frac{2(\rho_T \sigma_{CT} \sigma_{ET} / n_T + \rho_S \sigma_{CS} \sigma_{ES} / n_S)}{(\mu_{CT} - \mu_{CS})(\mu_{ET} - \mu_{ES})} \right]$$

$$\approx \hat{R}^2\left[\frac{S_{CT}^2/n_T + S_{CS}^2/n_S}{(\overline{C}_T-\overline{C}_S)^2} + \frac{S_{ET}^2/n_T + S_{ES}^2/n_S}{(\overline{E}_T-\overline{E}_S)^2} - \frac{2(r_T S_{CT}S_{ET}/n_T + r_S S_{CS}S_{ES}/n_S)}{(\overline{C}_T-\overline{C}_S)(\overline{E}_T-\overline{E}_S)}\right].$$

In order to construct a confidence interval, a distributional assumption is generally made. O'Brien *et al.* (1994) proposed the following CI:

$$\hat{R} \pm z_{1-\alpha/2}\sqrt{Var(\hat{R})}$$

where $z_{1-\alpha/2}$ is the 100(1-α/2)th percentile point of the standard normal distribution. This CI is based on the assumption that the ICER point estimator $\hat{R}$ has a normal distribution, and is symmetric around the point estimate of R.

One disadvantage of this method is that it is based on the assumption that the ICER point estimator is normally distributed. Although as Obenchain (1997) pointed out the ICER is asymptotically normally distributed since it is a "ratio estimator" in the sense of Cochran (1977), previous studies have shown this to be a questionable assumption (Chaudhary and Stearns, 1996). Criticism was also raised because the variance calculation is only approximate (Obenchain *et al.*, 1997). Obenchain (1997) commented that confidence intervals based upon Taylor series approximations are too narrow (i.e., anti-conservative) relative to the corresponding intervals from Fieller's theorem. In their comparison of four methods for constructing confidence intervals for cost-effectiveness ratios, Polsky *et al.* (1997) found that the Taylor series method gave confidence intervals that asymmetrically underestimated the upper limit of the interval. O'Brien *et al.* (1994) noted that accuracy of Taylor's approximation depends on the random variables $\overline{C}_T - \overline{C}_S$ and $\overline{E}_T - \overline{E}_S$ having small coefficients of variation.

Although this method is not believed to give particularly accurate results, it is, nevertheless, a significant improvement from the 'box' method because it takes into account the correlations between the costs and effectiveness. Also, the relative simplicity of both the derivation, and the computations required with this method, make it quite appealing.

2.2 Fieller's Theorem Method

A specific application of Fieller's Theorem (Fieller, 1954) allows the derivation of an exact confidence set for a ratio of normal means. A summary of the theorem can be found in Casella and Berger (1990). Fieller considered a situation where a random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from a bivariate normal distribution with parameters $(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ is available, and a confidence set for $\theta = \mu_Y / \mu_X$ is to be derived. If for $i = 1,\ldots,n$ we let

$Z_{\theta i} = Y_i - \theta X_i$ and

$\overline{Z}_\theta = \overline{Y} - \theta\overline{X}$, then

$\overline{Z}_\theta \sim N(0,V_\theta)$, where

$$V_\theta = \frac{1}{n}(\sigma_Y^2 - 2\theta\rho\sigma_Y\sigma_X + \theta^2\sigma_X^2).$$

The variance of $\overline{Z}_\theta$, V_θ, can be estimated by

$$\hat{V}_\theta = \frac{1}{n}(S_Y^2 - 2\theta S_{YX} + \theta^2 S_X^2),$$

where S_Y^2 and S_X^2 are the sample variances and S_{YX} is the sample covariance.

It can also be shown that $E[\hat{V}_\theta] = V_\theta$, $\hat{V}_\theta$ is independent of $\overline{Z}_\theta$, and $(n-1)\hat{V}_\theta / V_\theta \sim \chi_{n-1}^2$.

Thus, $\overline{Z}_\theta / \sqrt{\hat{V}_\theta} \sim t_{n-1}$, and therefore, the set $\left\{\theta: \frac{\overline{Z}_\theta^2}{\hat{V}_\theta} \le t_{n-1,1-\alpha/2}^2\right\}$ defines a 1-α confidence set for θ, where $t_{n-1,1-\alpha/2}$ is the 100(1-α/2)th percentile point of the t-distribution with n-1

degrees of freedom. For large n, the standard normal distribution can be used as an approximation of the t-distribution, and thus, the t-value can be replaced by the z-value.

Sacristán *et al.* (1995) proposed the use of Fieller's theorem for estimating the CI of the ICER, without providing a detailed derivation. They stressed the form of the CI which does not take covariance between costs and effectiveness into account (although they also considered the case with covariances). Furthermore, they provided a formula for a 95% CI, rather than a general formula applicable to any desired confidence level. Sacristán *et al.* (1995) pointed out that when data consists of paired samples or when the efficacy is expressed as a percentage, different formulas would need to be used.

By assuming that the numerator and denominator of the ICER estimator follow a bivariate normal distribution, Willan and O'Brien (1996) derived a procedure for calculating confidence intervals for ICERs based on Fieller's Theorem, using matrix algebra in the calculations. They considered sampled cost and binary effectiveness data, however, their derivation can be generalized to continuous effectiveness data. In their derivation they defined:

$\boldsymbol{x} = \begin{bmatrix} \bar{e}_T - \bar{e}_S \\ \bar{c}_T - \bar{c}_S \end{bmatrix}$ and $\hat{S}$ as the estimated variance-covariance matrix of $\boldsymbol{x}$. Letting

$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \hat{S}^{-1/2}\boldsymbol{x}$, $\hat{U} = \frac{1}{|\boldsymbol{y}|}\begin{bmatrix} y_1 & y_2 \\ -y_2 & y_1 \end{bmatrix}$, and $\hat{T} = \hat{U}\hat{S}^{-1/2}$, the 100(1-$\alpha$)% confidence interval for R is given by the slopes of the vectors:

$$\hat{T}^{-1}\left[1 \quad \frac{-(z_{1-\alpha/2}/|y|)}{\left\{1-(z_{1-\alpha/2}/|y|)^2\right\}^{1/2}}\right]' \text{ and}$$

$$\hat{T}^{-1}\left[1 \quad \frac{(z_{1-\alpha/2}/|y|)}{\left\{1-(z_{1-\alpha/2}/|y|)^2\right\}^{1/2}}\right]'$$

where $z_{1-\alpha/2}$ is the upper $(100\alpha/2)\%$ cut off point for the standard normal distribution. The above result was obtained by determining a CI for the slope of $\hat{T}\mu$ and then transforming the result back to get a CI for the slope of μ. The details of the derivation can be found in Willan and O'Brien (1996). Willan and O'Brien cautioned that no solution exists for $|y| < z_{1-\alpha/2}$, and explained that either small sample sizes or small expected differences can lead to no solution. In such cases, the sample size is inadequate to provide even reasonably precise estimates of R.

A derivation presented by Chaudhary and Stearns (1996) produces a "closed-form" CI. This derivation follows directly from a simple generalization of Fieller's theorem as it appears above, noting that in this case if we let

$\overline{X} = \overline{E}_T - \overline{E}_S$,

$\overline{Y} = \overline{C}_T - \overline{C}_S$, and

$\overline{Z} = \overline{Y} - R\overline{X}$ then

$$\overline{Z} \sim N(0, Var(\overline{C}_T - \overline{C}_S) + R^2 Var(\overline{E}_T - \overline{E}_S) - 2RCov[(\overline{C}_T - \overline{C}_S),(\overline{E}_T - \overline{E}_S)])$$

and thus,

$$\frac{(\overline{C}_T - \overline{C}_S) - R(\overline{E}_T - \overline{E}_S)}{\sqrt{Var(\overline{C}_T - \overline{C}_S) + R^2 Var(\overline{E}_T - \overline{E}_S) - 2RCov[(\overline{C}_T - \overline{C}_S),(\overline{E}_T - \overline{E}_S)]}} \sim N(0,1).$$

In order to obtain the confidence interval for R, the above expression is equated to $z_{\alpha/2}$, the 100(1-α/2)th percentile point of the standard normal distribution, and following some algebraic manipulation, a quadratic equation in R is found. Solving this quadratic equation for R, the two limits of the confidence interval can be approximated by:

$$R \in \hat{R}\left[\frac{(1 - z_{\alpha/2}^2 c_{nd}) \pm z_{\alpha/2}\sqrt{\{(c_{nn} + c_{dd} - 2c_{nd}) - z_{\alpha/2}^2 (c_{nn}c_{dd} - c_{nd}^2)\}}}{1 - z_{\alpha/2}^2 c_{dd}}\right], \text{ where}$$

$$c_{nn} = \frac{s_{CT}^2 / n_T + s_{CS}^2 / n_S}{(\overline{C}_T - \overline{C}_S)^2},\ c_{dd} = \frac{s_{ET}^2 / n_T + s_{ES}^2 / n_S}{(\overline{E}_T - \overline{E}_S)^2}, \text{ and}$$

$$c_{nd} = \frac{r_T s_{CT} s_{ET} / n_T + r_S s_{CS} s_{ES} / n_S}{(\overline{C}_T - \overline{C}_S)(\overline{E}_T - \overline{E}_S)}.$$

These confidence limits are equivalent to those previously suggested by Willan and O'Brien (1996).

Laska, Meisner, and Siegel (1997) also showed a brief derivation of the Fieller's theorem interval, making use of the chi-square distribution with one degree of freedom ($\chi_{(1)}^2$). However, they noted that since the variances and covariances are unknown and are replaced by their estimators, the F-distribution with the appropriate degrees of freedom should be used instead of the $\chi_{(1)}^2$. They pointed out that Willan and O'Brien (1996) used the $\chi_{(1)}^2$ value (or equivalently the square of the standard normal value) and

obtained a narrower Fieller interval. However, with large sample sizes, the difference becomes negligible.

The Fieller's theorem interval is preferred to the one which uses a Taylor's series expansion to approximate the variance of the ICER estimate, because it takes account of the skewness of the distribution of $\hat{R}$. Obenchain (1997) noted that similarly to the Taylor series approximation method, the Fieller's theorem approach recognizes that the ICER is a "ratio estimator" in the sense of Cochran (1977), and thus is asymptotically normally distributed. However, the Fieller's approach treats the numerator and denominator as a pair of correlated normal variables, and can thus be applied to small sample situations where the distribution of the ICER is actually highly skewed.

Chaudhary and Stearns (1996) comment that the confidence limits are approximate, since the two roots of the quadratic equation are imaginary in some samples. However, they conclude that this is rare if the coefficients of variation of the numerator and denominator are less than 0.3 (Cochran, 1977). Also, they feel that sometimes it may be hard to justify the assumption of bivariate normality, particularly when samples are small. Laska, Meisner, and Siegel (1997) discussed circumstances that may lead to solutions of the form $(-\infty,a)$ and (b,∞) (i.e., 'not [a,b]' which they call an exclusionary interval) or intervals that consist of the whole real line $(-\infty,\infty)$.

Chaudhary and Stearns (1996) found that the Fieller's theorem intervals were similar to those obtained using bootstrap methods. They felt that these methods, which account for the skewness in the distribution of the ratio estimator, are substantially

preferable to methods that assume the estimator is normally distributed. Using Monte Carlo experiments to compare different methods of obtaining confidence intervals for cost-effectiveness ratios, Polsky *et al.* (1997) found that overall probabilities of coverage for the Fieller's theorem, as well as the non-parametric bootstrap method, were more accurate than those for the Taylor series method or the box method, and the confidence intervals resulting from the former two methods, were more dependably accurate. Obenchain (1997) commented that confidence intervals calculated using the Fieller's theorem method are "bow tie" shaped regions on the cost-effectiveness plane, and consequently, these intervals are "too-narrow" when $(\overline{E}_T - \overline{E}_S, \overline{C}_T - \overline{C}_S)$ is not significantly different from (0,0). Obenchain (1997) also raised the issue that the Fieller method of forming ICER confidence intervals is not "rescaling commutative", meaning that rescaling an ICER statistic by a multiplicative factor changes its upper and lower confidence limits by a different factor. This kind of scale change can occur, for example, in converting a numerator cost difference from one currency into another.

2.3 Bootstrapping Approaches:

So far, two parametric approaches to finding a confidence interval for the ICER have been discussed. Both of these required what can be considered "questionable" assumptions. Briggs, Wonderling, and Mooney (1997) expressed concern that given the unknown nature of the ICER's sampling distribution, there is reason to be cautious of such parametric methods. An alternative to these approaches is to use non-parametric bootstrapping methods, developed in the late 1970's (Efron and Tibshirani, 1993). Such methods have been applied to cost-effectiveness analyses by many authors including Obenchain (1997), Chaudhary and Stearns (1996), Briggs, Wonderling and Mooney (1997) and Polsky *et al.* (1997). The bootstrapping approach can be applied to the problem of estimating the standard error, bias, or confidence limits for the ICER (Briggs, Wonderling, and Mooney, 1997).

The bootstrap method estimates the sampling distribution of a statistic through a large number of simulations, based on sampling with replacement from the original data (Briggs, Wonderling, and Mooney, 1997). The reasoning behind the method (in the case of a single random sample) is that the observed random sample is treated as an empirical estimate of the true probability distribution of the population by weighting each observation in the random sample by the probability $1/n$, where n is the sample size. Successive random samples of size n are then drawn from the original random sample with replacement, to provide the bootstrap re-samples. Let the number of bootstrap re-

samples taken be denoted by B. The statistic of interest is then calculated for each of these B re-samples, and these bootstrap replicates of the original statistic make up the empirical estimate of the statistic's sampling distribution, which can then be used to make inferences about the population parameter of interest (Briggs, Wonderling, and Mooney, 1997).

In the case of the ICER, because there are two variables per patient, bootstrapping requires sampling from a bivariate distribution. Furthermore, it will require sampling from two empirical distributions, corresponding to the two populations. Care must be taken to bootstrap each sample appropriately, since the ICER is estimated on the basis of four statistics from 2 samples (Briggs, Wonderling, and Mooney, 1997). Efron and Tibshirani (1993) advocate that the re-sampling mechanism mirrors that by which the original data were obtained. In the case of the ICER, where data on resource utilization and health outcome exist for treatment and standard with sample sizes n_T and n_S, respectively, the following algorithm (Chaudhary and Stearns, 1996) should be followed:

1) randomly sample n_T data pairs (C_{Ti},E_{Ti}) with replacement from the n_T treatment patients, and calculate $\overline{C}_T^*$ and $\overline{E}_T^*$, the bootstrap replicates of $\overline{C}_T$ and $\overline{E}_T$

2) randomly sample n_S data pairs (C_{Si},E_{Si}) with replacement from the n_S standard patients, and calculate $\overline{C}_S^*$ and $\overline{E}_S^*$, the bootstrap estimates of $\overline{C}_S$ and $\overline{E}_S$

3) calculate $\hat{R}_b^* = \dfrac{\Delta C^*}{\Delta E^*} = \dfrac{\overline{C}_T^* - \overline{C}_S^*}{\overline{E}_T^* - \overline{E}_S^*}$

4) repeat steps 1-3 B times, obtaining B independent bootstrap estimates of $\hat{R}^*$ ($\hat{R}_1^*,...,\hat{R}_B^*$) along with the corresponding B ($\Delta E^*, \Delta C^*$) points.

There are several ways of using the B estimates to construct confidence intervals:

a) normal-assumption method:

The sample variance of the B bootstrap estimates of R can be calculated as:

$$\tilde{v}(\hat{R}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{R}_b^* - \tilde{R})^2 \text{ , where } \tilde{R} = \frac{1}{B}\sum_{b=1}^{B}\hat{R}_b^*$$

and this result can be used to construct a symmetric confidence interval, assuming that $\hat{R}$ is normally distributed:

$\hat{R} \pm z_{\alpha/2}\sqrt{\tilde{v}(\hat{R})}$ (Chaudhary and Stearns, 1996)

b) percentile method:

The 100(α/2) and 100(1-α/2) percentiles of the empirical distribution of $\hat{R}$ can be used as the limits of the central confidence interval (Chaudhary and Stearns, 1996). This is accomplished by ordering the $\hat{R}_b^*$'s in ascending order, and finding the appropriate percentiles. Mathematically, the cumulative distribution of $\hat{R}_b^*$ is:

$\hat{P}(t) = \#\{\hat{R}_b^* < t\} / B$

where $\#\{\hat{R}_b^* < t\}$ denotes the number of bootstrap estimates $\hat{R}_b^*$ less than a given value t.

The percentile bootstrap confidence interval (which will be referred to as the central percentile interval) is:

$$\left\{\hat{P}^{-1}(\alpha/2), \hat{P}^{-1}(1-\alpha/2)\right\}$$

Chaudhary and Stearns (1996) discussed the fact that $\tilde{R}$ is a biased estimator of $\hat{R}$, and will be too high if the distribution of $\hat{R}$ is positively skewed. Thus, the percentile method will result in incorrect CIs. Briggs, Wonderling, and Mooney (1997) explain that $\hat{R}$ is a biased estimator of the population ICER, R, and that this bias is magnified during the bootstrap process. Thus, the expectation of the bootstrap sampling distribution is more biased an estimator of R than $\hat{R}$. They describe how the bootstrap process can be used to estimate the magnitude of the bias in $\hat{R}$ and to adjust for it when appropriate. Briggs, Wonderling, and Mooney (1997) caution against some types of bias-correction, and discuss circumstances when it is unnecessary. They point out that some types of bias correction need to be approached with caution, since using the bootstrap estimate of bias to adjust the sample ICER may inflate the mean squared error of the estimator.

c) bias-corrected percentile method:

A third bootstrap method, the bias-corrected percentile method suggested by Chaudhary and Stearns (1996), adjusts for the bias in $\tilde{R}$ as an estimate of $\hat{R}$. They point out that this method is therefore more appropriate when $\hat{R}$ is not distributed normally. The adjusted interval is given by:

$$\left[\hat{P}^{-1}\{\Phi(2\hat{z}_0 - z_{\alpha/2})\}, \hat{P}^{-1}\{\Phi(2\hat{z}_0 + z_{\alpha/2})\}\right]$$

where $\hat{z}_0 = \Phi^{-1}\{\hat{P}(\hat{R})\}$, and $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution, and $\hat{P}$ is defined as previously.

d) bias-corrected and accelerated method (BCa method)

Briggs, Wonderling, and Mooney (1997) propose using the bias-corrected and accelerated percentile method, first developed by Efron (1987), which they feel adjusts more satisfactorily for bias and skewness of the sampling distribution of $\hat{R}$. With this method, the adjusted percentiles are given by:

$$\alpha_1 = \Phi\left(\hat{z} + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right) \text{ and}$$

$$\alpha_2 = \Phi\left(\hat{z} + \frac{\hat{z}_0 + z_{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z_{(1-\alpha/2)})}\right).$$

The acceleration constant, a, adjusts for the skewness of the sampling distribution. Briggs, Wonderling, and Mooney (1997) propose using the jackknife estimate of a suggested by Efron and Tibshirani (1993):

$$\hat{a} = \frac{\sum_{i=1}^{n}(\overline{R}^{**} - \hat{R}_i^{**})^3}{6\left[\sum_{i=1}^{n}(\overline{R}^{**} - \hat{R}_i^{**})^2\right]^{3/2}}$$

where $\hat{R}_i^{**}$ is the jackknife replicate of the ICER with the ith observation removed, for $i=1,\ldots,n$, $\overline{R}^{**} = \sum \hat{R}_i^{**} / n$, and $n = n_S + n_T$.

e) percentile-*t* method

Finally, Briggs, Wonderling, and Mooney (1997) propose using the percentile-*t* method, which like the BCa percentile method accounts for skewness in the estimated sampling distribution, but in a very different way. First, each bootstrap replicate of the ICER is transformed into a standardized variable t^*:

$$t^{*b} = \frac{R^{*b} - \hat{R}}{\hat{\sigma}^{*b}},$$

where $\hat{\sigma}^{*b}$ is calculated for each replication using another round of bootstrapping, greatly increasing the number of computations required. The 100(α/2) and 100(1-α/2) percentiles from the obtained distribution of t^*, can then be used to calculate the interval as:

$$(\hat{R} - t^*_{1-\alpha/2}\sqrt{\widetilde{v}(\hat{R})}, \hat{R} - t^*_{\alpha/2}\sqrt{\widetilde{v}(\hat{R})})$$

where $\widetilde{v}(\hat{R})$ is calculated from the original bootstrap re-samples, as in the normal-assumption method. Because of the added computational complexities, this method is not frequently used, and has not been used in this study.

As Obenchain (1997) pointed out, "the bootstrap approach yields a rather dramatic graphical display of the variability in the cost and effectiveness differences when an entire study is literally redone hundreds of times", by plotting the ($\Delta E^*, \Delta C^*$) points for the individual re-samples. (See Figure 5). This graphical display shows the ICER CI as a 'wedge-shaped' region on the c/e plane. Obenchain (1997) proposed

dividing the c/e plane into five regions as shown in Figure 6, where the different regions span the full range of possible outcomes of cost-effectiveness studies, in terms of how favourable treatment is relative to standard.

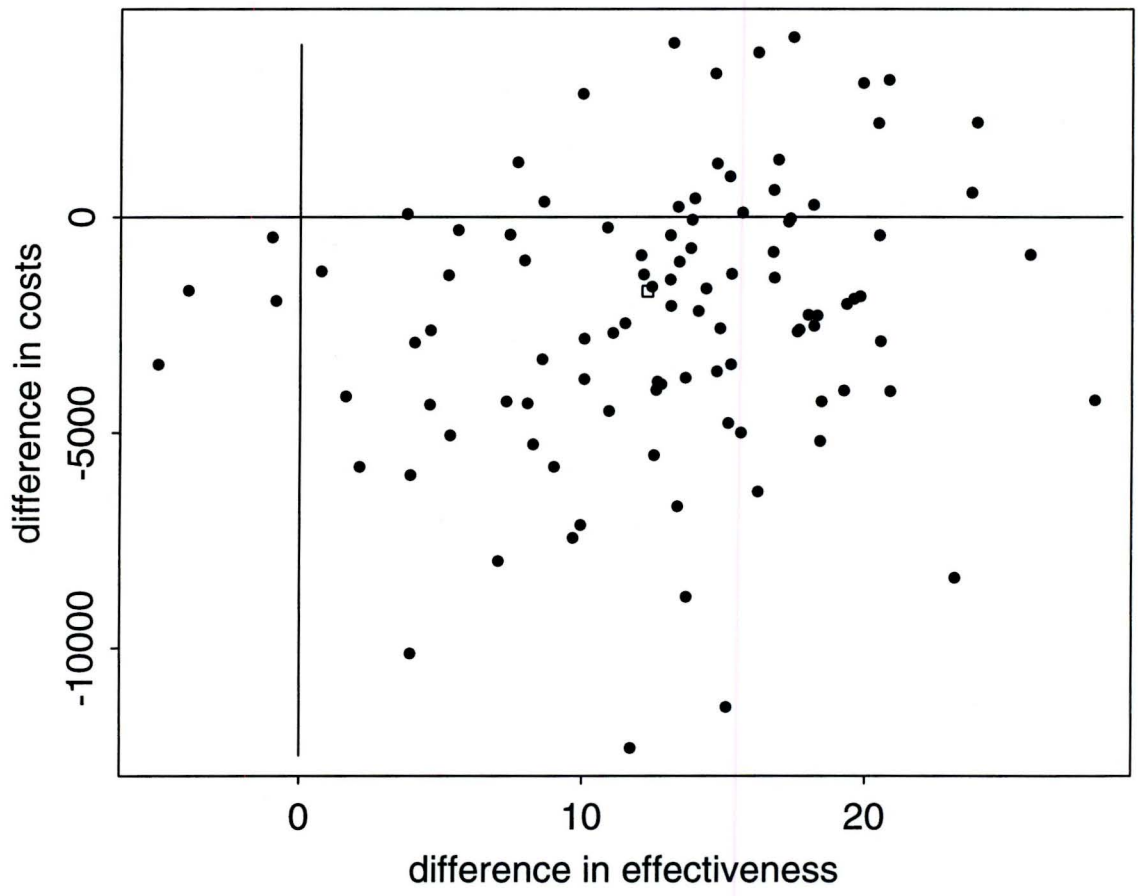


Figure 5: Graphical representation of the results of bootstrap re-samples (i.e. plot of $(\Delta E^*, \Delta C^*)$ pairs generated by bootstrap re-sampling) [Note that □ represents the point $(\Delta\overline{E}, \Delta\overline{C})$]

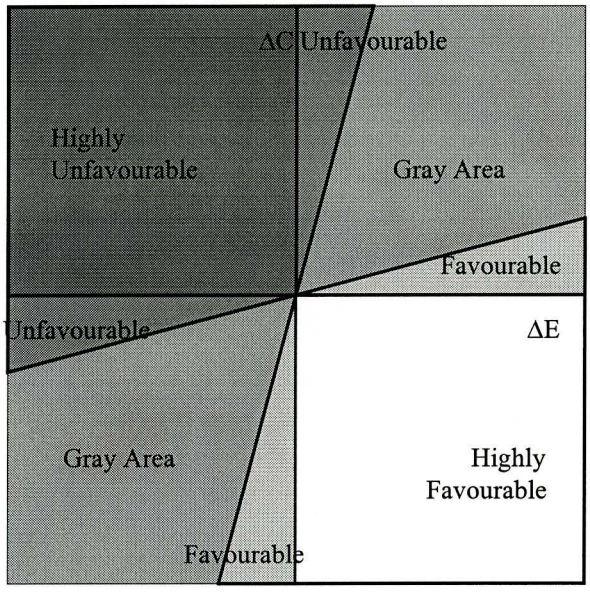


Figure 6: Division of the c/e plane proposed by Obenchain (1997)

Obenchain (1997) noted that we do not always find ourselves in the simple situation where all of the results generated in an ICER bootstrap analysis fall within quadrants 1 and 4, since some of the simulated effectiveness differences may turn out to be negative. However, points within the 2nd and 4th quadrant have negative slopes (negative ICERs), and points in the 3rd and 1st quadrant have positive slopes (ICERs). It has already been discussed (Chapter 1) that situations giving rise to points in the 2nd quadrant are very different from those giving rise to points in the 4th quadrant. Also, values in the 1st quadrant are interpreted differently from those in the 3rd quadrant, even if the slopes are exactly equal. This means that the ICER statistic is not, by itself, a sufficient statistic for making cost-effectiveness inferences. Obenchain (1997) proposed the use of ICER angles to deal with this problem.

The scales used along the horizontal and vertical axes of the c/e plane need to be standardized in order to define meaningful cost-effectiveness angles. This standardization can be accomplished in the following way (Obenchain, 1997):

$$x = \frac{(\overline{E}_T - \overline{E}_S)}{\sqrt{Var(E_{Ti}) + Var(E_{Sj})}} \text{ and } y = \frac{(\overline{C}_T - \overline{C}_S)}{\sqrt{Var(C_{Ti}) + Var(C_{Sj})}}$$

The ICER angle, θ, is then defined as the angle between the line segment joining the standardized point (x,y) with (0,0), and the minus-45-degree line. Contours of constant cost effectiveness can then be defined by pairs of line segments, joined at the origin, making equal angles +/- θ with the minus-45-degree line. Table 1 (reproduced from Obenchain, 1997) describes the possible situations that can arise, in terms of ICER angles, ICER slopes, and quadrants, linking these ideas together. It is important to note that the 60° and 120° lines were chosen arbitrarily. They are based on a similar idea that was used in defining the thresholds by Laupacis *et al.* (1992), however, the latter are more precisely defined and easier to justify.

Description	ICER Angle	ICER Slope	c/e Quadrant
Highly Favourable	$0° \leq \lvert\theta\rvert < 45°$	Negative	IV
Favourable	$45° \leq \lvert\theta\rvert < 60°$	Positive (extreme)	I or III
Mixed ("Gray Area")	$60° \leq \lvert\theta\rvert < 120°$	Positive (neither very large nor very small)	I or III
Unfavourable	$120° \leq \lvert\theta\rvert < 135°$	Positive (extreme)	I or III
Highly Unfavourable	$135° \leq \lvert\theta\rvert < 180°$	Negative	II

Table 1: Classification of situations arising from cost-effectiveness studies, reproduced from Obenchain (1997)

Using ICER angles, Obenchain (1997) defined two possible confidence intervals. In definition one of the ICER Angle Bootstrap CI, the bootstrap 100(1-α)% confidence region for cost-effectiveness is the wedge-shaped region subtending the smallest total angle at the origin and yet containing 100(1-α)% of the simulated cost-effectiveness pairs. According to a second definition, the bootstrap "central" 100(1-α)% confidence region for cost-effectiveness is the wedge-shaped region formed by excluding the top 100(α/2)% of simulated cost-effectiveness pairs with largest ICER angles and the bottom 100(α/2)% of simulated cost-effectiveness pairs with the smallest ICER angles. Although Obenchain (1997) proposes reporting CIs both in terms of angles and slopes, results reported in terms of angles would be rather difficult to interpret. It seems that, particularly in quadrant 1, it would be more beneficial to transform the angles obtained using the methods described by Obenchain (1997), into the corresponding slopes, and then compare the slopes to the thresholds defined by Laupacis *et al.* (1992) (or similar thresholds defined by policy makers), rather than comparing the angles to the arbitrarily chosen 60° and 120° angles suggested by Obenchain (1997).

Although it has not yet been done in the literature, similar corrections for bias could be used with the "angle methods" as they were with the "slope methods". It is proposed here, that bias correction and accelerated bias-correction should also be applied to calculation of ICER CIs, when the CIs are first found in terms of ICER angles which are subsequently converted to slopes (i.e., ICERs). The point estimate $\hat{R}$ needs to be expressed as an angle, and then the distribution of the angles obtained by bootstrapping

can be used to correct for the bias using the same protocol as with the ICERs (slopes). Noting that the idea of estimating the CI in terms of ICER slopes by the 'smallest wedge-shaped region' is based on the concept of a 'shortest interval', a similar method is proposed for dealing with the ICER (slope) estimates, when only the ICER (slope) estimates of the bootstrap re-samples are considered. Thus, it is proposed that the shortest CI containing 95% of the bootstrap ICER estimates can also be used to estimate the 95% CI, as an alternative to the "central" percentile method.

A commonly recognized advantage of the bootstrap methods is that possibly unrealistic assumptions about parametric forms for stochastic distributions need not be made, increasing the potential for accuracy and robustness (Obenchain, 1997). It has also been pointed out that with the bootstrap approach, it is of no consequence whether $\hat{R}$ has a "well-behaved" distribution (Willan and O'Brien, 1996). Another advantage of bootstrapping is that the ideas behind the method as well as the graphical display produced make this method easier to explain and appreciate than the previously discussed methods. A further advantage of the bootstrap approach discussed by Obenchain (1997), is that it allows estimation of probabilities, such as the probability that treatment is both more effective and less costly (i.e., probability that the true (ΔE, ΔC) point lies in the 4th quadrant), by finding the percentage of the bootstrap replications that produce points (ΔE^*, ΔC^*) in that region. This is a particularly useful method of analysis when the point estimate falls in the 4th quadrant, and there is a high probability that treatment is dominant, as compared with standard. Stinnett and Mullahy (1997) argued that in such

circumstances use of the ICER is misleading, and estimating the probability of dominance may be an appropriate alternative method of analysis. Laska, Meisner, and Siegel (1997) discussed statistical procedures for testing dominance, however, they relied mostly on simple parametric methods. Without making restrictive assumptions, the bootstrap approach of estimating probabilities may be preferable.

The historical limitation of the bootstrap, was the fact that it is computationally intensive. However, with the increased availability of fast computers, this is no longer an issue of great concern. The fact remains that a program (source code) has to be written, since currently there are no broadly available programs with commercial statistical software. Efron and Tibshirani (1993) describe some bootstrapping functions which are available from Splus function libraries, however, these may not be useful in the case of the ICER estimator. The only available program specific for the ICER is written in C and distributed by Eli Lilly and Company (Obenchain, 1995); however, this program is designed for analyzing binary effectiveness data only. Nonetheless, since the ideas and algorithms for bootstrapping presented in literature are uncomplicated, programming the method is not a difficult task.

A remaining issue is that all the observed data pairs for all patients in the two therapy groups are needed to construct bootstrap CIs, whereas the other methods described can be used when only summary statistics, such as means, variances, and covariances, are available. Furthermore, confidence limit values are sensitive to "parameters" such as B (the number of bootstrap replicates) and the random number seed

used by the computer for generating random numbers during the simulations. Obenchain (1997) suggests that full disclosure of such details is important, to allow verification of the results. Also, "sensitivity analyses" should be performed to ensure that results are not reported with inappropriate precision.

Briggs, Wonderling, and Mooney (1997) found that successive bootstrap estimates of bias and standard error suggest that these may be unstable, and thus they recommend a cautious interpretation of such estimates. They also found that the percentile and BCa methods gave fairly stable confidence limits after about 1000 replications. They pointed out that the BCa interval does not suffer from the problem of instability because it is based on the median bias, not the mean bias.

As all the other methods, the bootstrapping approach also has its limitations. Mainly, its validity rests on two asymptotics, discussed by Briggs, Wonderling, and Mooney (1997). The sample distribution tends to the population distribution as the original sample size approaches the population size, and given this, the bootstrap estimate of the sampling distribution of a statistic approaches the true sampling distribution of the parameter as the number of bootstrap replications, B, approaches infinity. Briggs, Wonderling, and Mooney (1997) also observe that bootstrap results should be interpreted cautiously pending further theoretical examination of the properties of the ICER statistic.

Finally, the use of ICER angles needs to be emphasized. In the literature, Obenchain (1997) and Obenchain *et al.* (1997) are the only ones who have employed this concept. Most of the studies which made use of the bootstrapping approach, although

they considered alternatives such as bias-correction, apparently did not make use of angles. This is not a problem if all the (ΔE^* , ΔC^*) points fall on one side of the vertical axis which is often the case. However, it is important that this be verified, and if necessary, that the angles are considered.

Chapter 3: Sample Size and Power

Although other authors have raised the issue of sample size for trial-based cost-effectiveness studies, O'Brien *et al.* (1994) first addressed this issue in the context of constructing statistical tests of economic hypotheses with predetermined level of significance and power. They considered testing the hypotheses:

H_O: $R=R_{max}$ vs H_A: $R<R_{max}$,

where R_{max} is a predefined "upper threshold" for the ICER, i.e., a maximum additional cost per unit effectiveness increase that the society (or policy maker) is willing to pay for. They proposed using the estimated $Var(\hat{R})$, calculated using the Taylor series approximation method, along with the normality assumption, to set up the test of the above stated hypothesis. If this test is performed, then the sample size required will depend not only on the magnitude of R_{max} and the amount of variation in $\hat{R}$, but also on the acceptable risk of Type I and Type II errors, which should be chosen prior to the study. Thus, they considered how prior expectation of variation in costs and effects could be used to determine the sample size required. Willan and O'Brien (1996) noted that with this approach it is required that policy makers, or those designing the trial, must establish threshold values, such as R_{max}, prior to conducting the study.

O'Brien *et al.* (1994) proposed that sample size calculations for stochastic economic studies should focus less on the required magnitude of a cost-effectiveness

outcome (i.e., hypothesis testing) and more on the degree to which increased sample size will improve the precision of the estimate by considering the width of the confidence interval. Sacristán *et al.* (1995) pointed out that CIs have an important practical consequence in calculating the sample size needed in cost-effectiveness studies. They proposed using Fieller's theorem in estimating the CIs. However, because of the complexity of the expression for the CI (as presented in Chapter 2.2) they did not consider sample size based on it, but rather assumed independence between costs and effects in their discussion of sample size. In their presentation of the use of Fieller's theorem in establishing CIs for the ICER, Willan and O'Brien (1996) suggested determining sufficiently large sample sizes to provide CIs of a predetermined width.

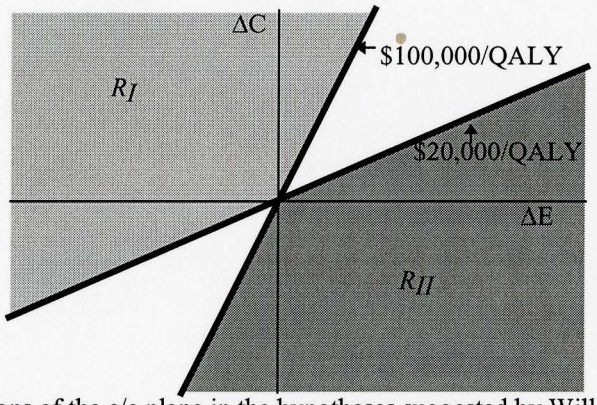


Figure 7: Regions of the c/e plane in the hypotheses suggested by Willan and O'Brien

Willan and O'Brien (1998) derived a formula for determining the sample size required to ensure that the confidence interval calculated using the Fieller's theorem method is narrow enough to "distinguish between two regions in the cost-effectiveness plane: one in which the new therapy is considered cost-effective (R_{II}) and one in which it

is not (R_I)". The two regions described, based on the threshold values suggested by Laupacis *et al.* (1992), are shown in Figure 7.

Defining $\mu = \begin{bmatrix} \mu_{ET} - \mu_{ES} \\ \mu_{CT} - \mu_{CS} \end{bmatrix}$, Willan and O'Brien considered testing H_1: $\mu \in R_I$ vs. A_1: $\mu \notin R_I$ and H_2: $\mu \in R_{II}$ vs. A_2: $\mu \notin R_{II}$, simultaneously, by plotting the 100(1-α)% CI. The authors noted that the probability of a Type I error (rejecting H_1 or H_2 when it is true) is limited by setting the level of the test (i.e., setting the confidence level for the interval), and also proposed setting the power by requiring that there be 1-α power of rejecting H_1 when H_2 is true and vice versa. This is achieved by ensuring that the angle spanned by the limits of the 100(1-α)% CI is no larger than the angle between the lines representing the upper and lower thresholds shown in Figure 7. Thus, the sample size desired is one sufficiently large to ensure that the CI is narrow enough so that it will lie either entirely outside R_I or entirely outside R_{II}. In the derivation, they assumed that the sample sizes are large enough so that $\boldsymbol{x} \sim N(\mu, \Sigma)$, where $\boldsymbol{x} = \begin{bmatrix} \overline{E}_T - \overline{E}_S \\ \overline{C}_T - \overline{C}_S \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_{ET} - \mu_{ES} \\ \mu_{CT} - \mu_{CS} \end{bmatrix}$, $\Sigma = n_S^{-1} \Sigma_S + n_T^{-1} \Sigma_T$; and, Σ_S and Σ_T are the variance-covariance matrices of $\begin{bmatrix} \overline{E}_S \\ \overline{C}_S \end{bmatrix}$ and $\begin{bmatrix} \overline{E}_T \\ \overline{C}_T \end{bmatrix}$ respectively. Assuming $n_S = n_T = n$, then $\Sigma = n^{-1} G$ where $G = \Sigma_S + \Sigma_T$.

The full derivation, which can be found in Willan and O'Brien (1998), leads to:

$$n = \frac{\left\{\left[\left\{\frac{z_{1-\alpha/2}}{\sin(\cos^{-1}[\boldsymbol{m}_1'G^{-1}\boldsymbol{m}_2 / \{\boldsymbol{m}_1'G^{-1}\boldsymbol{m}_1\boldsymbol{m}_2'G^{-1}\boldsymbol{m}_2\}^{1/2}]/2)}\right\}^2 - X_1^2\right]^{1/2} - Z\right\}^2}{\mu' G^{-1}\mu}$$

where m_1 and m_2 are the lower and upper thresholds respectively, and $\boldsymbol{m}_i = \begin{pmatrix}1 & m_i\end{pmatrix}'$; Z is a standard normal random variable; and X_1^2 is a random variable from a central chi-square distribution with 1 degree of freedom. To be conservative, Willan and O'Brien propose choosing some γ>0.5 and using $z_{1-\gamma}$ for Z and $x_{1-\gamma}^2$ for X_1^2, where $x_{1-\gamma}^2$ is the [100(1-γ)]th percentile of χ_1^2.

Because the sample size formula includes μ, Willan and O'Brien rearranged the formula to get:

$$\mu' G^{-1}\mu = \frac{\left\{\left[\left\{\frac{z_{1-\alpha/2}}{\sin(\cos^{-1}[\boldsymbol{m}_1'G^{-1}\boldsymbol{m}_2 / \{\boldsymbol{m}_1'G^{-1}\boldsymbol{m}_1\boldsymbol{m}_2'G^{-1}\boldsymbol{m}_2\}^{1/2}]/2)}\right\}^2 - X_1^2\right]^{1/2} - Z\right\}^2}{n}.$$

This equation defines, for a fixed value of n, a set of values μ, for which one would have adequate power (as described previously). These values are those that fall outside of the ellipse given by the equation, as shown in Figure 8.

Chaudhary and Stearns (1996) noted that sample sizes for randomized interventions are generally based upon a minimum detectable difference in the effectiveness measure, and that these samples may be considerably smaller than those required to obtain a precise estimate of the cost-effectiveness of the intervention. Furthermore, they pointed out that the modest sample sizes required to detect statistically

significant effects in a randomized trial may result in CIs for estimates of the ICER that are much wider than the boundaries obtained from deterministic sensitivity analyses. O'Brien *et al.* (1994), Sacristán *et al.* (1995), Chaudhary and Stearns (1996), and Willan and O'Brien (1996, 1998) mentioned the potential ethical issues arising from the fact that sample sizes required to answer cost-effectiveness questions will be considerably larger than those required for establishing effectiveness differences alone.

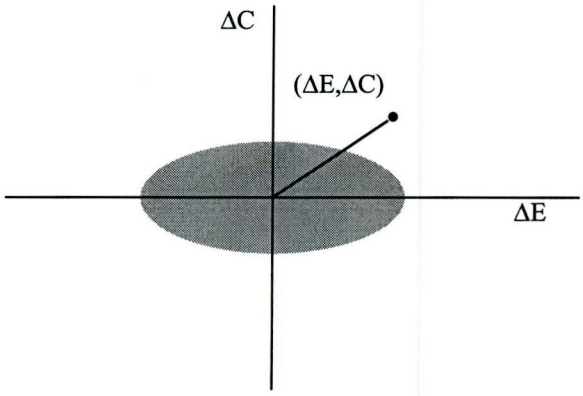


Figure 8: Ellipse which separates the c/e plane into regions with adequate and inadequate power

Chapter 4: Data and Methods

4.1 Data

The data was collected as part of a study to evaluate the benefits and the economic consequences of the use of chemotherapy with mitoxantrone plus prednisone in patients with symptomatic hormone-resistant prostate cancer (Tannock *et al.*, 1996; Bloomfield *et al.*, 1998). In the trial, 161 patients were randomized to initial treatment with mitoxantrone plus prednisone (M+P) or to prednisone alone (P). Although there was no significant difference in survival, the patients showed better palliation with M+P, with a clinically and statistically significant proportion of patients having relief of pain and improvement in health-related quality of life. Detailed retrospective chart review of resources used from randomization until death was undertaken for 114 out of the 161 patients. The 114 patients were enrolled at the three largest centers participating in the study: Calgary, Toronto, and Hamilton. 61 of these 114 patients were in the M+P ("treatment") group, and 53 were in the P ("standard") group. The cost-effectiveness analysis utilized the data on these 114 patients.

Health benefits of the two treatments were expressed in terms of quality adjusted life weeks (QALWs). Gold *et al.* (1996) defined quality adjusted life in the following way:

"A measure of health outcome which assigns to each period of time a weight, ranging from 0 to 1, corresponding to the health related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to a health state judged equivalent to death; these are then aggregated across time periods."

The core questionnaire of the European Organization for Research and Treatment of Cancer (EORTC), which included a disease specific module, was used in the trial. The utility was calculated from the rating scale value, which was the response to the global quality of life item in the questionnaire which asks patients "how would you rate your overall quality of life during the past week".

The resources collected included hospital admissions, outpatient visits, investigations, therapies (including all chemotherapy and radiation) and palliative care. Details of the computational methods used in arriving at the total costs for each patient can be found in Bloomfield *et al.* (1998).

The data analyzed here had been analyzed previously. These analyses included (i) a comparison of the effectiveness of the two treatments (Tannock *et al.*, 1996); (ii) an analysis of the cost-effectiveness, using the Fieller's theorem method (Bloomfield *et al.*, 1998); and (iii) an illustration of sample size calculation methods derived by Willan and O'Brien (1998).

4.2 Methods of Analysis

4.2.1 ICER Confidence Intervals

Summary statistics and plots of the data were obtained in Minitab and Splus. The point estimate of the incremental cost-effectiveness ratio was calculated as described in Chapter 1. Confidence intervals for the ICER were calculated using the following methods discussed in the earlier chapters:

i) Taylor's Series method;

ii) Fieller's Theorem method (both "closed form" formula and "matrix" method);

and the following bootstrapping methods:

iii) normal-assumption method using slopes;

iv) central percentile method using slopes;

v) bias-corrected percentile method using slopes;

vi) bias-corrected and accelerated method using slopes;

vii) narrowest CI method using slopes;

viii) normal-assumption method using angles;

ix) central percentile method using angles;

x) bias-corrected percentile method using angles;

xi) bias-corrected and accelerated percentile method using angles; and

xii) smallest wedge-shaped region method using angles;

The Taylor's series and Fieller's theorem CIs were calculated in SAS (the SAS programs are included in Appendix A). All of the bootstrapping and related calculations were performed in Splus. Twelve bootstraps were performed: three with B=100, three with B=500, three with B= 1000 and three with B=10,000. In each case all of the bootstrap approaches listed above were used to calculate a CI.

Validity of assumptions were evaluated by examining histograms and normal probability plots of the original data and some statistics (based on estimates from bootstrap re-samples). This part of the analysis was done using Minitab and Splus.

4.2.2 Probability of the ICER being in a "given region"

For each of the 12 bootstraps, the probability of the ICER being in the 4th quadrant or region R_{II} of Figure 7, was estimated by the proportion of the simulated $(\Delta E^*, \Delta C^*)$ points which fall within the given region.

4.2.3 Sample Size Calculations

The required calculations and the plot of the ellipse which separates the c/e plane into a region where all pairs $(\Delta E, \Delta C)$ will have enough power (as defined in the "sample size" section of the introduction) and a region of "inadequate power" were performed in Splus.

Chapter 5: Results

5.1 General Results

Table 2 provides some important summary statistics for the treatment (M+P) and standard (P alone) groups. Treatment is observed to have higher effectiveness, and lower (though not significantly) cost. In both groups there is a moderately low positive correlation, indicating a positive association between costs and effects (i.e., higher costs are associated with higher effectiveness, and lower costs with lower effectiveness). This weak relationship is also evident in the plots of costs versus effects seen in Figure 9.

Summary Statistic	Treatment (M+P)	Standard (P)
Number of patients (n_k)*	61	53
mean effectiveness ($\overline{E}_k$) (QALWs)	39.50	27.18
standard deviation of effectiveness (S_{Ek})	36.13	27.33
coefficient of variation of $\overline{E}_k$	0.12	0.14
95% CI for mean effectiveness†	(30.24,48.75)	(19.65,34.72)
mean total cost ($\overline{C}_k$)(CDN. $)	27,321.78	29,038.86
standard deviation of cost (S_{Ck})	19,860.70	20,426.75
coefficient of variation of $\overline{C}_k$	0.093	0.097
95% CI for mean costs†	(22234,32409)	(23407,34670)
sample correlation coefficient between costs and effects (r_k)	0.23	0.27
covariance between costs and effects	162,791.6	149,333.6

Table 2: Summary statistics (*k=T for the M+P group, k=S for the P group
†CIs calculated under assumption of normality of both populations)

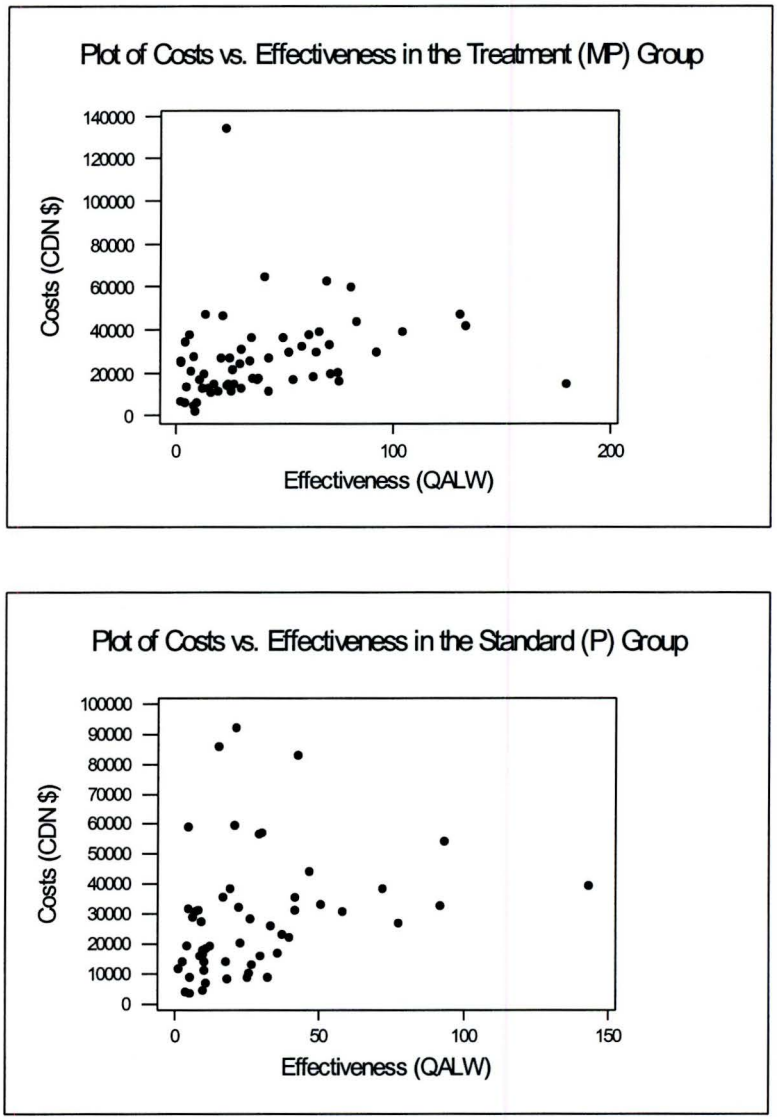


Figure 9: Plots of costs versus effectiveness for the two therapy groups

The histograms of costs and effects in the two groups seen in Figures 10 and 11, show that both costs and effects have right-skewed distributions. The skewness of treatment costs is to some extent due to a few individuals with particularly high costs.

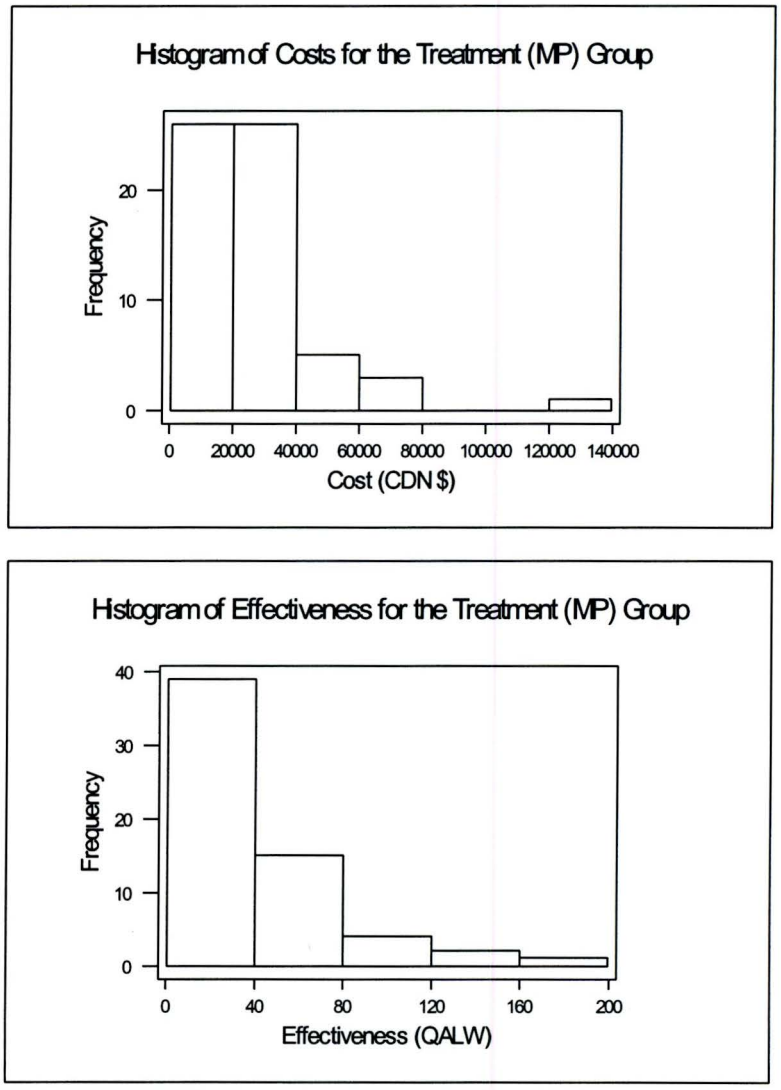


Figure 10: Frequency distributions of costs and effectiveness in the treatment group

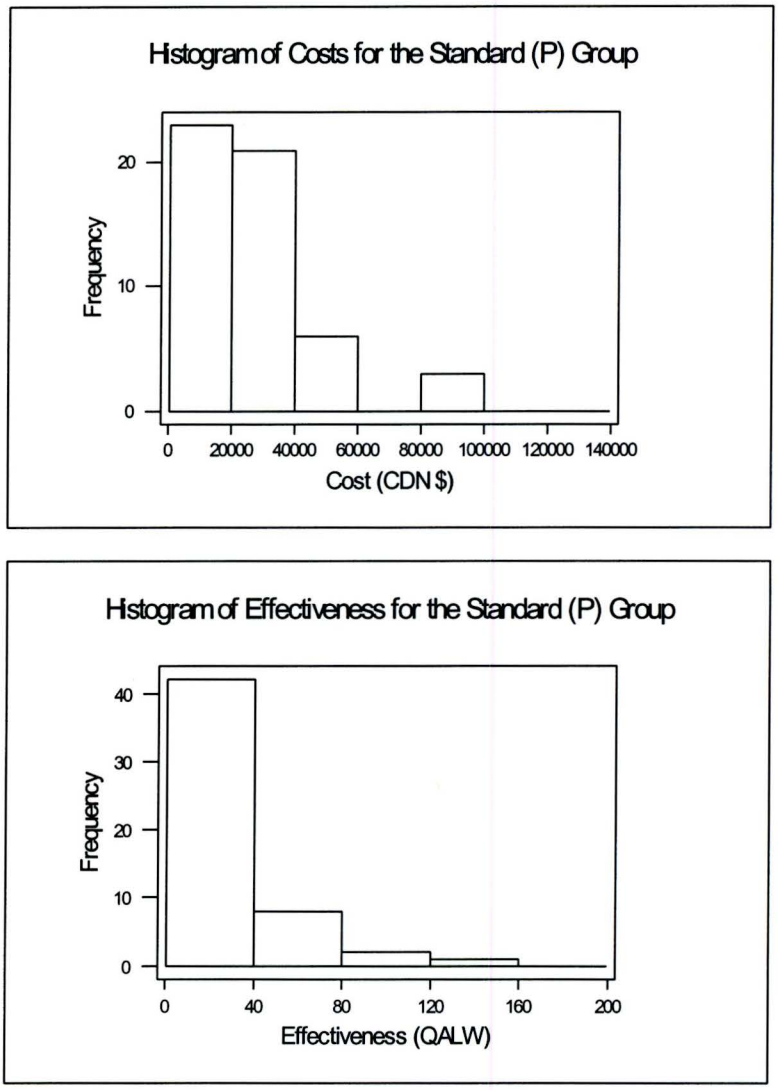


Figure 11: Frequency distributions of costs and effectiveness in the standard group

Table 3 summarizes some comparisons between the two groups, and provides estimates of the coefficients of variation for the numerator ($\overline{C}_T - \overline{C}_S$) and the denominator ($\overline{E}_T - \overline{E}_S$) of $\hat{R}$. The confidence intervals show that treatment appears to be significantly more effective than standard. The costs are not significantly different between the two groups. The sample coefficients of variation are high, particularly for the

difference in costs which is much greater than 0.5. The treatment arm has a higher average effectiveness and a lower average cost. Thus treatment is observed to be dominant and the point estimate of the ICER is negative. However, a possibility exists that the upper confidence limit may be positive. Since it is important to know how high the additional cost for a given improvement in effectiveness may be, the upper confidence limit is very meaningful to policy makers. There is also a potential for a positive lower limit, if there is a significant chance that treatment is less costly, but also less effective than standard. Although judging from the estimated CI for the difference in effects this may not be the case, a confidence interval of the ICER will answer questions regarding whether this appears to be the case or not.

Quantity	point estimate (95% CI)*	sample coefficient of variation
difference in mean effectiveness ($\overline{E}_T - \overline{E}_S$)	12.31 (0.3, 24.4)	0.48
difference in mean costs ($\overline{C}_T - \overline{C}_S$)	-1717.08 (-9207, 5773)	-2.21
ICER (CDN $/QALW)	-139.45	

Table 3: Comparisons between treatment and standard (* CIs calculated under "normality assumption" and assuming equal population variances in the 2 groups)

5.2 Approximate CIs for the ICER

Some of the methods used for constructing confidence intervals for the ICER rely on certain distributional assumptions. Histograms and normal probability plots of bootstrap estimates of the difference in costs, difference in effectiveness, the ICER

(slope), and ICER angles were produced to check the validity of these assumptions. Frequency distributions and their normal probability plots, based on bootstrap estimates of difference in mean effectiveness and difference in mean costs, show that the assumption of normality of the numerator and denominator of the ICER seems valid (see Figures 12 to 15). Histograms of the ICER bootstrap estimates, seen in Figure 16, and the corresponding coefficients of skewness, reveal that the distribution of $\hat{R}$ appears to be left-skewed. Furthermore, extreme (large or small) estimates of $\hat{R}$ do occur. Normal probability plots of these distributions (see Fig. 17) further support the notion that the estimator $\hat{R}$ does not appear to be normally distributed. Very extreme values of ICER angles are not possible. Frequency distributions and normal probability plots of ICER angles, seen in Figures 18 and 19, reveal that a normal probability model would seem more appropriate for the angles than it did for the slopes. However, the distribution of angles appears to be left skewed, and thus the assumption of normality may be invalid even for the angles.

The 95% confidence interval calculated using the Taylor's series approximation method is (-787.18, 508.29) CDN $ per QALW. Using the Fieller's theorem method, the 95% CI was found to be (-6096.65, 560.66). It is evident that the Taylor's method CI is narrower, and this difference in widths of the CIs is due to the fact that the lower limits of these two intervals are very different. The Fieller's interval is not symmetric about the ICER point estimate, with the lower limit being substantially further from the ICER point estimate than is the upper limit.

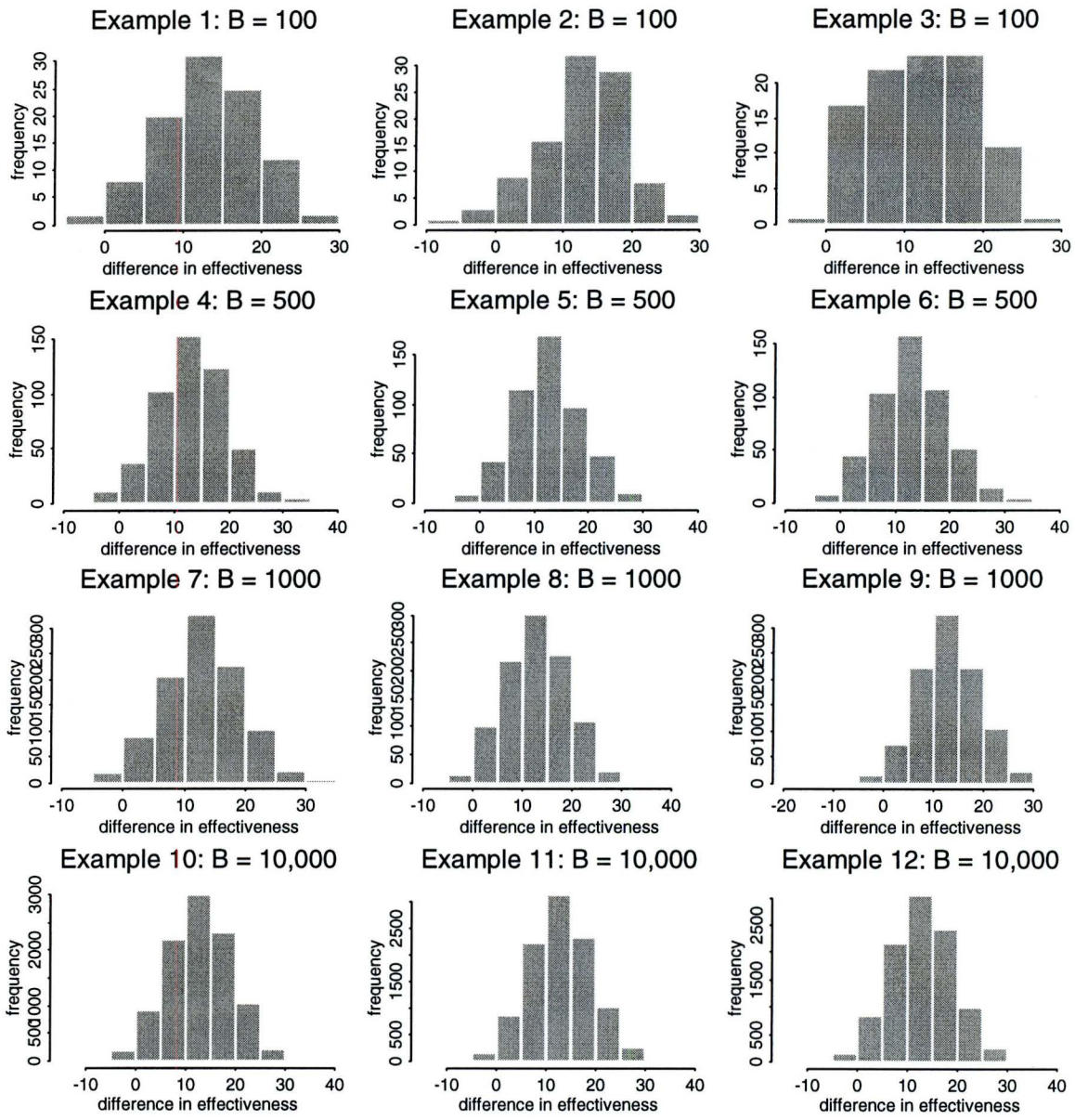


Figure 12: Histograms of B bootstrap estimates of the difference in effectiveness $(\overline{E}_T^* - \overline{E}_S^*)$ in QALWs. [The coefficients of skewness for the above 12 frequency distributions are: -0.16, -0.45, 0.05, 0.10, 0.13, 0.26, -0.03, 0.15, -0.006, 0.02, 0.05, and 0.006, respectively]

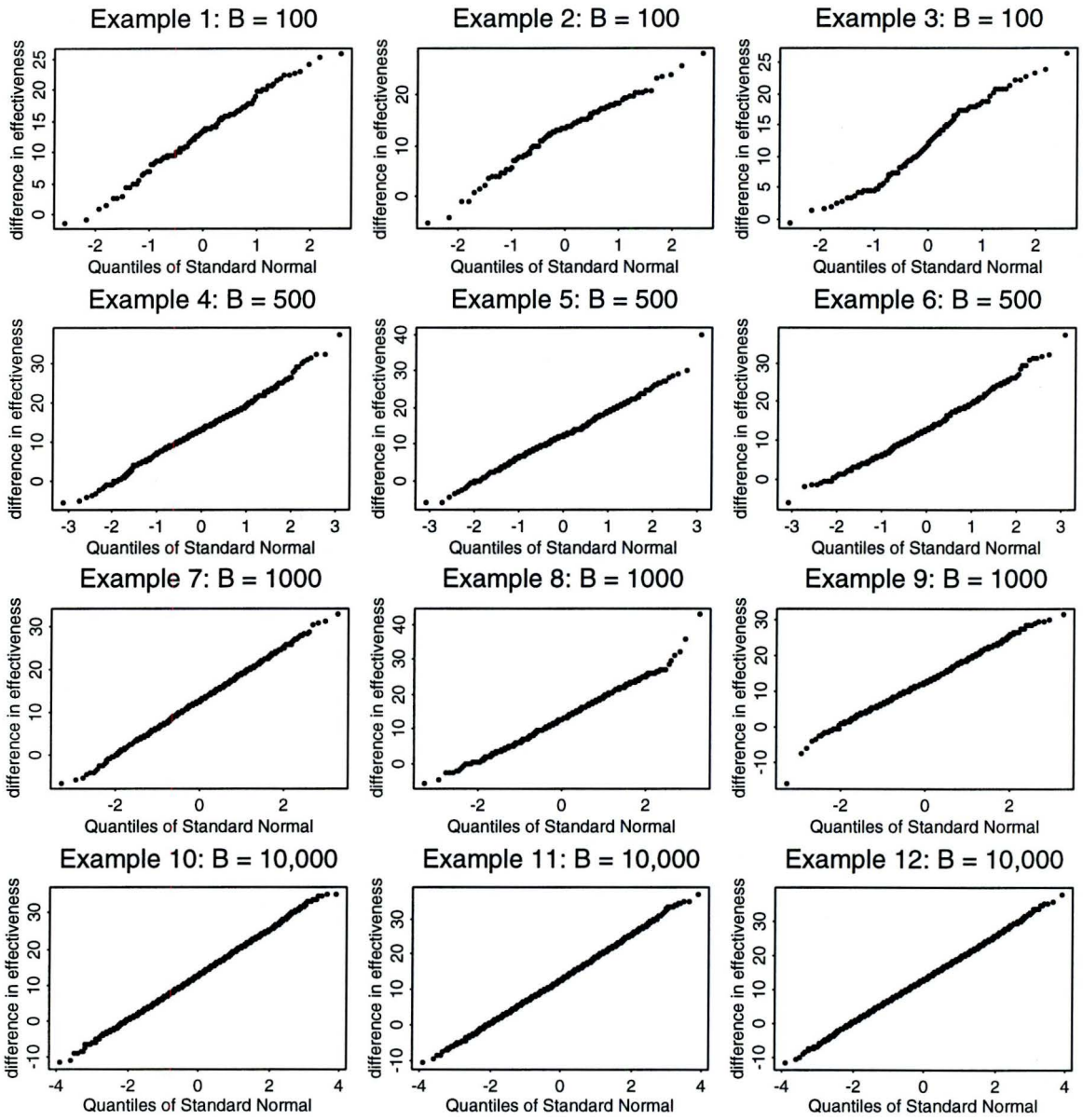


Figure 13: Normal probability plots for B bootstrap estimates of the difference in effectiveness $(\overline{E}_T^* - \overline{E}_S^*)$ measured in QALWs

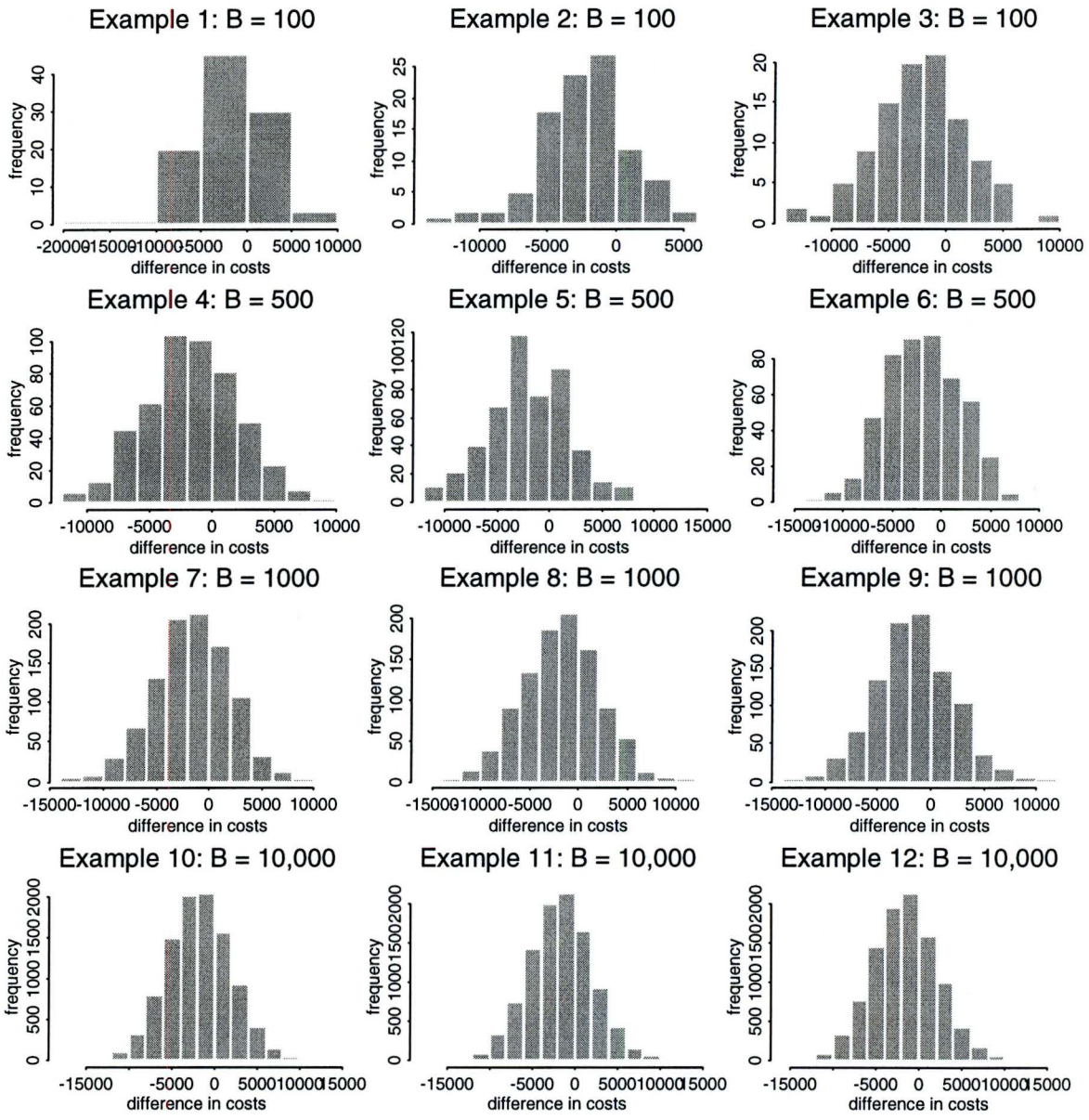


Figure 14: Histograms of B bootstrap estimates of the difference in costs $(\overline{C}_T^* - \overline{C}_S^*)$ in CDN $. [The coefficients of skewness for the above 12 frequency distributions are: -0.38, -0.48, -0.16, -0.008, 0.18, -0.04, -0.12, -0.003, 0.06, 0.06, 0.03, and 0.07, respectively]

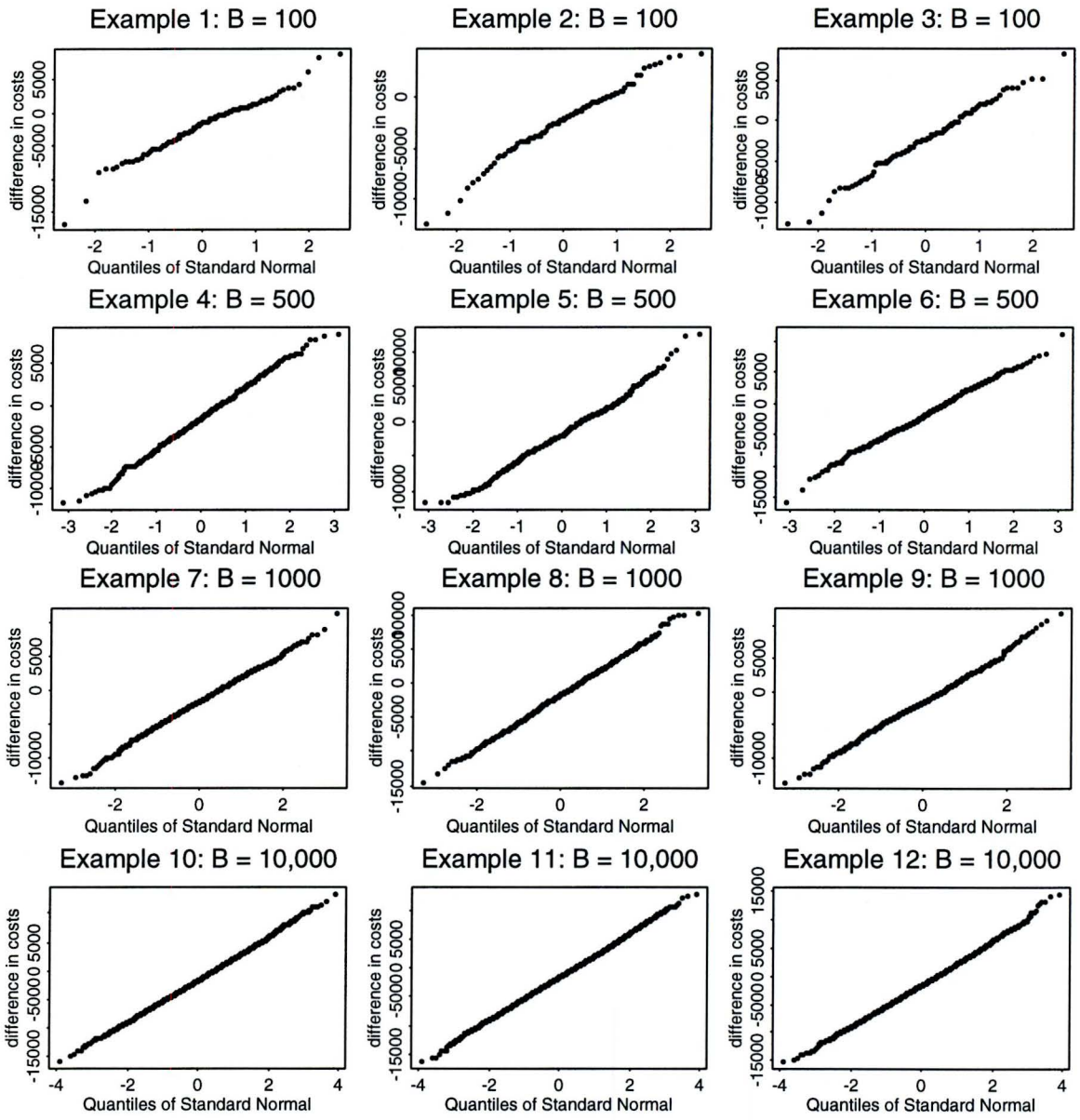


Figure 15: Normal probability plots for B bootstrap estimates of the difference in costs $(\overline{C}_T^* - \overline{C}_S^*)$ in CDN $

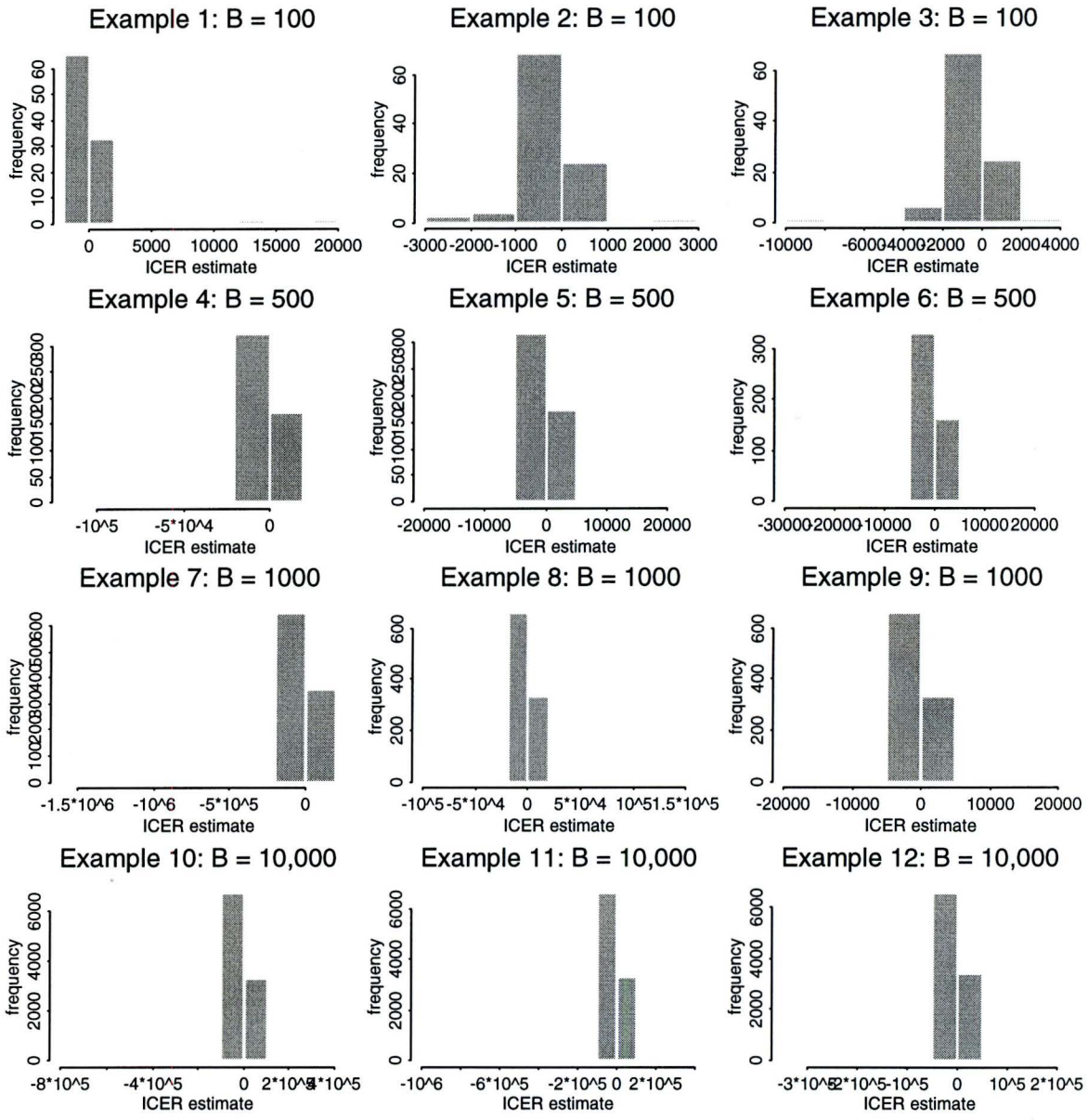


Figure 16: Histograms of B bootstrap estimates of the ICER (CDN $/QALW). [The coefficients of skewness for the above 12 frequency distributions are: 6.65, -1.30, -4.02, -14.59, 2.67, -0.92, -31.48, 14.18, 0.55, -50.85, -60.92, -30.07, respectively]

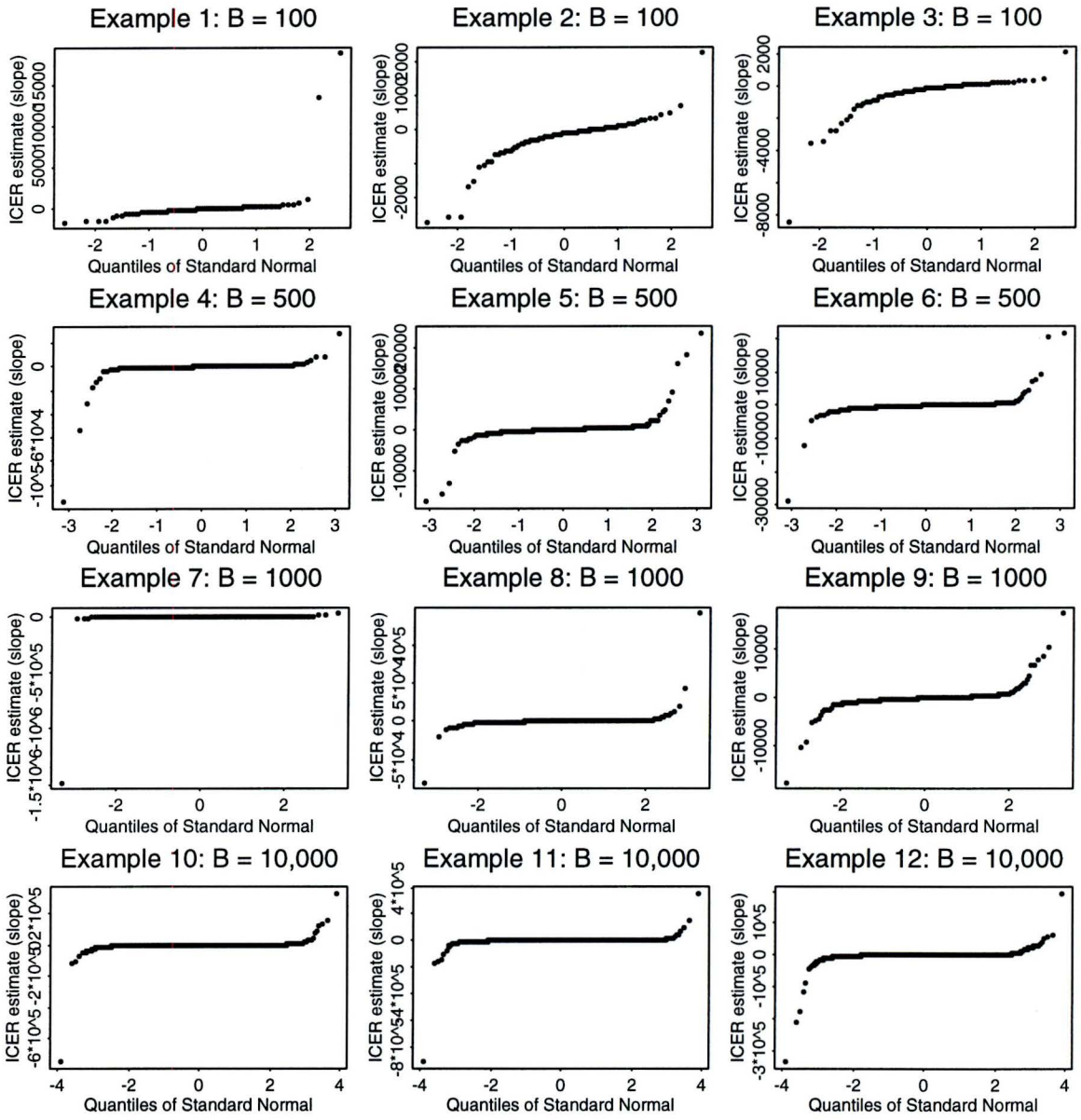


Figure 17: Normal probability plots for B bootstrap estimates of the ICER

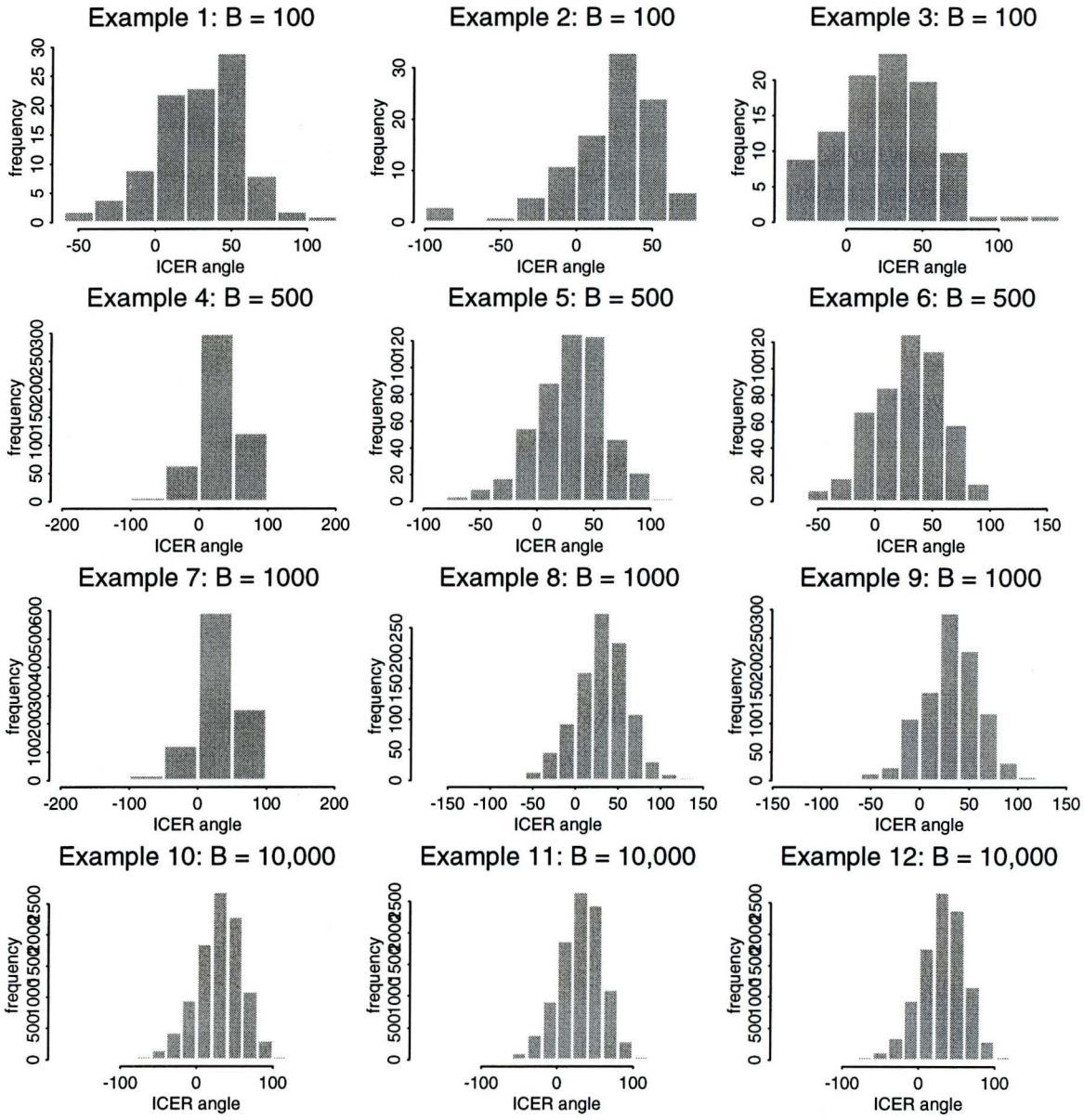


Figure 18: Histograms of B bootstrap estimates of ICER angles. [The coefficients of skewness for the above 12 frequency distributions are: -0.21, -1.49, 0.30, -0.62, -0.31, -0.04, -0.72, -0.42, -0.55, -0.66, -0.66, -0.61, respectively]

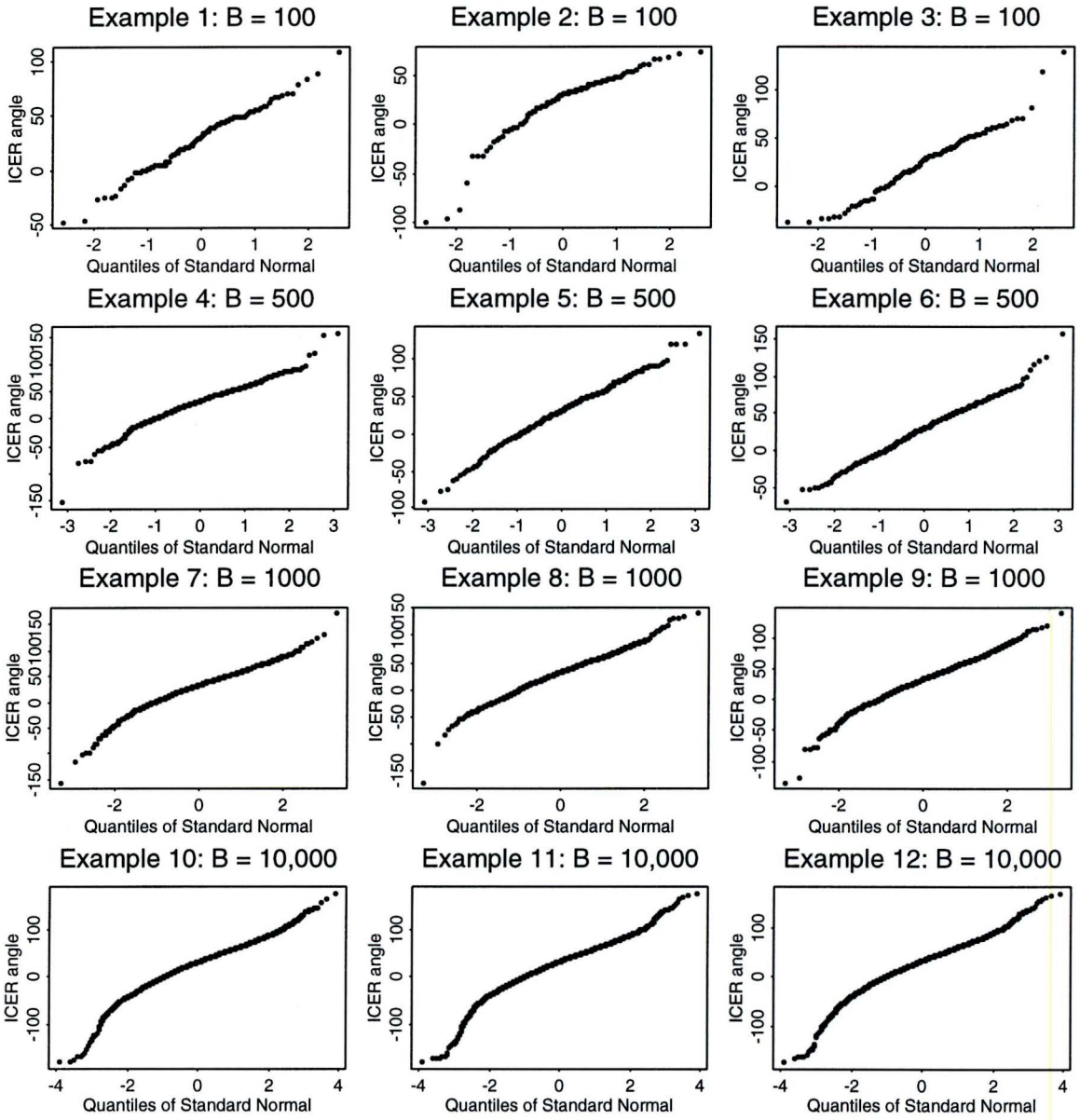


Figure 19: Normal probability plots for B bootstrap estimates of ICER angles.

Plots of the points $(\Delta E^*, \Delta C^*)$ generated by bootstrapping in the 12 examples are shown in Appendix C. The Bootstrap confidence intervals (and their widths) are shown in tables 4 and 5. Among the methods which relied on the slopes alone, results based on the percentile, bias-corrected, bias-corrected accelerated and the shortest confidence interval methods, were not drastically different from one another in most of the examples. The normal assumption method gave substantially different results from the remaining four methods, producing much wider confidence intervals. On average, the bias-corrected method provided intervals with both bounds slightly lower than those of the percentile method. The bias-corrected and accelerated intervals usually had bounds between those of the percentile and bias-corrected methods, but they were closer to those of the bias-corrected method.

Intervals based on ICER angles are quite a bit more variable, largely due to the fact that considering angles rather than slopes alone introduces the possibility of a positive lower confidence limit, if some of the bootstrap replications produce points in the 3rd quadrant. Again, the bias-corrected intervals were frequently lower than those produced by the central percentile method, and CIs produced by the bias-corrected accelerated method were often intermediate to those produced by the central percentile and bias-corrected methods. The 'smallest wedge shaped region' intervals were more stable than those obtained by the three methods discussed above, but there was still a considerable amount of variation among them. Intervals produced under the assumption of a normal distribution of ICER angles were most stable and narrowest.

Example #	B	Slope Methods				
		Normal Assumption	Percentile Method	Bias-corrected	Bias-corrected accelerated	Shortest CI
1	100	-4767, 4499	-1563, 835	-1580, 663	-1576, 711	-1723, 388
2	100	-1334, 1066	-2170, 448	-2170, 448	-2100, 452	-1711, 467
3	100	-2333, 2064	-3130, 291	-2617, 398	-2574, 406	-2766, 292
4	500	-12010, 11742	-1923, 783	-3056, 563	-2819, 567	-1613, 945
5	500	-4283, 4014	-1673, 1499	-1813, 988	-1684, 1061	-1684, 1027
6	500	-4332, 4064	-1724, 788	-1543, 1255	-1534, 1260	-1634, 810
7	1000	-92127, 91859	-1613, 788	-2021, 566	-1957, 573	-1482, 867
8	1000	-10929, 10660	-1903, 892	-2232, 601	-2194, 623	-1965, 735
9	1000	-2536, 2267	-1279, 791	-1534, 626	-1515, 634	-1279, 791
10	10,000	-17784, 17515	-2175, 763	-2390, 675	-2261, 694	-2002, 853
11	10,000	-20785, 20516	-2002, 643	-2361, 556	-2306, 569	-1581, 878
12	10,000	-10687, 10418	-1862, 757	-2374, 612	-2299, 630	-1731, 857

Table 4 a) Bootstrap CIs - Slope methods

Example #	B	Slope Methods				
		Normal Assumption	Percentile Method	Bias-corrected	Bias-corrected accelerated	Shortest CI
1	100	9266	2398	2243	2287	2111
2	100	2400	2618	2618	2552	2178
3	100	4397	3421	3015	2980	3058
4	500	23752	2706	3619	3386	2558
5	500	8297	3172	2801	2745	2711
6	500	8396	2512	2798	2794	2444
7	1000	183986	2401	2587	2530	2349
8	1000	21589	2795	2833	2817	2700
9	1000	4803	2070	2160	2149	2070
10	10,000	35299	2938	3065	2955	2855
11	10,000	41301	2645	2917	2875	2459
12	10,000	21105	2619	2986	2929	2588

Table 4 b) lengths of CIs - Slope methods

Example #	B	Angle Methods				
		Normal Assumption	Central	Bias-corrected central	Bias-corrected accelerated central	Smallest wedge-shaped region
1	100	-1559, 561	-1691, 426	-1710, 404	-1670, 449	-1723, 388
2	100	-2555, 736	1081, 238	1082, 238	-3414, 262	-2742, 286
3	100	-2444, 721	-2758, 371	-2292, 1085	-2385, 853	-2766, 292
4	500	-2616, 743	-74152, 526	6359, 463	23223, 507	-4114, 585
5	500	-2505, 729	-16805, 565	-52044, 554	-4474, 586	-2867, 597
6	500	-2124, 673	-3106, 479	-2351, 498	-1955, 531	-2167, 517
7	1000	-2380, 712	-10495, 526	5241, 450	-61992, 510	-2842, 567
8	1000	-2469, 724	-3550, 545	-6779, 517	-3568, 543	-2272, 639
9	1000	-1973, 647	-2804, 555	15667, 470	-4778, 533	-1638, 627
10	10,000	-2504, 729	-8017, 512	-12737, 492	-5366, 548	-6797, 521
11	10,000	-2309, 702	-3968, 507	-8246, 455	-4552, 492	-3898, 508
12	10,000	-2405, 715	-5175, 538	56685, 463	-7907, 517	-4239, 550

Table 5 a) Bootstrap CIs - Angle methods

Example #	B	Angle Methods				
		Normal Assumption	Central	Bias-corrected central	Bias-corrected accelerated central	Smallest wedge-shaped region
1	100	2120	2117	2114	2119	2111
2	100	3291	NA	NA	3676	3028
3	100	3165	3129	3377	3238	3058
4	500	3359	74678	NA	NA	4699
5	500	3234	17370	52598	5060	3464
6	500	2797	3585	2849	2486	2684
7	1000	3092	11021	NA	62502	3409
8	1000	3193	4095	7296	4111	2911
9	1000	2620	3359	NA	5311	2265
10	10,000	3233	8529	13229	5914	7318
11	10,000	3011	4475	8701	5044	4406
12	10,000	3120	5713	NA	8424	4789

Table 5 b) Lengths of bootstrap CIs - angle methods

5.3 Probabilities

The probability that the point (ΔE,ΔC) lies in the 4th quadrant, which is equal to the probability that the treatment dominates the standard, was estimated from the empirical distribution of $(\Delta\overline{E},\Delta\overline{C})$ by determining the percentage of bootstrap re-samples that resulted in a point in the 4th quadrant (as described in Chapter 2.3). These percentages are given for the 12 bootstrap simulation experiments in Table 6. Based on these results the probability seems to be approximately 0.66. The probability that the point (ΔE,ΔC) lies in a region where the treatment can be considered "highly favourable" or "favourable" (i.e., in region R_{II} of Figure 7) was approximated in a similar way. This probability is approximately equal to 0.95.

Example #	B	% Points in region R_{II} of Figure 7	% Points in the 4th quadrant
1	100	96	65
2	100	97	75
3	100	97	73
4	500	93.8	64.6
5	500	93.4	64.6
6	500	95.4	66.4
7	1000	94.2	64.7
8	1000	94.1	66.0
9	1000	94.2	66.3
10	10,000	94.67	66.80
11	10,000	94.79	66.23
12	10,000	94.30	65.68

Table 6: Estimates of probability that the ICER is in a given region

5.4 Sample Size Calculations

The ellipse defining the region for which $n = n_S = n_T = 200$ would be a sufficient sample size so that with a 95% confidence level, power would be adequate to draw conclusions about the cost-effectiveness of treatment, is shown in Figure 20a. The ellipse corresponding to a confidence level of .90 is shown in Figure 20b. The ellipses for sample sizes of $n_S = n_T = n = 57$ (which is close to the sample sizes obtained in this study) are shown in Figure 21.

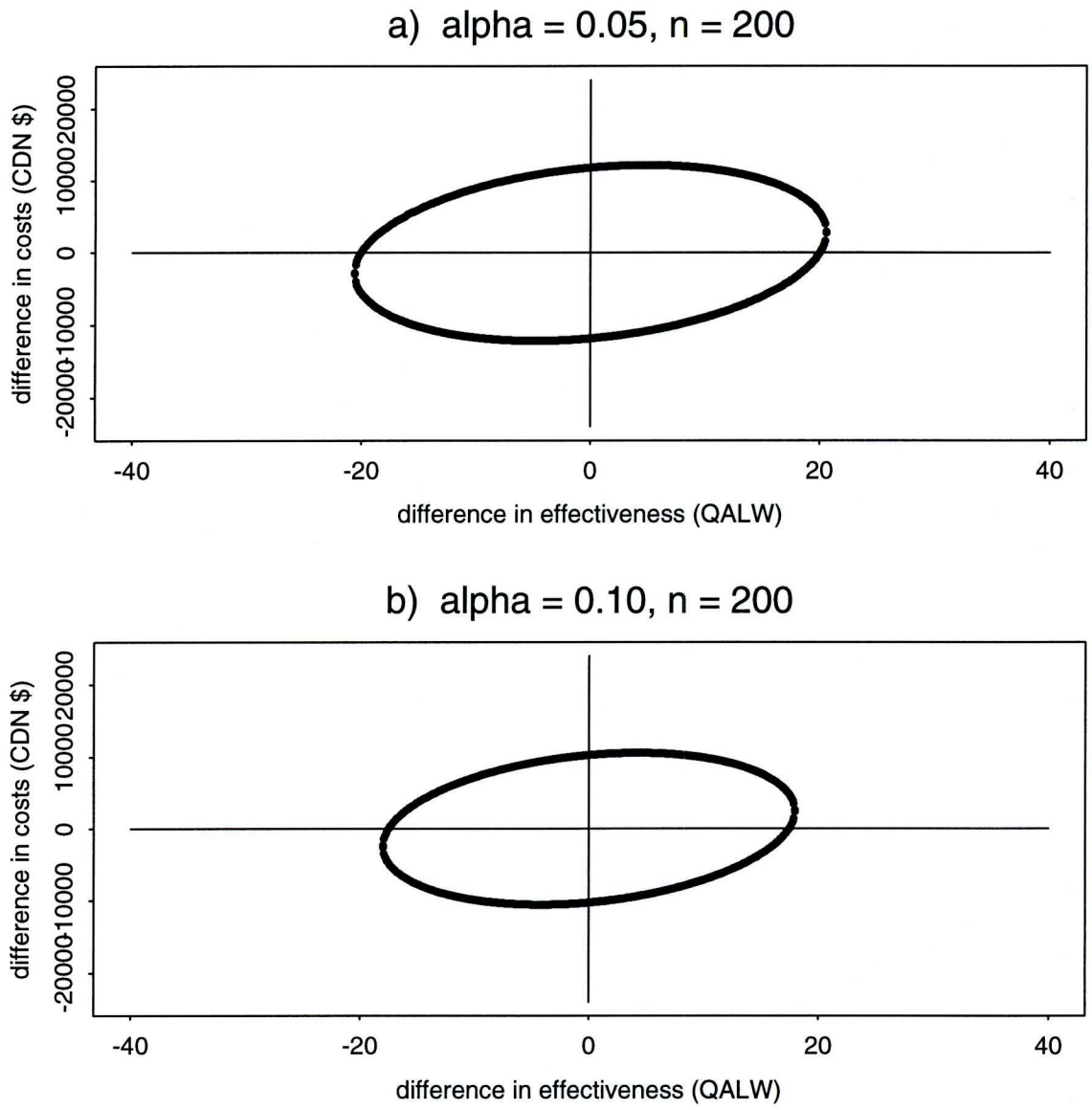


Figure 20: Ellipses of adequate power (as defined in Chapter 3) for $n = n_T = n_S = 200$. The first graph shows the ellipse corresponding to a 95% confidence level, and the second graph shows the ellipse corresponding to a 90% confidence level.

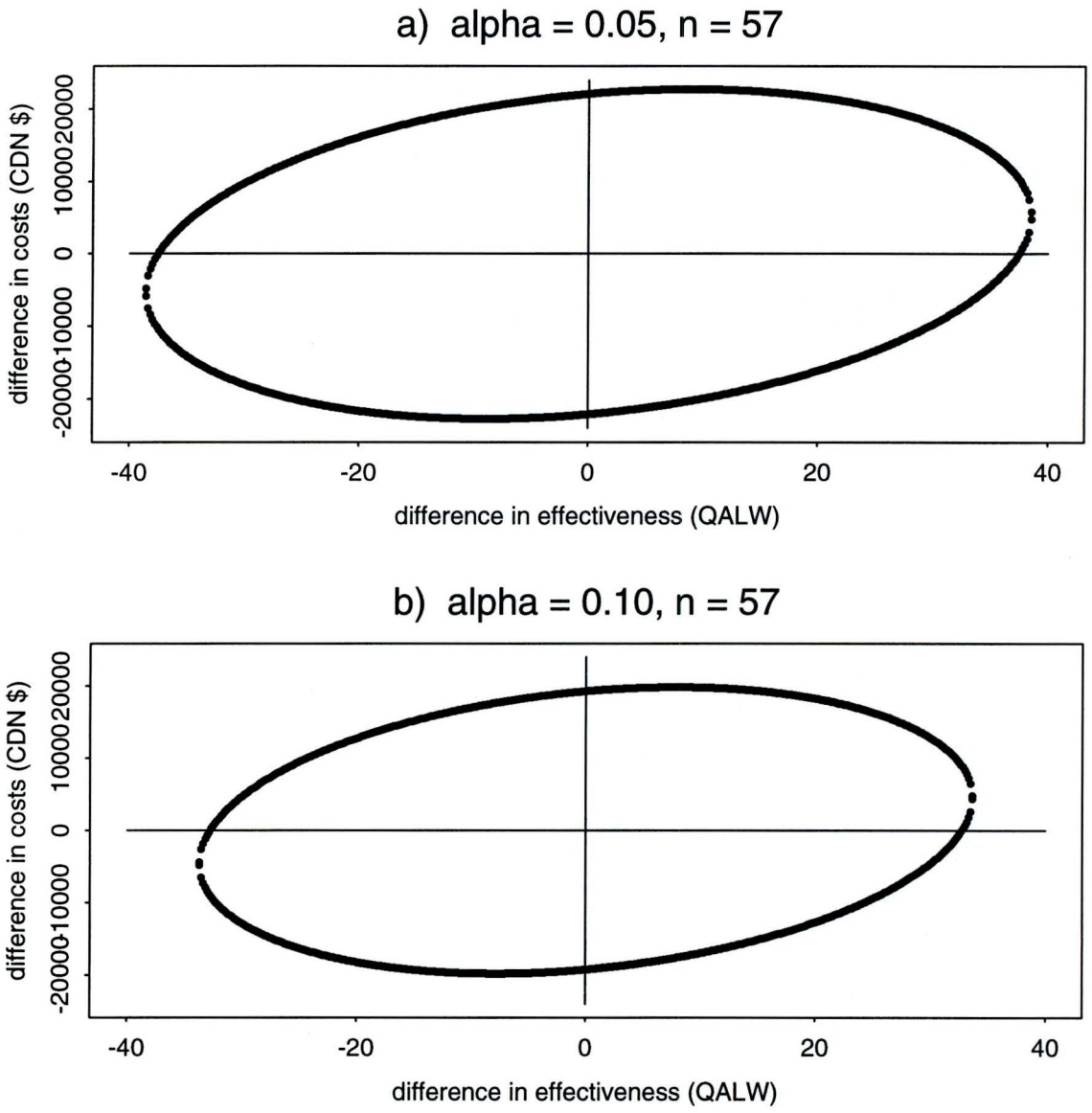


Figure 21: Ellipses of adequate power (as defined in Chapter 3) for $n = n_T = n_S = 57$. The first graph shows the ellipse corresponding to a 95% confidence level, and the second graph shows the ellipse corresponding to a 90% confidence level.

Chapter 6: Discussion

6.1 Assumptions

In considering the apparent validity of the methods, it is essential that the assumptions made in using the different methods be verified. Formal tests regarding distributional assumptions could be carried out. At the very least, an informal check, such as an examination of frequency histograms and normal probability plots, should be done to detect drastic deviations from the assumed distributions. Histograms of costs and effects in both treatment arms showed deviations from normality for both costs and effects in the two groups. However, histograms and normal probability plots of bootstrap re-sample estimates of $\overline{E}_T - \overline{E}_S$ and $\overline{C}_T - \overline{C}_S$ support the assumption that these two statistics have normal distributions. This is expected based on the Central Limit Theorem, and supports the findings of Willan and O'Brien (1996). The assumption of Fieller's theorem is that these two statistics have a bivariate normal distribution. The fact that the two statistics individually have marginal normal distributions is a necessary but not sufficient condition for their joint distribution to be bivariate normal. Further tests, such as those described by Johnson and Wichern (1992), could be used to verify the plausibility of bivariate normality of $\overline{E}_T - \overline{E}_S$ and $\overline{C}_T - \overline{C}_S$.

The assumption made by the Taylor's Series approximation and the normal-assumption bootstrap method which takes into account only the bootstrap ICER estimates is that $\hat{R}$ has a normal distribution. Histograms and normal probability plots of bootstrap estimates of $\hat{R}$ do not appear to support this assumption. The tails of the distribution are longer than in a normal distribution, resulting from the fact that very high and very low estimates of the ICER are sometimes obtained. Because of this, the histograms of bootstrap estimates of $\hat{R}$ are not very informative. The skewness coefficients of these frequency distributions reveal a left-skewness, particularly strong when B is high, since more extreme (very low) estimates of $\hat{R}$ are then obtained for some bootstrap replicates. Chaudhary and Stearns (1996) found that for moderate sample sizes the distribution of $\hat{R}$ is positively skewed in most cases. The limiting distribution of $\hat{R}$ is known to be normal. Chaudhary and Stearns (1996) commented that large sample results can be used if the sample sizes are greater than 30, and the coefficients of variation for the numerator and denominator are less than 0.1. Although the first condition is satisfied for the data set used in this study, the second one is not.

One of the methods based on using ICER angles assumes that the angles are normally distributed. Histograms and normal probability plots of bootstrap re-sample estimates of ICER angles, suggest that this assumption is more reasonable than the assumption of normality of $\hat{R}$. However, a slight left-skewness is observed in the

frequency distributions of the ICER angle estimates. More formal tests should be used to check whether this is a significant deviation from normality.

6.2 Comparison of Inferential Methods

Although it has already been shown that the assumption of normality of $\hat{R}$ does not seem justified in this case, it is interesting to compare the results of the two methods which relied on this assumption. The Taylor Series method has been criticized not only for its normality assumption, but also because the variance calculation is only approximate. Comparing the Taylor Series approximation CI to the bootstrap CI which relies on the assumption of normality of $\hat{R}$, reveals that the Taylor Series interval is markedly more narrow, due to the fact that the Taylor Series approximation gives a lower estimate of variance of $\hat{R}$, than does the bootstrapping method. This seems to indicate that the Taylor Series approximation may underestimate the variance. This finding agrees with Obenchain's (1997) comment that Taylor Series CIs are too narrow. Polsky *et al.* (1997) also found that the Taylor series method gave anti-conservative intervals, however, they showed that this was due to asymmetric underestimation of the upper limit. In this study, the problem appears to be asymmetric overestimation of the lower limit.

Although it appears that the Taylor series estimate of variance of $\hat{R}$ is too small, the bootstrap estimate is also questionable. Briggs, Wonderling, and Mooney (1997) found that bootstrap estimates of standard error are unstable (because they are susceptible

to the effect of "unusual" observations). This is supported by the fact that the intervals which rely on this estimate are very unstable and increase in width as the number of bootstrap replications increases.

In comparison to the Taylor series interval, the CI calculated using the Fieller's theorem approach is wider, having a considerably smaller lower confidence limit. This is because the Fieller's theorem method accounts for the negative skewness of the distribution of $\hat{R}$. Obenchain (1997) noted that because Fieller's Theorem intervals are "bow-tie shaped intervals" they also tend to be too narrow when $(\overline{E}_T - \overline{E}_S, \overline{C}_T - \overline{C}_S)$ is not significantly different from (0,0). In this study, the Fieller's theorem intervals do not appear to be too narrow as compared to bootstrap intervals which rely on estimates of the ICER (ratio) only. However, they are narrower than some of the angle method intervals.

Chaudhary and Stearns (1996) found Fieller's theorem intervals to be similar to bootstrap percentile, and bias-corrected percentile intervals based on ICER slopes only. In this study, although the percentile, bias-corrected percentile and bias-corrected accelerated percentile methods for slopes gave similar results, these intervals had a higher lower limit then the interval based on Fieller's theorem. However, since a considerable proportion of bootstrap replicates gave estimates of $(\overline{E}_T - \overline{E}_S, \overline{C}_T - \overline{C}_S)$ in the 2nd and 3rd quadrants, the methods which consider only ICER slopes cannot be considered reliable. This suggests that methods which made use of ICER angles should be superior to those which only utilized the bootstrap estimates of $\hat{R}$.

Plots of $(\Delta E^*, \Delta C^*)$ points resulting from bootstrap replications reveal that bootstrap methods for constructing CIs which consider ICER angles need to be used in this study. Several methods were proposed which made use of ICER angles. The method which assumed the ICER angles were normally distributed gave narrower intervals than the others. However, as the normality assumption seems questionable, this method cannot be considered very reliable. Because of the slight left skewness of the distribution, the "central CIs", which do not correct for the bias, may also be inaccurate. Furthermore, as Obenchain (1997) pointed out, calculating CIs using this method divides the bootstrap observations rather artificially using the 180°-line in the 2nd quadrant. The bias-correction and accelerated bias-correction should provide an improvement over the "central CI" method by accounting for the bias caused by skewness in the distribution of ICER angles, however the artificial division of points in the 2nd quadrant still exists. Furthermore, it should be noted that these intervals, particularly the lower limits, are highly unstable. The smallest wedge-shaped region method should provide fairly reliable confidence limits in cases where bootstrapping results in $(\Delta E^*, \Delta C^*)$ points in all four quadrants. Although this is perhaps the most reliable of all the bootstrap methods for this data, it also gives quite variable results, even when many replications (B = 10,000) are performed.

Because the point estimate $(\Delta \overline{E}, \Delta \overline{C})$ falls in the 4th quadrant, it could be argued that it does not make sense to report results using the ICER. Several authors (including Siegel *et al.*, 1996; Stinnett and Mullahy, 1997) have pointed out that the magnitude of a

negative ICER conveys no useful information. This suggests that the lower limit which is (usually) in 4th quadrant is uninterpretable, except to show dominance. However, the upper limit is of great importance since it represents the highest possible cost per unit outcome consistent with the study's results. Thus, perhaps it would be useful to find one sided intervals, placing more emphasis on the upper confidence limit. Although the upper limit from some of the methods could be interpreted as a upper limit of a 100(1-α/2)% one-sided confidence interval, not all of the methods could be used in this way. It was pointed out that the lower limits of some of the angle-method intervals were quite unstable. In light of the fact that the magnitude of the lower limit may not be considered interpretable, perhaps this is not of great concern. Several authors have noted that any method which hinges on the magnitude of negative ICER estimates should be interpreted cautiously, as it relies on "meaningless numbers" (Stinnett and Mullahy, 1997; Briggs, Wonderling and Mooney, 1997). This suggests that of all the angle methods, the "smallest wedge shaped region" (and in particular its upper limit) is the most useful of the confidence interval results. Based on the examples with B = 10,000, the upper limit of the confidence interval of the ICER can, thus, be estimated by approximately CDN $530/QALW, which is an "acceptable" value according to the threshold proposed by Laupacis *et al.* (1992).

In a situation where the ICER point estimate is negative, and dominance appears to be highly likely, estimating the probabilities of dominance or "cost-effectiveness" (based on predefined thresholds of what is cost-effective) is very useful . Alternatively,

one may consider a hypothesis test of dominance, such as the ones proposed by Laska, Meisner, and Siegel (1997). The advantage of the probability estimation approach is that bootstrap results can be used, whereas the hypothesis testing approaches derived by Laska, Meisner, and Siegel (1997) made use of parametric methods.

6.3 Further Issues

Since the results from cost-effectiveness analyses will likely be used increasingly for clinical decision-making, improvements in the design and analysis of cost-effectiveness studies are required. New recommendations for conducting cost-effectiveness analysis have recently been published by the U.S. Panel on Cost-Effectiveness in Health and Medicine (Weinstein *et al.*, 1996), yet many issues regarding the design and analysis of stochastic cost-effectiveness studies remain unresolved.

It is agreed upon that studies intended to provide cost-effectiveness information should be designed appropriately. As Willan and O'Brien (1996) point out, collecting cost data adds to the cost and duration of a study. Although it is agreed upon that both patient specific costs and effects should be measured, it is not clear how these quantities should be measured. Prices for health care utilization are very difficult to measure, and are generally estimated. Several authors have discussed some of the difficulties involved in price estimation, and the bias that can result from methodological choices. Neumann *et al.* (1997) discuss that once all the component resources are identified, they must be

assigned a value. However, it is difficult to know the actual value of the resources consumed, and approximations are often used by analysts. Many researchers (Weinstein *et al.*, 1996; Elliott and Harris, 1997) stress the importance of discounting of costs, which is a method of adjusting costs for different timing.

There are also some difficulties in measuring effects. Cost-effectiveness of diverse medical interventions needs to be evaluated in similar terms, in order to allow decision makers to determine which of many competing interventions produces the greatest overall gain in health for the resources used. This means that the costs and effects of many alternative interventions should be measured in the same units. Since quality-adjusted life years (QALYs) incorporate both prolongation and quality of life, these are ideal units for comparing interventions that benefit patients in very different ways. However, there are many problems and discrepancies in measuring QALYs (Sacristán *et al.*, 1995) and methods for determining quality weights (used in the calculation of QALYs) continue to be an active area of research and debate (Neumann *et al.*, 1997).

Having obtained data on costs and effects of alternative interventions, the analyst must decide whether to analyze the data in its original state, or whether the data needs to be transformed. Obenchain (1997) discussed the usefulness of log transformations to symmetrize highly skewed cost distributions. The impact of such transformations on the analysis and results needs to be examined further (though Obenchain has discussed some of the issues).

Although it is clear that point estimates are insufficient, proper statistical analysis of cost-effectiveness data is a complicated issue. First of all, there is still some debate over the type of analysis that is most useful for decision-making, and benefits of using the ICER are still questioned by some. In a discussion of some of the difficulties associated with interpreting negative ICERs, Stinnett and Mullahy (1997) stated that they prefer reporting economic evaluations in terms of a net benefit rather than as a C/E ratio. Rittenhouse (1995) feels that cost-effectiveness analyses may yield inconsistent results, and proposes using the "more formally developed method" of cost-benefit analysis. However, many researchers studying cost-effectiveness analysis agree that the incremental cost-effectiveness ratio is a useful method of judging the cost-effectiveness of competing therapies, and in particular, that estimation of CIs for the ICER is the "appropriate focus for cost-effectiveness analysis" (Briggs and Fenn, 1997).

Presentation of results using confidence intervals has obvious appeal, however as the discussion of the potential methods of calculating CIs for the ICER shows, it is a very complicated issue. Although much has been accomplished in improving the statistical methodology in this area, little is still known about the relative advantages and disadvantages of the alternative approaches. Several authors have provided theoretical arguments in an attempt to compare the validity of the different methods. Briggs, Wonderling, and Mooney (1997) provided an extensive explanation regarding bias-correction in bootstrap estimation methods. Chaudhary and Stearns (1996) and Briggs,

Wonderling, and Mooney (1997) considered the number of bootstrap replications required for stable confidence limit estimates from bootstrap methods.

Which methods are the most reliable under what circumstances, and how they should be applied, remains debatable. The alternative methods need to be tested by Monte Carlo simulation methods, under various assumptions. Polsky *et al.* (1997) conducted such a study (but without varying some important assumptions, and comparing only a few of the available methods), and Briggs, Wonderling, and Mooney (1997) report that they are currently working on such a study.

Some of the researchers studying the statistical methods of analyzing cost-effectiveness data have pointed out that sensitivity analysis should still be incorporated, even when sampled costs and effects are available from the same clinical trial. Willan and O'Brien (1996) felt that sensitivity analysis is particularly useful for assessing uncertainty about the external validity and generalizability of study results to the real world. The potential lack of generalizability of the results of cost-effectiveness analyses has been questioned, especially because of the uncertainty in price estimates, which may vary depending on the institution participating in the trial. Sensitivity analysis can be used to test the robustness of results to changes in costs.

Further issues that need to be addressed regard the reporting of results of cost-effectiveness analyses. At this stage, since the methods are still being developed and debated, perhaps what is most important is that details of both the design and the analysis are reported. Neumann *et al.* (1997) discussed the importance of providing explicit

information about how studies were conducted, including the sources and characteristics of the data used, and what assumptions were made. This type of information is important in judging the generalizability of results. Obenchain (1997) stressed the importance of reporting "technical details" of bootstrap analyses, such as the number of replications and the random number seed used.

Beyond the statistical issues, those concerned with the applicability of CEA to decision making, need to consider many other questions. For instance Baltussen, Leidl, and Ament (1996) discussed the importance of taking the age of potential users of the health care intervention under consideration into account during decision-making based on cost-effectiveness ratios. Phillips and Hotlgrave (1997) described issues of relevance to the cost-effectiveness of preventive medicine.

Economic analyses will play an increasingly influential role in the allocation of limited resources. Because resources are limited, consideration of opportunity cost, is believed to be of great importance (Elliott and Harris, 1997; Willan and O'Brien, 1996). The opportunity cost takes into account the benefits forgone by committing resources to one program instead of others (Willan and O'Brien, 1996). This has prompted Willan and O'Brien (current research) to consider a model for cost-effectiveness which incorporates not only the incremental cost-effectiveness of a treatment relative to standard, but also where the additional resources would come from, and the effects of such a resource reallocation. This is another issue that needs to be taken into account during policy making based on cost-effectiveness analyses.

As the above discussion indicates, cost-effectiveness analysis and its use in decision-making are very complex. Statisticians should develop methods for the proper statistical analysis of cost-effectiveness data, including the presentation of results in a format that is useful to decision makers. Although much progress has been made, considerable work remains to be done in this area.

References

Adams M.E., McCall N.T., Gray D.T., Orza M.J., and Chalmers T.C. (1992) Economic analysis in randomized control trials, *Medical Care* **30**, 231-238.

Baltussen R., Leidl R., and Ament A. (1996) The impact of age on cost-effectiveness ratios and its control in decision making, *Health Economics* **5**, 227-239.

Black W.C. (1990) The CE Plane: a graphic representation of cost-effectiveness, *Medical Decision Making* **10**, 212-214.

Bloomfield D.J., Krahn M.D., Neogi T., Panzarella T., Smith T.J., Warde P., Willan A.R., Ernst S., Moore M.J., Neville A., and Tannock I.F. (1998) Economic evaluation of chemotherapy with mitoxantrone plus prednisone for symptomatic hormone resistant prostate cancer, based on a Canadian randomized trial with palliative endpoints. Submitted to *Journal of Clinical Oncology.*

Briggs A., and Fenn P. (1997) Trying to do better than average: a commentary on 'statistical inference for cost-effectiveness ratios', *Health Economics* **6**, 491-495.

Briggs A.H., Wonderling D.E., and Mooney C.Z. (1997) Pulling cost-effectiveness up by its bootstraps: a non-parametric approach to confidence interval estimation, *Health Economics* **6**, 327-340.

Casella G., and Berger R.L. (1990) *Statistical Inference*, Wadsworth and Brooks/Cole, Pacific Grove, California.

Chaudhary M.A., and Stearns S.C. (1996) Estimating Confidence Intervals for cost-effectiveness ratios: an example from a randomized trial, *Statistics in Medicine* **15**, 1447-1458.

Cochran W.G. (1977) *Sampling Techniques*, John Wiley and Sons, New York.

Efron B. (1987) Better bootstrap confidence intervals (with discussion), *Journal of the American Statistical Association* **82**, 171-200.

Efron B., and Tibshirani R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Elliott S.L., and Harris A.H. (1997) The methodology of cost-effectiveness analysis: avoiding common pitfalls, *Medical Journal of Australia* **166**, 636-639.

Fieller E.C. (1954) Some problems in interval estimation (with discussion), *Journal of the Royal Statistical Society, Series B* **16**, 175-185.

Gold M.R., Siegel J.E., Russell L.B., and Weinstein M.C. (1996) *Cost Effectiveness in Health and Medicine*, Oxford University Press, Oxford.

Guyatt G.H., Jaeschke R.Z., Heddle N., Cook D.J., Shannon H., and Walter S. (1995) Basic statistics for clinicians: Interpreting study results: Confidence intervals, *Canadian Medical Association Journal* **152**, 169-173.

Johnson R.A., and Wichern D.W. (1992) *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, California.

Laska E.M., Meisner M., and Siegel C. (1997) Statistical inference for cost-effectiveness ratios, *Health Economics* **6**, 229-242.

Laupacis A., Feeny D., Detsky A., and Tugwell P. (1992) How attractive does a new technology have to be to warrant adoption and utilisation? Tentative guidelines for using clinical and economic evaluations, *Canadian Medical Association Journal* **146**, 473-481.

Neumann P.J., Hermann R.C., Berenbaum P.A., and Weinstein M.C. (1997) Methods of cost-effectiveness analysis in the assessment of new drugs for Alzheimer's disease, *Psychiatric Services* **48**, 1440-1444.

Obenchain R.L., Melfi C.A., Croghan T.W., and Buesching D.P. (1997) Bootstrap analyses of cost effectiveness in antidepressant pharmacotherapy, *Pharmacoeconomics* **11**, 464-472.

Obenchain R.L. (1997) Issues and algorithms in cost-effectiveness inference, *Biopharmaceutical Report* **5**, 1-7.

Obenchain R.L. (1995) ICERconf: Confidence intervals for incremental cost-effectiveness ratios. (A manual available with the program ICERconf: version 9609), Eli Lilly and Company, Indianapolis.

O'Brien B.J., Drummond M.F., Labelle R.J., and Willan A. (1994) In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care, *Medical Care* **32**, 150-163.

Phillips K.A., and Hotlgrave D.R. (1997) Using cost-effectiveness/cost-benefit analysis to allocate health resources: a level playing field for prevention? *American Journal of Preventive Medicine* **13**, 18-25.

Polsky D., Glick H.A., Willke R., and Schulman K. (1997) Confidence intervals for cost-effectiveness ratios: a comparison of four methods, *Health Economics* **6**, 243-252.

Rittenhouse B.E. (1995) Potential inconsistencies between cost-effectiveness and cost-utility analyses. An upstairs/downstairs socioeconomic distinction, *International Journal of Technology Assessment in Health Care* **11**, 365-376.

Sacristán J.A., Day S.J., Navarro O., Ramos J., and Hernández J.M. (1995) Use of confidence intervals and sample size calculations in health economic studies, *The Annals of Pharmacotherapy* **29**, 719-725.

Siegel J.E., Weinstein M.C., Russell L.B., Gold M.R. (1996) Recommendations for reporting cost-effectiveness analyses, *JAMA* **276**, 1339-1341.

Stinnett A.A., and Mullahy J. (1997) The negative side of cost-effectiveness analysis (letter to the editor), *JAMA* **277**, 1931-1932.

Tannock I.F., Osoba D., Stockler M.R., Ernst D.S., Neville A.J., Moore M.J., Armitage G.R., Wilson J.J., Venner P.M., Coppin C.M., and Murphy K.C. (1996) Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: a Canadian randomized trial with palliative end points, *Journal of Clinical Oncology 14*, 1756-1764.

van Hout B.A., Al M.J., Gordon G.S., and Rutten F.F. (1994) Costs, effects and c/e-ratios alongside a clinical trial, *Health Economics* **3**, 309-319.

Weinstein M.C., Siegel J.E., Gold M.R., Kamlet M.S., and Russell L.B.(1996) Recommendations of the Panel on Cost-effectiveness in Health and Medicine, *JAMA* **276**, 1253-1258.

Weinstein M.C., and Stason W.B. (1977) Foundations of cost-effectiveness analysis for health and medical practices, *New England Journal of Medicine* **314**, 253-255.

Willan A.R., and O'Brien B.J. (1996) Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem, *Health Economics* **5**, 297-305.

Willan A.R., and O'Brien B.J. (1998) Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. Submitted to *Health Economics*.

Appendix A: SAS Programs

Program 1: Data input

```
options ls=78 ps=60 nodate nonumber;
libname  SASF '/2/biernac';

data SASF.pmh_icer;
      infile '/2/biernac/icerf.dat';
      input id trt costs qalw_tr util_tr;
      attrib id label='Patient Identifier'
            trt label='Allocation; 1=M+P; 2=P'
            costs label='Cost for Individual Patient'
            qalw_tr label='Qual. Adj. Weeks, Trans. Utility'
            util_tr label='Transformed Utility';
run;

Proc contents data=SASF.pmh_icer;
      title 'Contents of SASF.PMH_ICER';
run;
```

Program 2: Plots and summary statistics

```
proc sort;
      by TRT;
run;

proc plot;
      by TRT;
      plot COSTS*QALW_TR;
      title 'Costs versus Quality Adjusted Life Weeks';
run;

proc means;
run;

proc means;
      by TRT;
run;
```

Program 3: Creation of a SAS file with covariances

```
options ls=78 ps=60 nodate nonumber;
libname SASF '/2/biernac';

proc corr data=SASF.pmh_icer nocorr cov outp=SASF.COVSF;
      var costs qalw_tr;
      by trt;
run;
```

Program 4: Creation of a matrix of summary statistics (including covariances)

```
options ls=78 ps=60 nodate nonumber;
libname SASF '/2/biernac';

proc IML;
      reset print;
      use SASF.covsf(type=cov);
      list all;
      read all into x;
      print x;
      reset storage="SASF.matrix2";
      store;
quit;
```

Program 5: Calculation of Taylor's Series CI and Fieller's Theorem CI (closed form)

```
options ls=78 ps=60 nodate nonumber;
libname SASF '/2/biernac';

proc IML;
      reset noprint;
      reset storage="SASF.matrix2";
      load X;
      Ct=X[2,2];
      Mct=X[3,2];
      Met=X[3,3];
      Vct=X[1,2];
      Vet=X[2,3];
      Nt=X[5,2];
      Cc=X[7,2];
      Mcc=X[8,2];
      Mec=X[8,3];
      Vcc=X[6,2];
      Vec=X[7,3];
      Nc=X[10,2];
```

```
      R=(Mct-Mcc)/(Met-Mec);
      Cnn=(Vct/Nt+Vcc/Nc)/(Mct-Mcc)**2;
      Cdd=(Vet/Nt+Vec/Nc)/(Met-Mec)**2;
      Cnd=(Ct/Nt+Cc/Nc)/((Mct-Mcc)*(Met-Mec));
      VR1=(R**2)*(Cnn+Cdd-2*Cnd);
      z=1.96;
      lb1=R-z*sqrt(VR1);
      ub1=R+z*sqrt(VR1);
      int1=lb1||ub1;
      print "The point estimate of the incremental cost-effectiveness
      ratio is:", R;
      print "The Taylor method confidence interval is", int1;
      arg=(Cnn+Cdd-2*Cnd)-(z**2)*(Cnn*Cdd-Cnd**2);
      ub2=R*((1-(z**2)*Cnd)-z*sqrt(arg))/(1-(z**2)*Cdd);
      lb2=R*((1-(z**2)*Cnd)+z*sqrt(arg))/(1-(z**2)*Cdd);
      int2=lb2||ub2;
      print "The Fieller's method confidence interval is", int2;
quit;
```

Program 6: Calculation of Fieller's theorem CI (matrix form)

```
options ls=78 ps=60 nodate nonumber;
libname SASF '/2/biernac';

proc IML;
      reset noprint;
      reset storage="SASF.matrix2";
      load;
      Ct=X[2,2];
      Mct=X[3,2];
      Met=X[3,3];
      Vct=X[4,2]**2;
      Vet=X[4,3]**2;
      Nt=X[5,2];
      Cc=X[7,2];
      Mcc=X[8,2];
      Mec=X[8,3];
      Vcc=X[9,2]**2;
      Vec=X[9,3]**2;
      Nc=X[10,2];
      R=(Mct-Mcc)/(Met-Mec);
      x11=Vet/Nt+Vec/Nc;
      x12=Ct/Nt+Cc/Nc;
      x22=Vct/Nt+Vcc/Nc;
      s1=x11||x12;
      s2=x12||x22;
      S=s1//s2;
      xbar=Met-Mec//Mct-Mcc;
      call eigen(LAMBDA,GAMMA,S);
      inv_RS=T(GAMMA)*INV(SQRT(DIAG(LAMBDA)))*GAMMA;
      y=inv_RS*xbar;
      lengthy=sqrt(y[1]**2+y[2]**2);
      u1=(1/lengthy*y[1])||(1/lengthy*y[2]);
      u2=(1/lengthy*(-y[2]))||(1/lengthy*y[1]);
```

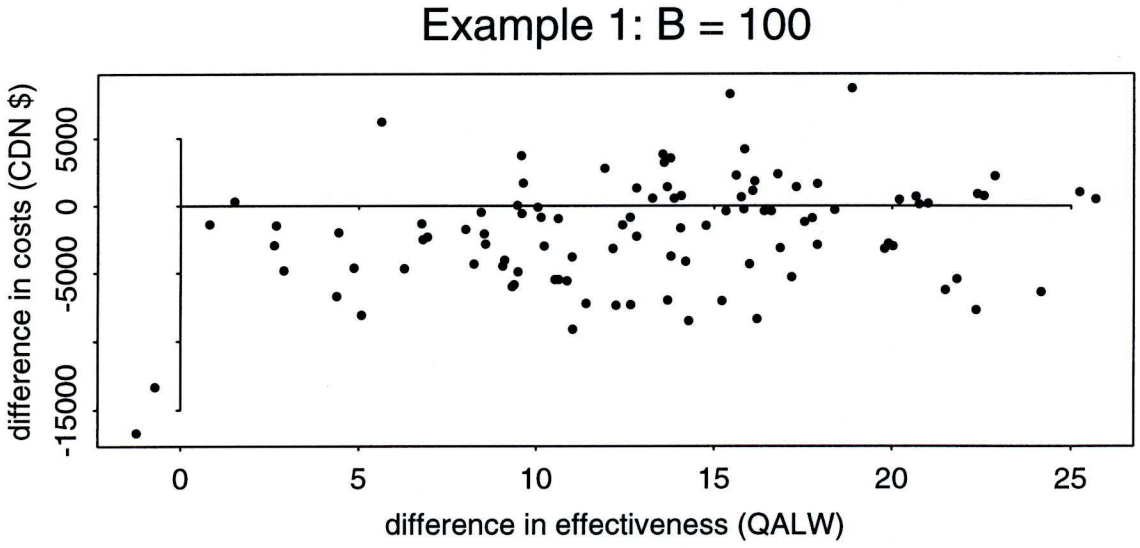
```
        U=u1//u2;
        T=U*inv_RS;
        z=1.96;
        x11=1;
        x12=-(z/lengthy)/(1-(z/lengthy)**2)**(1/2);
        x1=x11//x12;
        v1=T**(-1)*x1;
        x21=1;
        x22=(z/lengthy)/(1-(z/lengthy)**2)**(1/2);
        x2=x21//x22;
        v2=T**(-1)*x2;
        s1=v1[2]/v1[1];
        s2=v2[2]/v2[1];
        int3=s1||s2;
        print "The point estimate of the incremental cost-effectiveness
        ratio is:", R;
        print "The Fieller's method confidence interval is", int3;
quit;
```

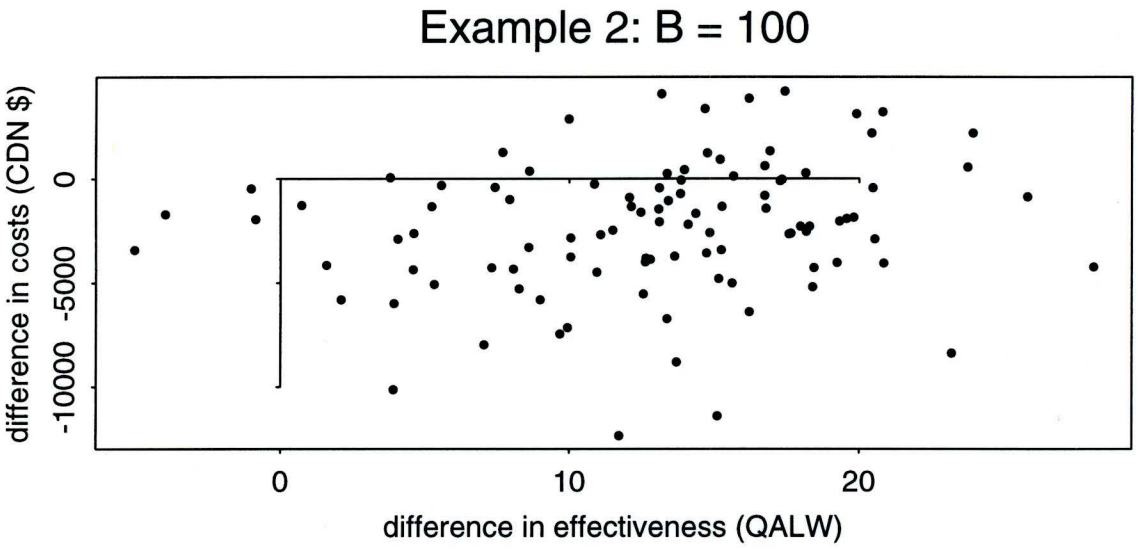
Appendix B: Random number seeds

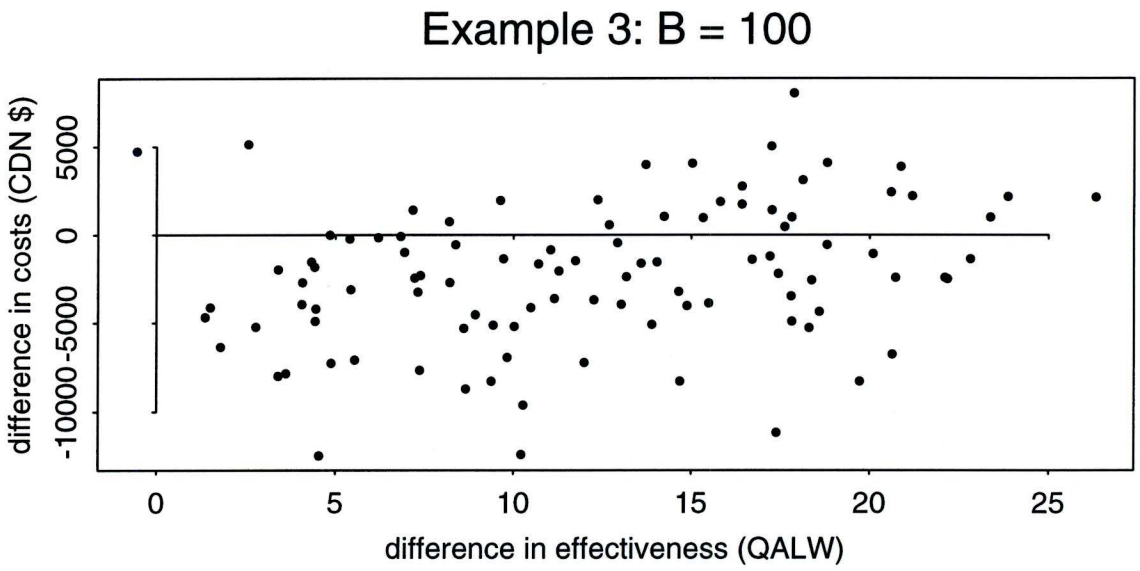
The random number seeds from Splus used for the 12 bootstrapping examples are given in the following table:

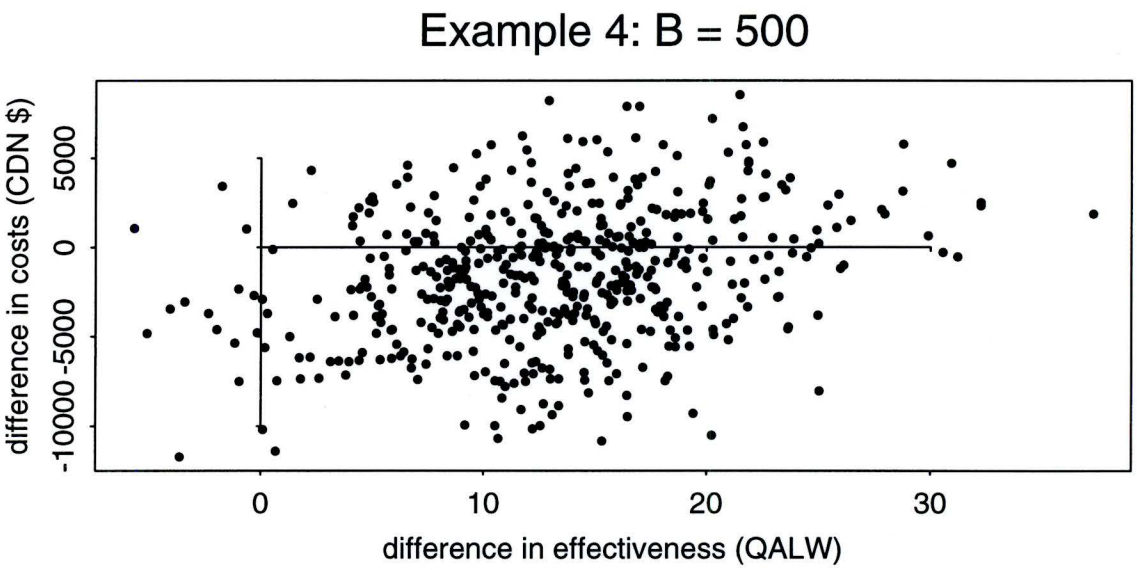
Example number	Random number seed vector
1	53 43 13 20 46 0 26 63 18 24 15 2
2	21 6 28 13 27 1 28 48 21 61 39 2
3	53 48 7 10 13 1 6 47 39 2 16 0
4	21 43 56 0 28 2 13 41 55 31 51 3
5	53 63 51 41 26 3 38 50 9 49 2 1
6	21 36 34 44 32 1 17 38 61 22 14 2
7	53 24 12 25 28 2 3 6 58 32 0 3
8	53 49 49 20 33 2 37 32 39 28 40 0
9	53 10 36 45 37 2 27 24 46 63 10 2
10	53 35 35 52 10 3 8 23 35 52 27 3
11	53 29 40 3 51 1 39 56 2 19 41 2
12	53 23 1 43 56 0 27 55 16 53 18 2

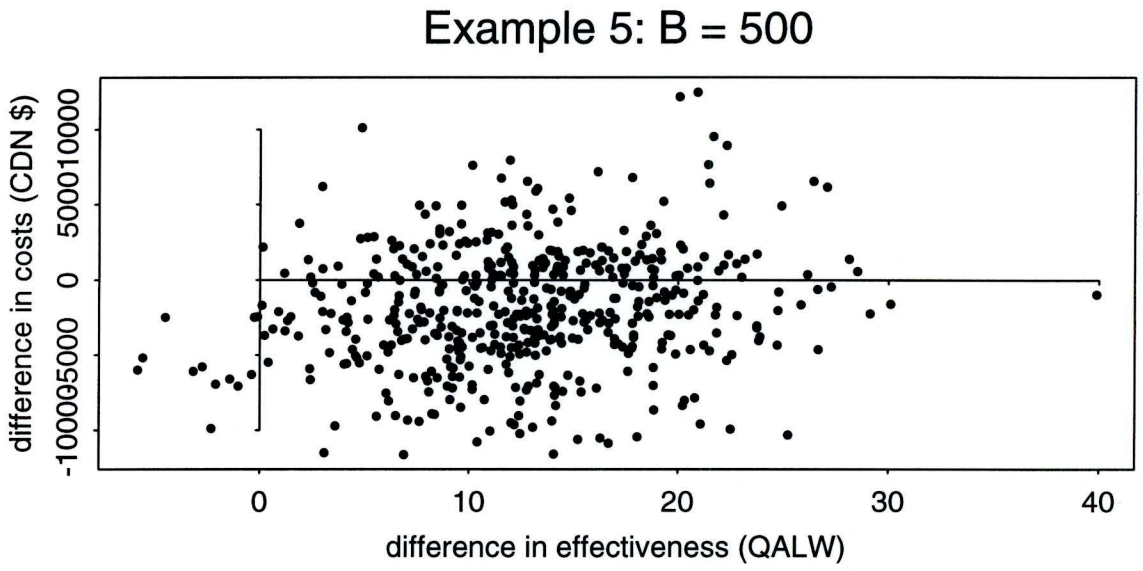
Appendix C: Plots of Bootstrap Re-sample Results

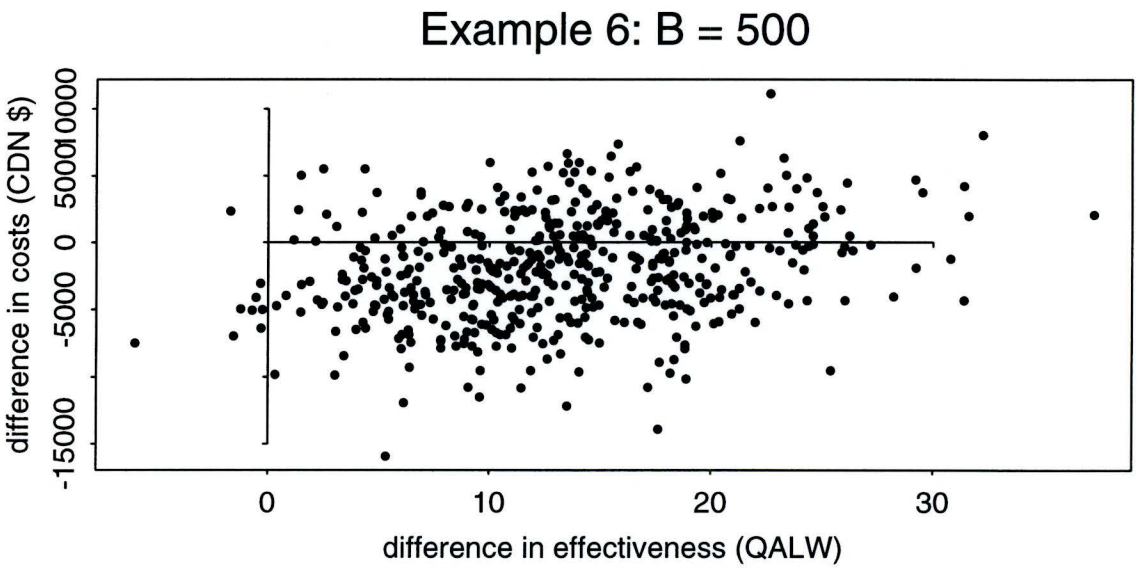


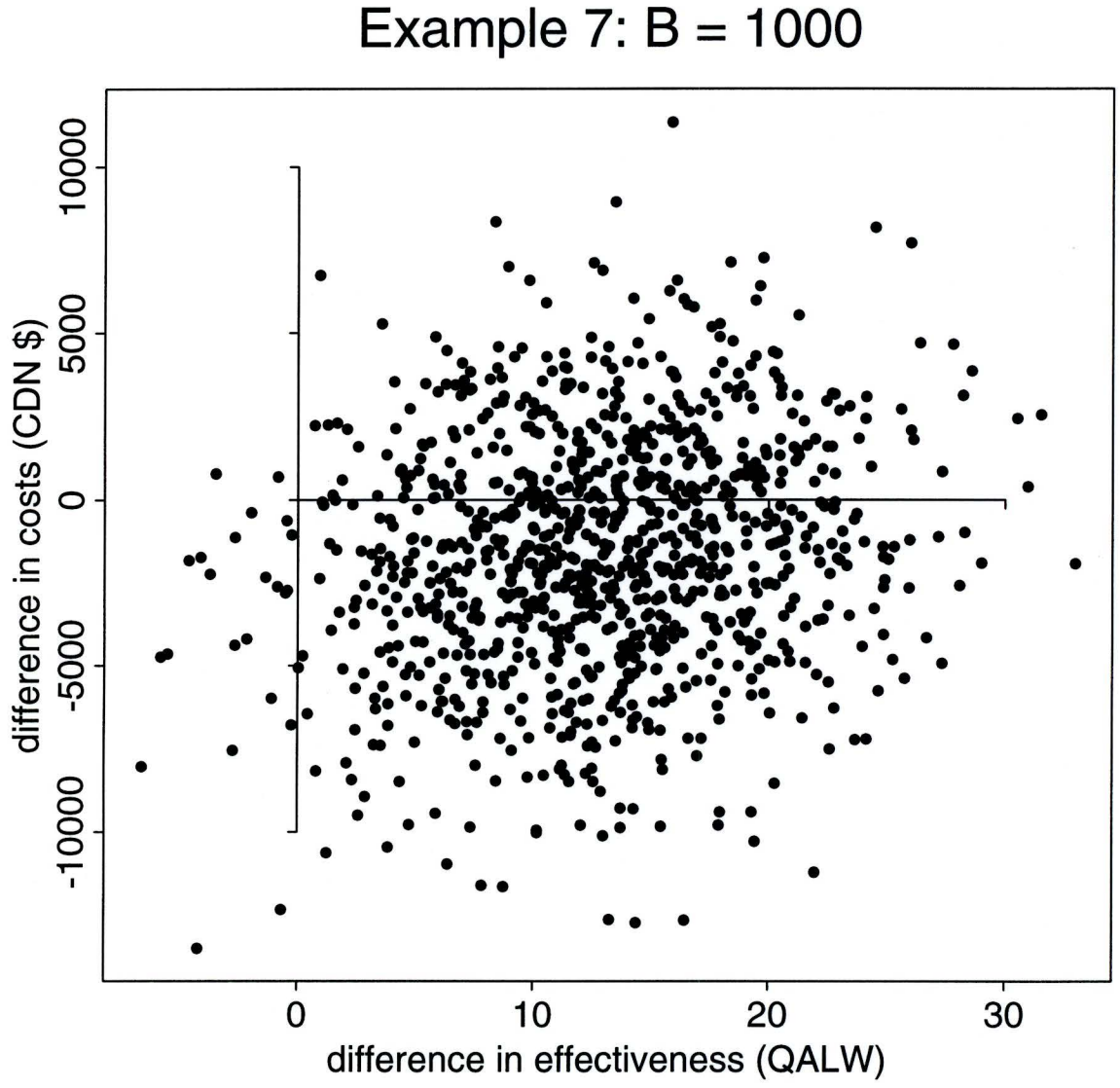


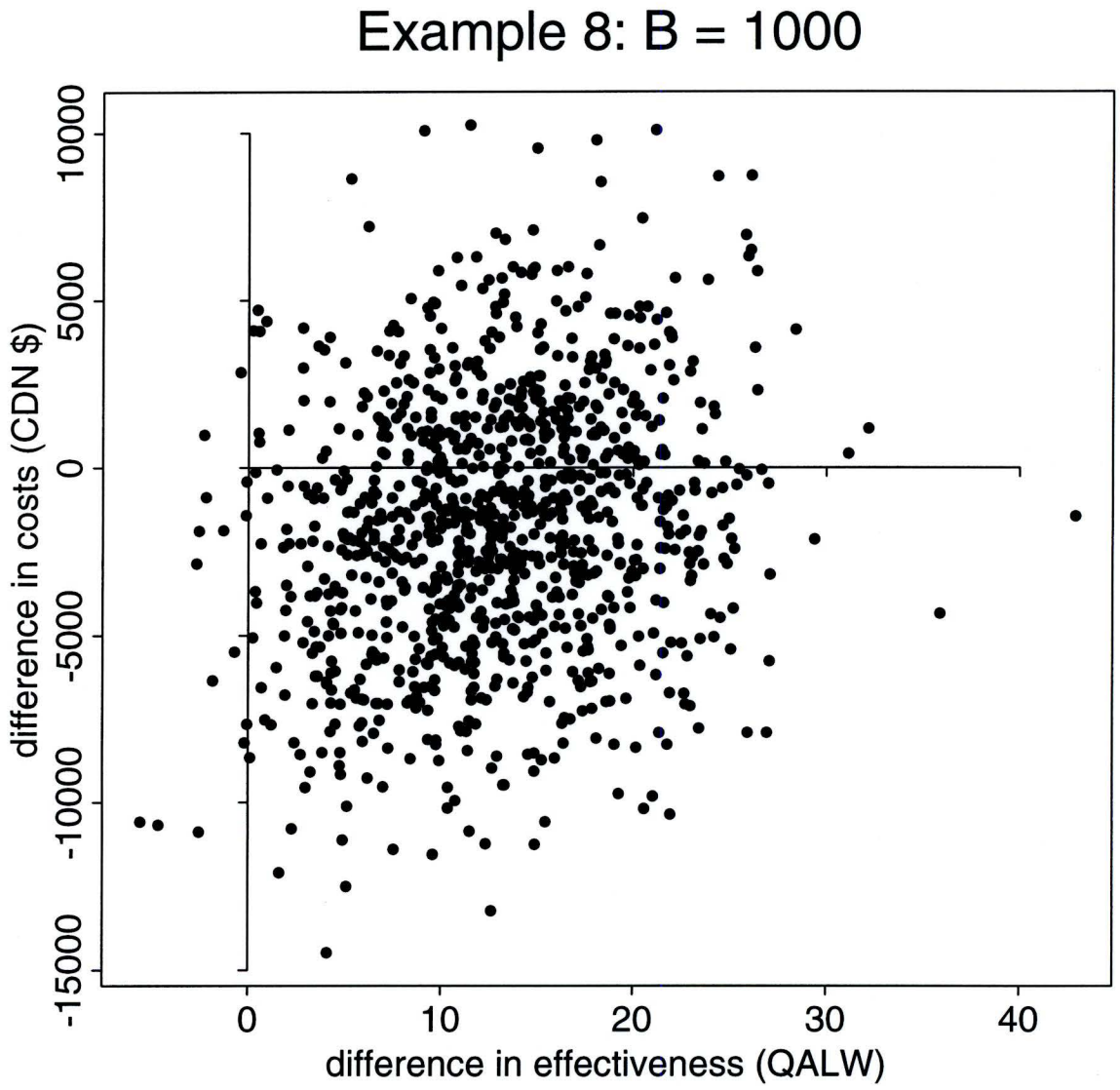

Example 3: B = 100
difference in costs (CDN $)
-10000
-5000
0
5000
0
5
10
15
20
25
difference in effectiveness (QALW)

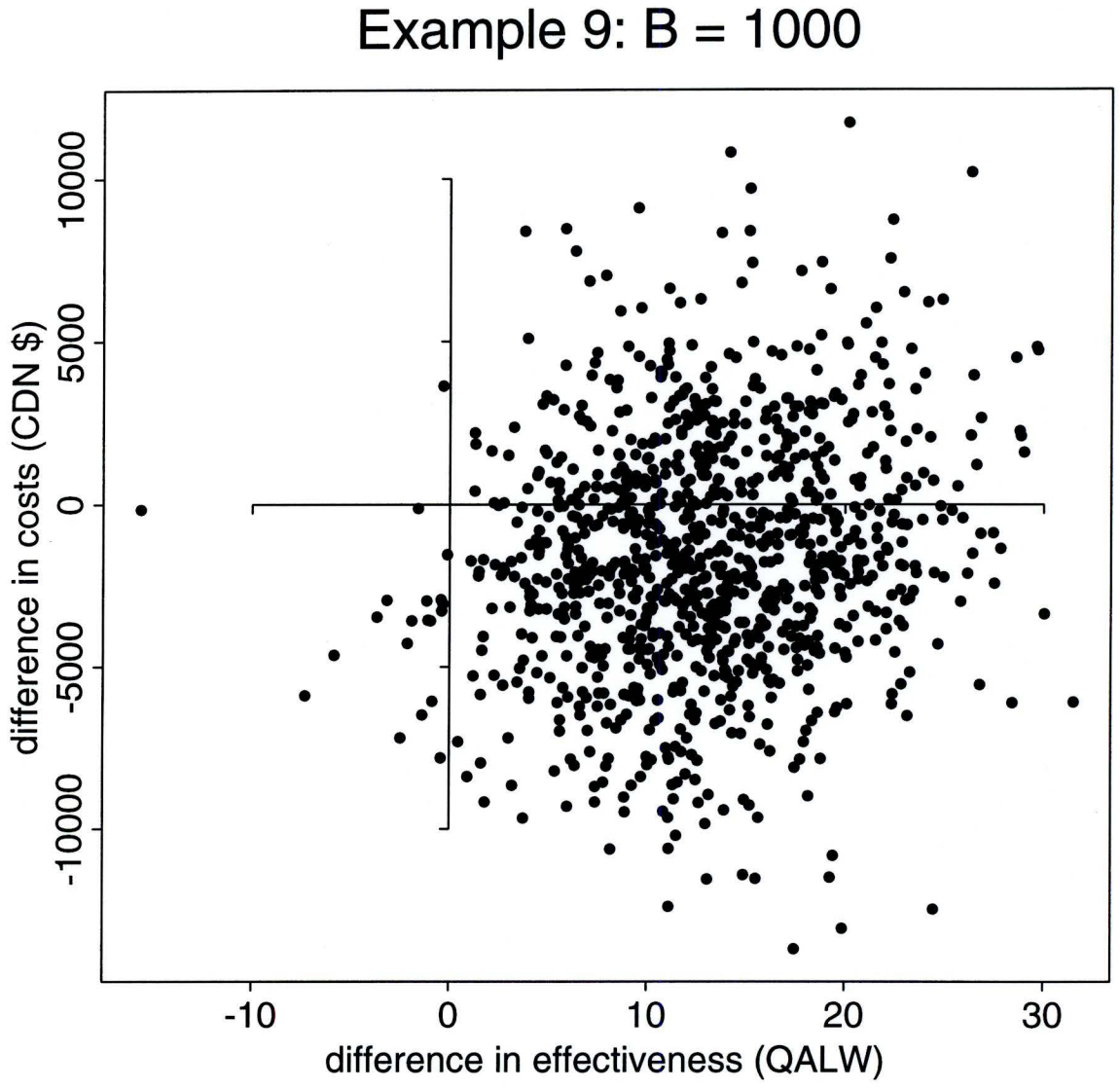

Example 4: B = 500
difference in costs (CDN $)
-10000
-5000
0
5000
0
10
20
30
difference in effectiveness (QALW)

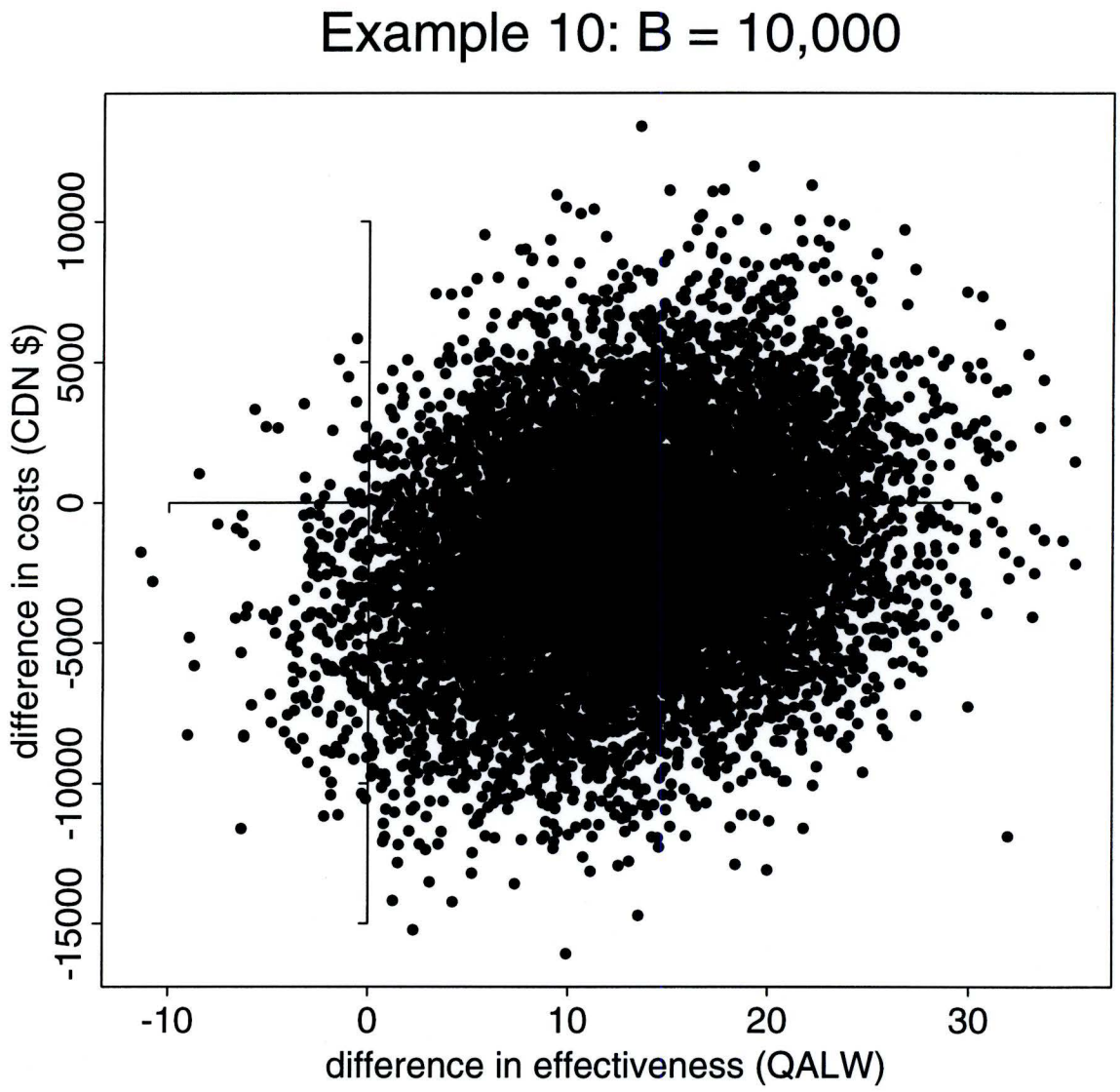

Example 5: B = 500
difference in costs (CDN $)
-10000 -5000 0 5000 10000
0 10 20 30 40
difference in effectiveness (QALW)

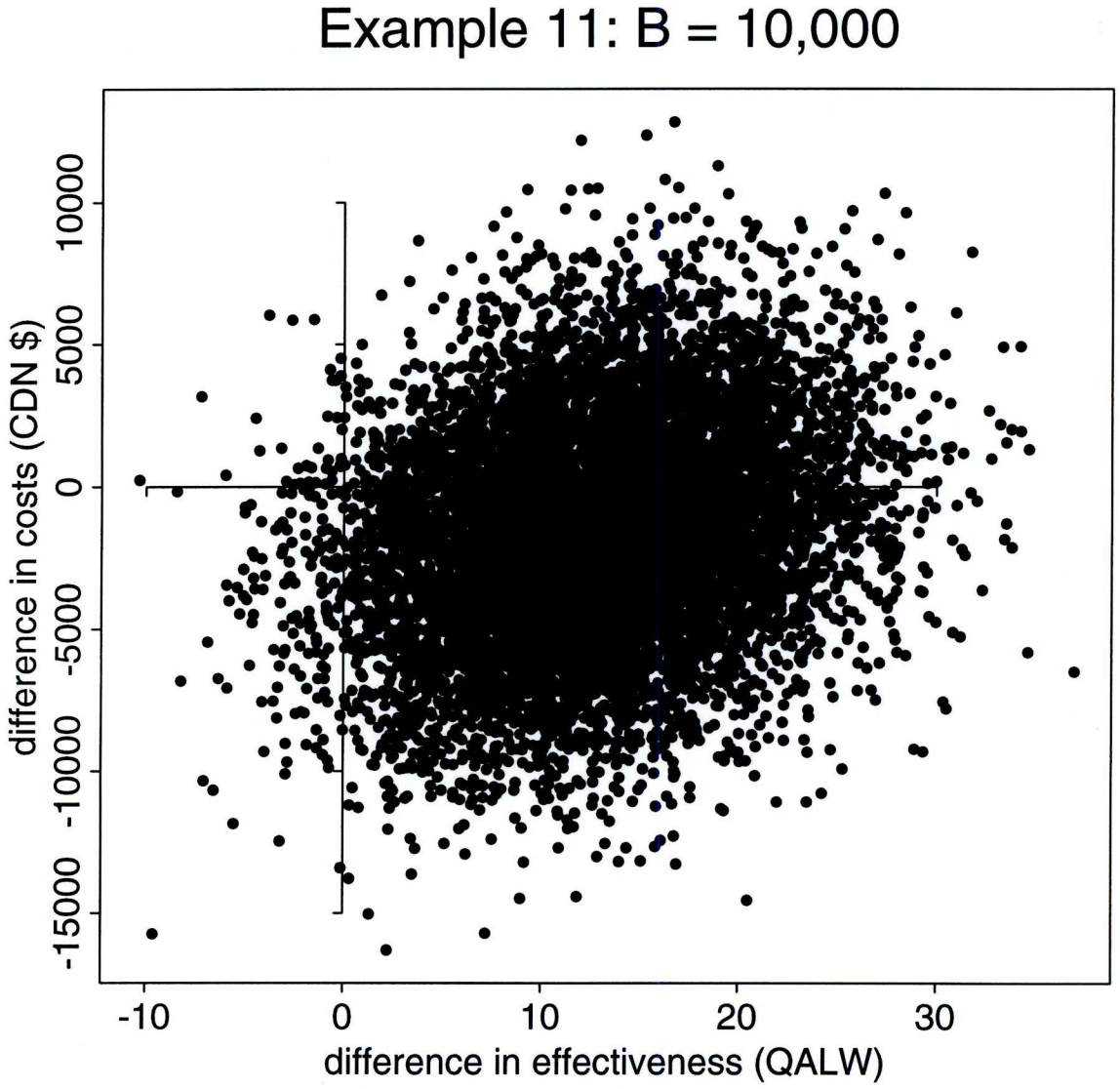

Example 6: B = 500
difference in costs (CDN $)
-15000 -5000 0 5000 10000
0 10 20 30
difference in effectiveness (QALW)

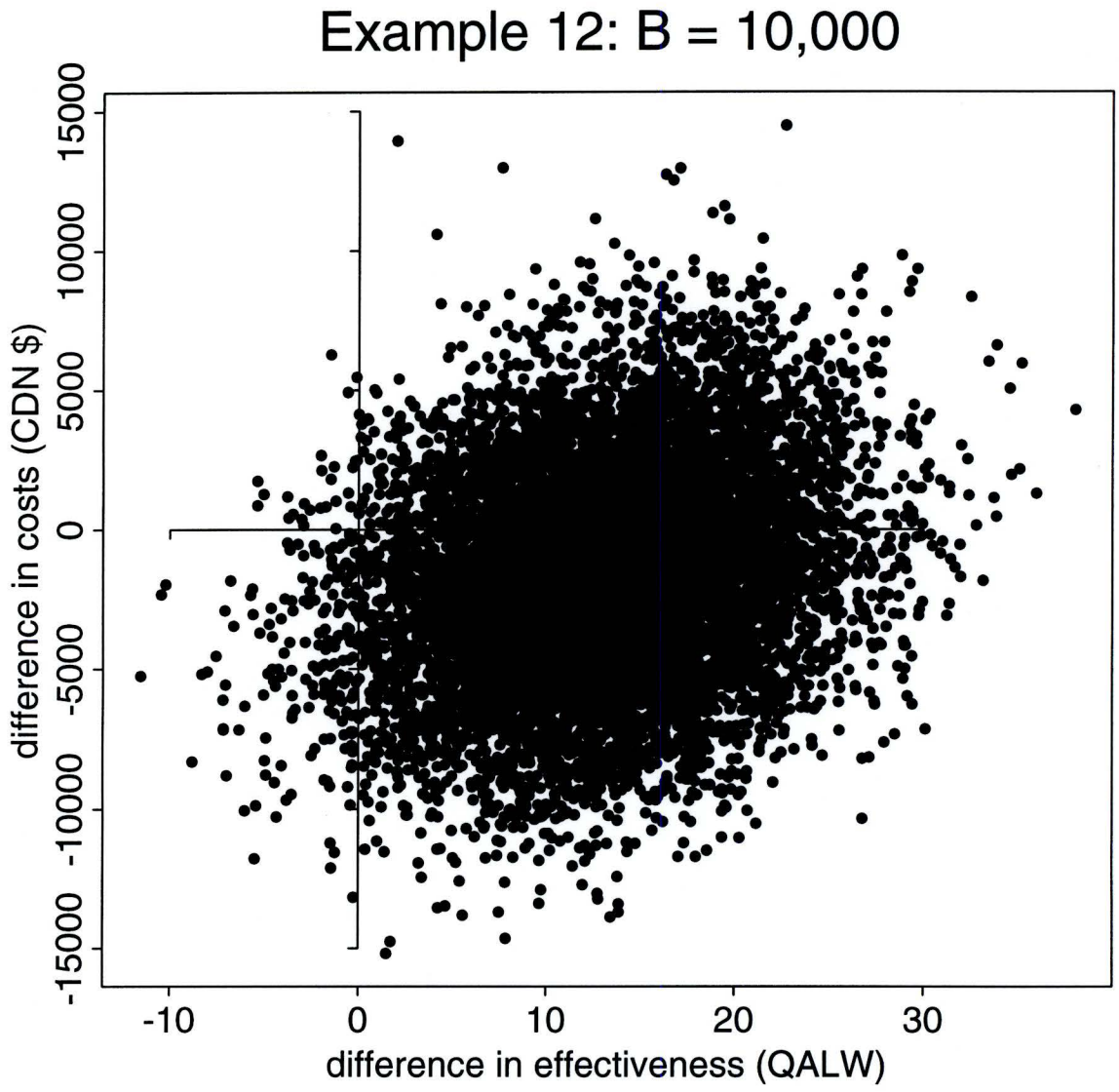

Example 7: B = 1000
difference in costs (CDN $)
10000
5000
0
-5000
-10000
0
10
20
30
difference in effectiveness (QALW)


Example 8: B = 1000
difference in costs (CDN $)
10000
5000
0
-5000
-10000
-15000
0
10
20
30
40
difference in effectiveness (QALW)


Example 9: B = 1000
difference in costs (CDN $)
10000
5000
0
-5000
-10000
-10
0
10
20
30
difference in effectiveness (QALW)


Example 10: B = 10,000
difference in costs (CDN $)
10000
5000
0
-5000
-10000
-15000
-10
0
10
20
30
difference in effectiveness (QALW)


Example 11: B = 10,000
difference in costs (CDN $)
10000
5000
0
-5000
-10000
-15000
-10
0
10
20
30
difference in effectiveness (QALW)


Example 12: B = 10,000
difference in costs (CDN $)
15000
10000
5000
0
-5000
-10000
-15000
-10
0
10
20
30
difference in effectiveness (QALW)