Revisiting methods for calculating confidence region for ICERs ? Are Fieller's and bootstrap methods really equivalent ?

Carole SIANI[(a,b,*)], Christian de PERETTI[(a)] and Jean-Paul MOATTI[(b)]

(a) GREQAM, 2, rue de la Charité, 13002 Marseille, France.
(b) Unité INSERM 379, 232 boulevard Sainte Margueritte, 13009 Marseille, France.
(*) Correspondence to: Carole Siani, INSERM U379, 232 boulevard Sainte Margueritte, 13009 Marseille, France,
Tel.: +33 (0)4 91 22 35 02; Fax: +33 (0)4 91 22 35 04,
E-Mail: siani@marseille.inserm.fr.

SUMMARY

Previous Monte-Carlo studies of the Literature which compared the performances of Fieller's and bootstrap methods, concluded that they had similar performance for calculating confidence regions for the incremental cost-effectiveness ratio. However, in these studies, the number of simulations used was insufficient to provide a great accuracy, the data dealt with were configured in such a way that the difference between average effects of the two treatments or the (mean costs difference, mean effects difference) pair were highly significant and the miscoverage was measured calculating the mean miscoverage on several Data Generating Process and any conclusion can be drawn from these results. In this paper, we focus on the performance of Fieller's and bootstrap methods in the problematic cases, frequently occurring in practice, of the difference between average effects of the two treatments approaching statistically zero or of the (mean costs difference, mean effects difference) pair also approaching statistically zero using Monte-Carlo simulations. These simulations permit to quantify the difference of performance between the methods, they show that the non reordered bootstrap method perform worse than Fieller's method in these problematic cases, and that it should be better to use "reordered" bootstrap methods. In addition, they confirm that the re-ordered bootstrap method and Fieller's method have similar performance most of the time. Nevertheless, our Monte Carlo experiments show that Fieller's method performs significantly better than re-ordered bootstrap method in case of (mean costs difference, mean effects difference) pair approaching statistically zero. Consequently, since Fieller's method seems to be the best method, we study it in detail and we prove some theorems that permit to show that Fieller's method is mathematically applicable in all the situations and is always usable for decision-making, even in the problematic cases.

Key-words: Uncertainty, incremental cost-effectiveness ratio, confidence regions, Fieller, bootstrap.

1 INTRODUCTION

In recent years, stochastic data on both costs and effectiveness of alternative medical strategies have been simultaneously available at the level of individual patients, for example through collection of costs data alongside clinical trials [1, 2]. Such data give the opportunity to summarise uncertainty associated with the results of a cost-effectiveness (CE) analysis in the form of confidence regions, and there has been a growing body of health economics literature dealing with the methodological problems for calculating such confidence intervals for cost-effectiveness ratios. Various methods for calculating confidence regions for the incremental cost-effectiveness ratios (ICERs) have been explored [3, 4, 5, 6, 7]: they are either based on the estimator of the ICER density (Taylor's method, nonparametric non reordered bootstrap methods) or alternatively on the bivariate density function of the pair composed by the mean costs difference and the mean effects difference (the "box" method, the ellipse method and Fieller's method and reordered bootstrap). Current conclusions of the literature tend to suggest that both reordered bootstrap and Fieller's methods are the most appropriate and it has been argued, on the basis of Monte-Carlo simulations, that both methods obtained similar results [8, 9, 10, 20] except when the mean effects difference and the mean costs difference are unsignificant.

However, as it will be detailed in section 2.2, it could be hypothesised that methods based on the density function of the ICER estimator, such as non reordered bootstrap methods, may become unstable or even mathematically inapplicable in the case of the difference between average effects of the two treatments approaching statistically zero or in the case of the (mean costs difference, mean effects difference) pair also approaching statistically zero. Fieller's and reordered bootstrap methods, based on the bivariate density function of the pair, may not encounter these problems.

In practice, many empirical studies may indeed correspond to such cases because clinical trials are often designed to detect small differences in effectiveness between treatments and medical innovations often imply some deterioration of the CE ratio for a limited improvement in effectiveness compared to standard treatment.

In this paper, Monte-Carlo simulations are used to compare the performances of all bootstrap (non reordered and reordered) and Fieller's methods for calculating confidence regions of ICER in the problematic cases in which differences in clinical effectiveness are close to statistical unsignificance or in which both the differences in clinical effectiveness and the differences in costs between the two health care programs are close to statistical unsignificance. Monte-Carlo simulations will be applied to empirical data issued from a randomised clinical trial (RCT) that corresponds to the first type of situation and to data extrapolated from this same source in order to correspond to the latter type of situation. In both problematic cases, we will show that mathematical limitations of the non reordered bootstrap methods should lead to consider Fieller's method and the reordered bootstrap method, as the methods of choice. Lastly, a second set of experiments is carried out varying the skewness and kurtosis coefficients, using the gamma distribution, and for different sample sizes, so as to compare the performance of the reordered percentile method and Fieller's method in extreme cases, and we will show that Fieller's method performs better in these situations.

Since Fieller's method seems to be the best method, we study it in detail and we prove some theorems that permit to show that Fieller's method is mathematically applicable in all the situations and is always usable for decision-making, even in the problematic cases.

2 METHODS

In this section, after the definition of the ICER, we present the various methods used for calculating confidence regions for the ICER: non reordered and reordered bootstrap methods as well as Fieller's methods. Then, we present the empirical data that were used for processing Monte Carlo simulations in order to compare the performances of the various methods. These data illustrate a typical case in which clinical differences are only close to statistical significance. These data were also translated for illustrating the case where costs differences are close to statistical unsignificance in addition to the small clinical differences between the two treatments already present in the data. In the last subsection, methodology of these simulations is detailed.

2.1 Definition of the ICER

In cost-effectiveness analysis, one (or more) new treatments (T_1) are compared to (one or more) standard treatments (T_0) on the two-fold basis of the cost and the medical effects of each treatment. In this context, the appropriate summary measure of cost-effectiveness is the ICER which is the ratio of the mean costs difference and the mean effects difference between both the treatments. The ICER can be estimated as follows, on the basis of data collected from the two groups of patients:

$$\hat{R} = \frac{\overline{C_1} - \overline{C_0}}{\overline{E_1} - \overline{E_0}} = \frac{\Delta\bar{C}}{\Delta\bar{E}} \ ,$$

where $\overline{C_1}$, $\overline{C_0}$ are the sample mean of the costs and $\overline{E_1}$, $\overline{E_0}$ in the two treatments arms are the sample mean of effects.

2.2 Nonparametric bootstrap methods

Generally speaking, bootstrap methods have been particularly prized because the bootstrap law constitutes a better approximation of the law of the statistic of interest than the asymptotic law [11].

Nonparametric bootstrap method has the substantial advantage of making no parametric assumptions about the sampling distribution of costs and effects. This method consists in building up an empirical estimate of the sampling distribution of the ICER estimator, by resampling from the original data in the following way:

1. Sample with replacement n_1 (cost, effect) pairs and n_0 pairs respectively, from the sample of patients who underwent treatment (T_1) and treatment (T_0) respectively. It should be noted that we make a drawn from the (cost, effect) pair for each treatment group as to preserve the correlation between costs and effects.

2. Calculate $\bar{C}_1^*$, $\bar{E}_1^*$, $\bar{C}_0^*$ and $\bar{E}_0^*$ the bootstrap simulations of $\bar{C}_1$ and $\bar{E}_1$, $\bar{C}_0$ and $\bar{E}_0$ respectively.

3. Calculate the bootstrap replicate R_b^* of the ICER given by the equation:

$$R_b^* = \frac{\overline{C_1^*} - \overline{C_0^*}}{\overline{E_1^*} - \overline{E_0^*}} \ .$$

4. After repeating this three-stage process many times (denoted B), we obtain a vector of bootstrap estimates $(R_1^*, \ldots, R_B^*)$ which is an empirical estimate of the sampling distribution of the ICER estimator.

Once the sampling distribution of the ICER estimator has been estimated, there exist several approaches for estimating the bounds of the confidence interval, such as the percentile method [6, 12], the percentile-t method [6, 12] and the Bias Corrected and Accelerated method [6, 13], which the two latter methods take into account any asymmetry of the distribution. The percentile-t method under consideration involves a double level of bootstrap simulations for estimating the standard error of the ICER.

All these non reordered bootstrap methods have however, a general limitation. They are inapplicable if $\mu_{\Delta E} = 0$ statistically. In that case, the theoretical ratio is not statistically defined (i. e. $R = \pm\infty$) and a confidence interval given numerically of the form $[R^L, R^U]$ has no mathematical sense; the findings of the cost-effectiveness analysis are therefore meaningless. In addition, some of these methods such as the percentile-t method require estimating the variance of the ratio, which is an additional cause of instability when $\mu_{\Delta E}$ approaches statistically zero. Finally, bootstrap methods have the disadvantage of excluding confidence regions having the form $]-\infty, R^L] \cup [R^U, +\infty[$. These restrictions tend to make non reordered bootstrap methods quite inappropriate when the mean effects difference approaches statistically zero or when the pair composed by the mean costs difference and the mean effects difference approaches statistically $(0, 0)$.

To solve these problems, some authors [6, 18, 20] have suggested to use the "reordered" bootstrap method, in particular, the "reordered" percentile method. This method consist of applying the previous four steps of the nonparametric non reordered bootstrap method and in a last step to re-order the negative ICER resulting from negative effects at the top of the ordered list if ICERs instead of at the bottom. Indeed, many authors reproached to bootstrap methods the misplacement of negative values of the ICER of QIV at the left of bootstrap distribution, if we number the quadrants of the CE plane from I (the upper right) counterclockwise to IV (the lower right), this artificially reduces both upper and lower bootstrap confidence intervals for the ICER thereby invalidating coverage probability.

In this paper, we have tested nonparametric non reordered bootstrap methods, associated with the percentile method, the percentile-t method and the BCA method and the reordered percentile method.

2.3 Method based on Fieller's theorem

2.3.1 General theory

This analytic method, is based on the joint distribution function of the (mean costs difference, mean effects difference) pair which is assumed to follow a bivariate Gaussian distribution. This method involves calculating confidence regions using the pivotal function technique, which consists in resolving a second degree equation in the ICER. We briefly recall the general context of Fieller's theorem [3]. It is assumed here that X_1 and X_2 are two random normally distributed variables, such that:

$$X \sim N(\eta, \Omega) \text{ with } X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix},$$

and it is proposed to determine a $(1-\alpha)$ confidence region for $\frac{\eta_1}{\eta_2}$.

To find the $(1-\alpha)$ confidence region for $\frac{\eta_1}{\eta_2}$, the following inequation must be solved:

$$Q(\rho) \leq 0. \tag{1}$$

where

$$Q(\rho) = x\rho^2 + y\rho + z \ ,$$

with $x = X_2^2 - k_{1-\alpha}\omega_2^2$, $y = 2(k_{1-\alpha}\omega_{12} - X_1X_2)$ and $z = X_1^2 - k_{1-\alpha}\omega_1^2$. The roots of the polynomial function Q (denoted R^L and R^U) are given by the following formulae:

$$R^L = \frac{X_1X_2 - k_{1-\alpha}\omega_{12} - \sqrt{(k_{1-\alpha}\omega_{12} - X_1X_2)^2 - (X_2^2 - k_{1-\alpha}\omega_2^2)(X_1^2 - k_{1-\alpha}\omega_1^2)}}{X_2^2 - k_{1-\alpha}\omega_2^2} \ ,$$

$$R^U = \frac{X_1X_2 - k_{1-\alpha}\omega_{12} + \sqrt{(k_{1-\alpha}\omega_{12} - X_1X_2)^2 - (X_2^2 - k_{1-\alpha}\omega_2^2)(X_1^2 - k_{1-\alpha}\omega_1^2)}}{X_2^2 - k_{1-\alpha}\omega_2^2}.$$

If the variances and covariances are unknown, they can be replaced by their estimators, in which case $k_{1-\alpha}$ is interpreted as the $(1-\alpha)$ quantile of a Fisher distribution with the appropriate degree of freedom.

2.3.2 Application to the ICER

We assume that (C_j, E_j) is a random vector with mean (μ_{C_j}, μ_{E_j}), variance $(\sigma^2_{C_j}, \sigma^2_{E_j})$ and correlation λ_i for j=0 and 1.
The variables used in Fieller's method correspond to the following values:

$$\begin{aligned}
X_1 &= \Delta\bar{C}, \\
X_2 &= \Delta\bar{E}, \\
\omega_1^2 &= \sigma^2_{C_0}/n_0 + \sigma^2_{C_1}/n_1, \\
\omega_2^2 &= \sigma^2_{E_0}/n_0 + \sigma^2_{E_1}/n_1, \\
\omega_{12} &= \lambda_1\sigma_{C_1}\sigma_{E_1}/n_1 + \lambda_0\sigma_{C_0}\sigma_{E_0}/n_0.
\end{aligned}$$

The $(1-\alpha)$ confidence regions for the ICER can have different forms (see table 1). This requires a detailed study and the demonstration of several theorems that will be made in the next section.

2.3.3 The various forms of the confidence regions

Depending on the sign of the coefficient before the second degree term of the polynomial function (denoted x) indicating its concavity and depending on the sign of the discriminator of the polynomial function (denoted Δ), the various forms of the confidence regions obtained with Fieller's method are shown in Table 1. where R^L, R^U denote the roots of the polynomial

	$\Delta < 0$	$\Delta = 0$	$\Delta > 0$
$x > 0$ convex function	impossible case	impossible case	$\left[R^L, R^U\right]$
$x = 0$ linear function	impossible case	$\mathbb{R}$	$\left]-\infty, R^U\right]$ if $y > 0$ $\left[R^L, +\infty\right[$ if $y < 0$
$x < 0$ concave function	$\mathbb{R}$	$\mathbb{R}$	$]-\infty, R^U] \cup [R^L, +\infty[$

Table 1: Form of the confidence region

function Q. In case where $x > 0$, we have $R^L < R^U$, otherwise if $x < 0$, we have $R^L > R^U$. Lastly, if $x = 0$, then $R^L = R^U$.

It should be noted that the condition $x < 0$ corresponds to the case where $\mu_{\Delta E}$ is not significantly different from zero at level α, and geometrically, the $(\Delta\bar{C}, \Delta\bar{E})$ pair is close to the vertical axis. The sign of x determines the statistical distance between the mean effects difference and zero. As regards the sign of Δ, it measures the statistical distance between the (mean costs difference, mean effects difference) and the origin of the CE plane (see theorem 2).

The table 2.3.3 was studied in details and each form of confidence region was rigorously interpreted and proved in [21], in case of the difference between average effects of the two treatments approaching statistically zero or in case of the (mean costs difference, mean effects difference) pair also approaching statistically zero (for the interpretations, see appendix 2.3.4 and for the proofs see see appendix A).

Thus, from the theoretical point of view, this table makes it possible to conclude Fieller's method does not a priori suffer from the restrictions previously underlined with non re-ordered bootstrap methods: Fieller's method is applicable all the time without no condition and the confidence region can have the form of the complement of an interval. Only the normality hypothesis of the $(\Delta\bar{C}, \Delta\bar{E})$ pair could raise problem but this will be tested below through Monte-Carlo simulations (see paragraph 3.1).

2.3.4 Interpretation of the various forms of confidence regions obtained with Fieller's method

The "impossible" cases

The cases denoted "impossible" in Table 2.3.3 correspond to cases that cannot occur. They are explained by the following theorems 1 and 1 bis (for the proof see appendix A). These theorems permit to analyse this case that was considered by Heitjan et al. [19] but that was not commented in their paper.

Theorem 1

$$x > 0 \Rightarrow \Delta > 0$$

Theorem 1 bis

$$x \geq 0 \Rightarrow \Delta \geq 0.$$

I.e. when the polynomial function is convex (or linear), the discriminator is always positive (or null), and the $(1 - \alpha)$ confidence region takes the form of an interval such as $[R^L, R^U]$.

The complement of an interval

In the case of $\Delta > 0$ and $x < 0$, the confidence region has the form of the complement of an interval, that is, the form $]-\infty, R^U] \cup [R^L, +\infty[$. To understand the intuition of this form of confidence region for the ICER, it can be interpreted as the complement of the confidence interval obtained calculating the confidence region for the inverse of the ICER (*i.e.* the ratio between the mean effects differences and the mean costs differences). We have shown that there is no mathematical problems about it (see Table 2.3.3). However, the form of this kind of confidence region can seem problematic for decision-making purposes, because it extends into more than one quadrant and it contains infinite values, and it has not been dealt with in the literature. In fact, it is sufficient to proceed in exactly the same way as for confidence intervals extending into only one quadrant. The confidence region for the ICER corresponds to an angular sector on the CE plane. If this angular sector is located to the left (respectively to the right) of the straight line associated with the ceiling ratio corresponding to some maximum value of the ICER that society is prepared to pay to

achieve the additional effectiveness, the new therapy is dominated (respectively dominant). Lastly, when the ceiling ratio belongs to the confidence region the two treatments are not significantly different on the basis of the ICER.

In this context, even if negative or infinite values belong to the angular sector (*i.e.* this angular sector contains the Y-axis), we clearly see that there are any problems at the decision-making level.

The whole real line

In the case of $\Delta < 0$ and $x < 0$, some authors have pointed out that "Fieller's method sometimes produces imaginary results" because R^L and R^U were calculated with the formula 2 and 2. Other authors have arged that "no solution exists" (among others, see Willan and O'Brien [14]). In fact, in this case, R^L and R^U must not be directly calculated by the formula 2 and 2 but rather by resolving the inequation 1 and the confidence region obtained is the whole real line, since Q is concave.

It can seems surprising to obtain $\mathbb{R}$ as a $(1-\alpha)$ confidence region, but theoretically, there is absolutely no contradiction with the definition of a confidence region: $\mathbb{R}$ is a possible realisation of a $(1-\alpha)$ confidence region, (that is a random region that contains the ICER $(1-\alpha)\times 100$ times over 100). But in practice, how to interpret this result ? Theorem 2 gives equivalent conditions that permit to answer this question (for the proof see appendix A.2).

Theorem 2

Denoting $X = \begin{pmatrix} \Delta\bar{C} \\ \Delta\bar{E} \end{pmatrix}$, and $\Omega = \mathrm{Var}(X)$, the following statements are equivalent:

1. $\Delta > 0 \Leftrightarrow \|\Gamma X\|_2^2 > k_{1-\alpha}$ with $\Gamma = \Omega^{-1/2}$, $\|.\|_2$ indicates the Euclidean norm and $k_{1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom and also corresponds to the $(1-\alpha')$ quantile of the chi-squared distribution with two degrees of freedom (with $\alpha' > \alpha$ [1]).
2. $\Delta > 0 \Leftrightarrow$ to reject the hypothesis $(H_0) : (\mu_{\Delta C}, \mu_{\Delta E}) = (0,0)$ for a test of size $\alpha' > \alpha$.
3. $\Delta > 0 \Leftrightarrow X$ is located outside the ellipse defined by the following equation: $\|\Gamma X\|_2^2 = k_{1-\alpha}$.

Theorem 2 indicates that when the discriminator is negative or null, the ratio is poorly defined: the direction of the sector is not statistically defined (the ratio is close to the form "$\frac{0}{0}$") and the confidence region obtained is the whole real line. This does not mean that we have no information about the ratio; on the contrary, we have the following precise piece of information about it: the ratio is not statistically defined because, either both the treatments are equivalent on the two-fold basis of the cost and the medical effects of each treatment, or the sample size is not sufficiently large to distinguish between them. In this latter case, two other possibilities are available: we can either work on larger groups of patients, or we can calculate a confidence interval with a weaker confidence level.

One minus the confidence level (*i.e.* $1-\alpha$) is the probability that the true value for the ICER does not belong to the confidence region: it is the probability of making an error when giving the possible values for the ICER. Thus, the confidence level can be interpreted as a risk of error when decision-making is made. This level has to be chosen *a priori*, independently

[1] because $\|\Gamma X\|_2^2 \sim \chi^2(2)$. For example, for $\alpha = 0.05$, we have $\alpha' = 0.15$.

of the data, before estimating the ICER. Thus, the second solution proposed above causes a problem: decreasing the confidence level, depending on the form of the confidence region, makes α depend on the data (contrary to what should be done), and in addition increases the risk of error, leading to a risk level that may not be acceptable for the decision-maker. In our sense, it is preferable to consider that the treatments are indistinguishable (on the basis of the data), by choosing the whole real line as a confidence region, rather than taking a decision with a high risk of misusing the results of the analysis.

In any event, finding the whole real line is informative. Fieller's method has the great advantage of detecting the case in which the treatments are not distinguishable on the two-fold basis of the cost and the medical effects of each treatment, contrary to bootstrap methods (non reordered and reordered), which provide numerical results all the time, but yield results that may have no mathematical sense: theoretically, when $(\mu_{\Delta C}, \mu_{\Delta E}) = (0, 0)$, the bootstrap method is not consistent, [2], and in practice, when $(\mu_{\Delta C}, \mu_{\Delta E})$ is close to zero $(0, 0)$, the results will be very biased and random [3].

Nevertheless, in order that Fieller's method should provide a region different from the whole real line, it is sufficient to check that $(H_0) : (\mu_{\Delta C}, \mu_{\Delta E}) = (0, 0)$ is rejected at level α' (see statement 2) and this condition is less restrictive than a classical test of size α (since α' is greater than α).

2.4 Empirical data used in Monte-Carlo simulations

We present a preliminary study of the empirical data issued from a randomised clinical trial that will be used in our Monte Carlo simulations (see Section 2.5). These data illustrate a typical case, in which clinical differences are only close to statistical significance. In this trial (multi-centric trial Pegase 01 initiated by the National French Federation of Anti Cancer Regional Centers), high dose chemotherapy supported by recombinant hematopoietic growth factors and blood stem cell transplantation was compared with a conventional chemotherapy control group in the context of breast cancer for high risk patients (with an axillary invasion of eight or more lymph nodes). The main variable for measuring effectiveness was length of survival without relapse during the follow-up period (equal to three years here). In this trial, direct medical costs were measured for each patient on the basis of physical units for each cost component weighted by the unit prices of these resources expressed in 2000 French Francs (FF). The descriptive statistics are summarised in table 2. In our example, the ICER of the new treatment obtained from these data is equal to 21967 FF per month gained without relapse.

After noting that the coefficient of variation summarises the relative proximity of an estimate to zero, it can be seen from table 2 that the coefficient of variation of the difference between mean effects was almost equal to 0.5, this means that the difference between mean effects are not significant. This exactly correspond to one of the two problematic cases for estimating the ICER that we intended to study.

In view of the Skewness and the Kurtosis (see Table 3), it can also be noted that the costs data are skewed and leptokurtic, which suggests that these data are not normally distributed. To check this point, we tested whether the Skewness was equal to zero and

[2] When $(\mu_{\Delta C}, \mu_{\Delta E}) = (0, 0)$, the estimated ICER follows a Cauchy distribution (for all sample sizes). In this case, the bootstrap confidence region, based on the estimated ICER distribution, will be close to zero (the location of the Cauchy distribution hump) whereas the true ICER value is indefinite.

[3] When $(\mu_{\Delta C}, \mu_{\Delta E})$ is not significantly different from $(0, 0)$, the estimated ICER distribution is close to a Cauchy distribution. In this case, the bootstrap confidence region tends to be close to zero whatever the true value for the ICER.

Group variable	Sample mean	s.e.	c.v.	c.c.	Skewness	Kurtosis
Treatment: n=145				0.06		
Cost (FF)	117077.67	22070.43	0.19		1.57	7.95
Effect (months of life gained without relapse)	33.48	14.5	0.43		-0.03	2.40
Control: n=155				-0.14		
Cost (FF)	34206.37	17461.15	0.51		6.80	65.90
Effect (months of life gained without relapse)	29.71	15.3	0.52		0.51	2.38
Difference				-0.03		
Cost (FF)	82871.30	2307.90	0.03		0.19	0.08
Effect (months of life gained without relapse)	3.77	1.7	0.46		0.01	0.01

s.e. denotes the standard error, c.c. corresponds to an estimator of correlation coefficient between costs and effects for each treatment and between mean costs difference and mean effects difference, c.v. denotes the coefficient of variation of the costs, of the effects for each treatment, of the mean costs difference and of the mean effects difference, Skewness and Kurtosis denote estimators of the Skewness and Kurtosis coefficients.

Table 2: Descriptive statistics of the clinical trial

whether the kurtosis was equal to 3 and performed a Jarque-Bera test [16] with both hypothesises combined. The bootstrap version of these tests were used. They show that all the data are not normally distributed except for the effects of the treatment group for which the p-value was equal to = 0.263 in the Jarque-Bera test (the costs data of the two treatments groups provides a p-value equal to zero in the three tests and the effects of the control group gives a p-value equal to 0.015 in the Jarque-Bera test). Thus, these data will led us to test the impact of their non normality on the performances of the methods, in particular with Fieller's method which requires this normality hypothesis.

2.5 Methodology of the Monte-Carlo simulations

Monte-Carlo simulations were carried out to assess the performance of Fieller's and bootstrap methods (non reordered and reordered) for calculating confidence region for the ICER with a nominal confidence level equal to 0.95, principally when applied in problematic situation in which the mean effects difference approaches statistically zero or when the pair composed by the mean costs difference and the mean effects difference also approaches statistically $(0, 0)$,

A first experiment focus on the impact of the distance between the $(\Delta\bar{C}, \Delta\bar{E})$ pair and the origin of the CE plane, on the performance of the various methods. A second set of experiments is carried out for different sample sizes, varying the skewness and kurtosis coefficients using a gamma distributions.

2.5.1 First experiment: different statistical location of the $(\Delta\bar{C}, \Delta\bar{E})$ pair

This first set of experiments is carried out for different statistical locations of the $(\Delta\bar{C}, \Delta\bar{E})$ pair more or less close to $(0, 0)$, for the same sample sizes as the real data, that is, 145 for

the treatment group and 155 for the control group, and using a non-Gaussian distribution inspired by the real data issued from a RCT.

Three possible locations are considered: in one case, the $(\Delta\bar{C}, \Delta\bar{E})$ is far from the origin of the CE plane (denoted case 1), in another case corresponding to the problematic case in which differences in clinical effectiveness are close to unsignificance, the pair is close to the costs-axis (denoted case 2) and in the last case corresponding to the problematic case in which both the differences in clinical effectiveness and the differences in costs between the two treatments are close to unsignificance, the pair is close to the origin of the CE plane (denoted case 3). It should be noted that case 1, which is not problematic, is given as an illustration to check that all methods perform well in that case as it has been shown in the literature. The distance between the $(\Delta\bar{C}, \Delta\bar{E})$ pair and the origin of the CE plane is determined from the coefficient of variation of $\Delta\bar{C}$ and that of $\Delta\bar{E}$, denoted $cv(\Delta\bar{C})$ and $cv(\Delta\bar{E})$ respectively (for the values, see table 3).

The real data already corresponds to the case denoted 2 and in this case, the distribution used is the empirical distribution obtained by resampling from the real data (that is, the uniform law applied to the (cost,effect) pair of the data). To obtain cases 1 and 3 respectively, we transform the real data so that the coefficient of variation of the difference between mean effects become equal to 0.05 and so that the coefficient of variation of the difference between mean costs became equal to 0.46 respectively and the empirical distribution used is obtained by resampling from modified data. The data are translated so that the standard errors of the modified data is identical to that of the original data as follows: $E'_1 = E_1 + 34.2$ for obtaining case 1 ($C'_1 = C_1 - 77853.6$ for obtaining case 3 respectively) and other data remain unchanged, where E'_1 (C'_1 respectively) denotes the modified effect (cost respectively) data for (T_1) . Table 3 sums up the various cases studied.

	Location of the $(\Delta\bar{C}, \Delta\bar{E})$ pair	$cv(\Delta\bar{C})$	$cv(\Delta\bar{E})$
Case 1	far from the origin of the CE plane	3%	5%
Case 2	close to the costs-axis of the CE plane	3%	46%
Case 3	close to the origin of the CE plane	46%	46%

Table 3: The various cases studied in Monte-Carlo simulations with an empirical distribution

We have carried out $B = 999$ bootstrap replications and $R = 10,000$ Monte-Carlo simulations, except for the percentile-t method, with which we performed only $1,000$ Monte-Carlo simulations because this method is very costly in computing time and as we will see, the results were too unsatisfactory for it to be worth attempting to achieve greater accuracy. We keep the same random numbers sequence for the Monte-Carlo experiments as for the bootstrap resampling procedure so as to reduce the experimental errors.

2.5.2 Second experiment: high skewness and kurtosis values

A second set of experiments is carried out varying the skewness and kurtosis coefficients using the gamma distribution, and for different sample sizes, so as to compare the performance of the reordered percentile method and Fieller's method in extreme cases. Both variables, costs and effects, follow a gamma distribution with five different choices of the parameters (α, β, γ). The first case is chosen so that the first three moments of the gamma distribution are equal to those of the real variables (using the formula linking α, β and γ with the moments of the variable following a gamma distribution), the successive (α, β, γ) triples are obtained by dividing the previous α by four, and recalculating β and γ so that the means

and the standard deviations of the distributions remain unchanged. Consequently, higher moments (such as skewness and kurtosis) will change. We performed 10,000 Monte-Carlo experiments. For each experiment (*i.e.* each value for (α, β, γ)), we consider four sample sizes $T_1 = T_0 = T \in \{150, 75, 37, 19\}$, the sample size of 150 corresponding roughly to the sample size of the real data. The various parameters of the distribution of costs and effects of each treatment are summarised in the table 4. Likewise, corresponding estimates of the skewness and kurtosis coefficients of costs and effects are summarised in table 5, as well as estimates of the skewness and kurtosis coefficients of $\Delta\bar{C}$ for $T = 150$ and $T = 19$ respectively.

DGP	$\alpha(C_1)$	$\beta(C_1)$	$\gamma(C_1)$	$\alpha(C_0)$	$\beta(C_0)$	$\gamma(C_0)$
1	4444	0.22	-944	15.38	3.90	-30.28
2	1111	0.44	-455	3.85	7.80	-0.29
3	278	0.88	-211	0.96	15.60	14.70
4	69	1.76	-89	0.24	31.20	22.20
5	17	3.52	-28	0.06	62.40	25.95

With the DGP 3, the parameters of the distribution of C_1 and C_0 are not indicated in the table since they are the same as for the DGP 1. In the column "DGP", the first number corresponds to the number of the model and the second number corresponds to the DGP under consideration.

Table 4: The parameters of the variables for the various DGPs using gamma distributions

DGP	$S(C_1)$	$S(C_0)$	$S(\Delta\bar{C})$ $T = 150$	$S(\Delta\bar{C})$ $T = 19$	$K(C_1)$	$K(C_0)$	$K(\Delta\bar{C})$ $T = 150$	$K(\Delta\bar{C})$ $T = 19$
1	0.03	0.51	0.02	0.05	3.00	3.39	0.01	0.08
2	0.06	1.02	0.03	0.09	3.01	4.56	0.01	0.10
3	0.12	2.04	0.07	0.19	3.02	9.24	0.02	0.17
4	0.24	4.08	0.13	0.37	3.09	27.97	0.06	0.44
5	0.48	8.16	0.26	0.74	3.35	102.87	0.19	1.51

Table 5: Skewness and Kurtosis coefficients of costs and effects for the various DGPs using gamma distributions

2.5.3 Criteria for assessing the performances of the various methods

The first criterion used to assess the performances of the seven computation methods studied was the overall probability of coverage (i.e. the percentage of samples where the true (mean costs difference, mean effects difference) pair fell inside the estimated confidence region) with its standard error. The procedure with the best confidence region was the one that came closest to the nominal coverage level of 0.95. The average length of the confidence across the simulated samples region with its standard error as well as the average angle of the confidence sector with its standard error were also used as criteria for evaluating the methods. Generally, the length of a confidence interval is a satisfactory criterion, since the narrower this interval is, the more efficient the method will be due to some likelihood considerations. In the case of Fieller's method, the length of the confidence region is no longer a satisfactory criterion since in case of a confidence region having the form a the complement of an interval, this length will be infinite even if the region is optimal with

respect to the likelihood of the $(\mu_{\Delta C}, \mu_{\Delta E})$ pair. In this latter case, we rather consider the average angle of the confidence sector with its standard error for avoiding this problem: the smallest the angle is, the more efficient the method will be.

3 RESULTS OF MONTE-CARLO SIMULATIONS

3.1 First experiment

With the location case 1, we observed that all the methods performed well. Consequently, we only report in tables 4 and 5 the performances criteria in location cases 2 and 3 respectively. The same notations as for paragraph 2.2 are preserved.

With case 2, all the methods have relatively good performances except for the percentile-t

Method	Coverage	Length ($\times 10^3$FF)	Angle (^{o})
Fieller	0.948(0.002)	∞	0.00467 (0.00021)
reordered Percentile	0.948(0.002)	∞	0.00465 (0.00020)
Percentile	0.973(0.002)	326054.57(428734.06)	72.87(88.36)
Percentile-t	0.743(0.007)	4670597.16(20931815.94)	172.98(34.84)
BCA	0.920(0.002)	∞	42.81(76.63)

The values in parenthesis represent the standard error of the estimators.

Table 6: Performances of the methods on location case 2 using nonparametric distribution

Method	Coverage	Length ($\times 10^3$FF)	Angle (^{o})
Fieller	0.952(0.002)	∞	63.26(85.40)
reordered Percentile	0.971(0.002)	∞	50.99(61.36)
Percentile	0.982(0.002)	19878.46(29634.71)	119.14(84.76)
Percentile-t	0.887(0.007)	234050.73(917526.56)	178.57(15.50)
BCA	0.920(0.002)	∞	103.41(88.61)

The values in parenthesis represent the standard error of the estimators.

Table 7: Performances of the methods on location case 3 using nonparametric distribution

method which become very unstable (see table 4).

The poor performances of the percentile-t method (the coverage is smaller than 0.80 and the average angle is around equal to 170^o) is due to the the fact of "studentizing" the estimated ICER statistic that makes it unstable and farther from a pivotal statistic than the estimated ICER.

As regards the non reordered percentile method, the theoretical study might have led us to expect that this method would give poor results. This method is usually criticised because it does not take the estimator bias into account. Among other authors, Briggs et al. [6] have pointed out that: "the ICER estimator is a biased estimator and the percentile method of interval estimation does not adjust for bias". But usually, when we talk about bias, we understand that it is a bias due to a translation of the estimator distribution which changes the confidence interval, with respect to the true value of the parameter and this parameter can no longer belong to this interval. Besides, in our case, the bias is also due to a distortion of the distribution, which does not cause the confidence interval to shift with

respect to the true value of the parameter. This explains why this method gives satisfactory results in our case with a coverage equal to 0.973 and an angle equal to 72.87^o.

As regards Fieller's and re-ordered percentile methods, they have similar performance and they perform quite perfectly with a coverage equal to 0.950 and with an almost null angle.

With case 3, in addition to what has been said previously, the existence of the variance of $\hat{R}$ used in the percentile-t method to "studentize" the statistic is not guaranteed and this explains why this method gave poor results in terms of the three criteria (see table 5). As in case 2, we observed that this method performed less efficiently than the percentile method. As regards non reordered bootstrap methods, they became more unstable than in case 2. With percentile and BCA methods, we observed that the coverage is almost similar to the location case 2 but the average angle of the confidence sector increased a lot and was greater than 100^o.

Fieller's method performs quite well with a coverage almost equal to 0.95 and with around a twice smallest angle than with non reordered percentile and BCA methods (see table 5). The only problem with Fieller's method could come from the normality hypothesis of the $(\mu_{\Delta C}, \mu_{\Delta E})$ pair especially as data are often strongly non Gaussian in practice. With the sample size of our data, the results showed that there is no problem. Although the behaviour of the density function of this pair with small samples lay out of the scope of this paper, it has been studied with Monte-Carlo simulations. We have shown that even with data strongly non Gaussian and when the sample size is relatively small (from 30), the density function of the $(\mu_{\Delta C}, \mu_{\Delta E})$ pair is close to the normality thanks to the Central Limit Theorem; this involves that Fieller's method can be applied all the same and performs almost perfectly. To conclude, Monte-Carlo simulations have shown that Fieller's method performs well, in all the situations tested and even in the most problematic ones.

The reordered percentile method performs less satisfactorily than Fieller's method. However, the performance is still satisfactory since the case under consideration should be problematic. A detailed Monte-Carlo study based on extreme distributions for the data and small sample size is therefore required and is made in the next section.

3.2 Second experiment

The results for the Fieller's method and the reordered bootstrap method for different sample sizes, varying the skewness and kurtosis coefficients using a gamma distributions are presented in table 8 for a 0.95 confidence level. The results of Fieller's method are convincing, and show that this method is robust in all situations even with strongly skewed and leptokurtic data, and with relatively small sample size (except for very small sample sizes around 20).

In situations that are not too much extreme, the reordered percentile method performs satisfactorily but less than Fieller's method. However, when the skewness increases and/or the sample size decreases, the performance of this method get worse much more quickly than for Fieller's method.

	Reordered percentile method			Fieller's method		
Sample size	DGP	Coverage	Angle (^{o})	DGP	Coverage	Angle (^{o})
150	1	0.948	0.005	1	0.948	0.005
	2	0.948	0.005	2	0.945	0.005
	3	0.947	0.005	3	0.948	0.005
	4	0.941	0.005	4	0.940	0.005
	5	0.935	0.005	5	0.947	0.005
75	1	0.947	0.007	1	0.951	0.007
	2	0.944	0.007	2	0.944	0.007
	3	0.945	0.007	3	0.947	0.007
	4	0.935	0.043	4	0.944	0.025
	5	0.929	0.150	5	0.940	0.097
37	1	0.942	0.009	1	0.944	0.010
	2	0.943	0.009	2	0.944	0.010
	3	0.938	0.117	3	0.949	0.280
	4	0.928	0.405	4	0.941	0.496
	5	0.922	0.189	5	0.939	0.801
19	1	0.933	0.013	1	0.940	0.013
	2	0.932	0.301	2	0.941	0.229
	3	0.927	0.301	3	0.939	0.877
	4	0.922	0.554	4	0.930	1.471
	5	0.915	0.444	5	0.934	1.182

The number of Monte-Carlo simulations being equal to $10,000$, the standard error of the coverage is equal to 0.002.

Table 8: The performance of Fieller's and reordered bootstrap methods, using gamma distributions

4 DISCUSSION

We have focused on the performances of Fieller's and bootstrap methods, which are usually considered as the best methods for calculating confidence regions for the ICER. Other methods such as the ellipse method and Taylor's method have not been dealt with because they have worst performances than bootstrap and Fieller's methods: it is well known that the ellipse method is too approximate and that Taylor's method is restrictive and constraining because of the normality hypothesis only valid in the asymptotic context.

The previous recommendations of the literature suggest that Fieller's and the non reordered bootstrap methods have not equivalent performances from the theoretical point of view, in the case of the difference between average effects of the two treatments approaching statistically zero or in the case of the (mean costs difference, mean effects difference) pair also approaching zero. However, no paper in the literature made Monte Carlo experiments in this situation. This paper permits to confirm and quantify the effect of the discontinuity problem (when the mean effects difference is insignificant) on the methods as classical bootstrap methods. Our Monte-Carlo simulations clearly show that the non reordered bootstrap methods encounter serious problems in these latter' cases and limitations to their use: bootstrap methods become unstable when the mean effects difference approaches zero statistically and the percentile-t method in particular is very unstable and not suitable with ratios.

On the reverse, Fieller's method is quite perfect and stable in a variety of situations even with skewed data and/or in the case of the (mean costs difference, mean effects difference) pair also approaching zero. In addition, the normality hypothesis is not very restrictive for the application of Fieller's method due to the central limit theorem.

Previous Monte-Carlo studies of the Literature [8, 9] which compared the performances of Fieller's method and of reordered percentile methods concluded that they had similar performances when the mean effects difference is insignificant but mean costs difference is significant. Indeed, this seems to have been the case because first, the number of Monte-Carlo simulations used is too small. Second, previous studies mostly dealt with data configured in such a way that the difference between average effects of the two treatments or the (mean costs difference, mean effects difference) pair were highly significant. And third, in the studies of Polsky et al. [8] and of Briggs et al. [9], the Monte-Carlo experiment was performed in different populations defined, among other things, by levels of correlation between costs and effects, distributions of costs, distribution of effects or the coefficients of variation of the mean costs difference and of the mean effects difference. But since the results were obtained by calculating the mean miscoverage across the various populations, problematic cases did not appear and were hidden in the other cases. However, what makes Monte-Carlo simulations particularly useful is precisely to show how the performances of the methods vary depending on the various parameters of the DGP and to detect problematic cases.

As suggested by Glick *et al.* [20], our Monte Carlo experiments confirm that the reordered bootstrap method has quite satisfactory performances as for Fieller's method, when the mean effects difference is unsignificant and the mean costs difference is significant.

A last question remains: what happens when the mean costs difference and the mean effects difference are both insignificant ? Previous papers suggested that Fieller's method provides too wide region to be useful. On the contrary, our theorem 2 shows that this region is informative since it indicates that both the treatments are equivalent. In addition, no Monte Carlo study of the literature studied the performance of the reordered bootstrap method in this situation. Consequently, we have carried out Monte Carlo simulations in this problematic situation to assess and to compare the performance of the reordered bootstrap and Fieller 's method in this situation.

Monte-Carlo simulations clearly show that Fieller's method performs satisfactorily, except in extreme cases for skewness and kurtosis combined with a small sample size whereas the reordered bootstrap method encounters much more coverage distortion than Fieller's method in these latter cases. This problem can be due to the fact that nonparametric methods are not able to estimate accurately distributions when the sample size is small, especially the tails of distributions. A solution should be to use parametric bootstrap methods, however, the problem becomes "which parametric distribution to choose ?". Lastly, a second additional source of coverage distortion should be that, the direction of the ICER is undefined when the mean costs difference and the mean effects difference are both insignificant, and then, the confidence region can be any one, but can never be the whole real line, as for Fieller's method.

In recent years, there have been growing criticisms about the use of confidence intervals to represent uncertainty associated with results from cost-effectiveness studies. For example, Willan and Lin [17] Stated that "the confidence intervals for the ICER i) can include undefined values or ii) may even be completely undefined". Problem i) refers to confidence regions which contain infinite values (i.e. when $\mu_{\Delta E}$ is statistically equal to zero) having the form of the complement of an interval. We have shown that there are no mathematical problems and no problem at the decision-making level if we work with confidence regions of this form, obtained with Fieller's method. Problem ii) may refer to confidence regions

having the form of the whole real line given by Fieller's method (see table 2.3.3). It can seem surprising and meaningless to obtain the whole real line as a $1-\alpha$ confidence region but theoretically, there is absolutely no contradiction with the definition of a confidence region: $\mathbb{R}$ is a possible realisation of a $1-\alpha$ confidence region, that is a realisation of a random region that contain the ICER $(1-\alpha)\times 100$ times over 100. In fact, this kind of confidence region corresponds to the case where $(\mu_{\Delta C}, \mu_{\Delta E})$ cannot be statistically distinguished from $(0,0)$ and the direction of the confidence sector is not statistically defined. This form of confidence region is informative since it indicates that the data are configured in such a way that either both the treatments are equivalent, either the sample size is not sufficiently large for distinguishing between them, on the two-fold basis of the cost and the medical effects of each treatment. In this case, two other possibilities are available: we can either work on larger groups of patients, or we can calculate a confidence interval with a weaker coverage. Anyway, no other method is able to give more information than Fieller's method which has the great advantage of detecting the cases where any answer can be brought from the statistical point of view contrary to bootstrap methods which provide numerical results all the time, even if they have no mathematical sense (i.e. when the ICER is not statistically defined or infinite). To conclude, Fieller's method is applicable all the time whatever the form of the confidence region obtained and the criticisms i) and ii) above are not really valid.

Some problems have been also pointed out with negative ratios Among other authors, Heitjan et al. [18] reproached to bootstrap methods the misplacement of negative values of the ICER of QIV at the left of bootstrap distribution, if we number the quadrants of the CE plane from I (the upper right) counterclockwise to IV (the lower right), this artificially reduces both upper and lower bootstrap confidence intervals for the ICER thereby invalidating coverage probability, and Briggs [10] questioned about how such a confidence interval should be interpreted. This effectively shows what happens when bootstrap methods are applied in the particularly unstable case where the mean effects difference approaches zero statistically. For using bootstrap methods all the same, Heitjan et al [18] has had the original idea to set ICERs of QIV equal to $+\infty$ and to set ICERs of QII equal to zero so that cost-effectiveness states be appropriately ordered. Other authors such as Briggs et al. [6] has suggested to order negative ICER resulting from negative effects at the top of the ordered list if ICERs instead of at the bottom. But, both these methods are artificial because they do not make bootstrap method stable. In fact, it is not the bootstrap method which is a poor method, but it is rather the fact to apply the bootstrap on an unstable distribution, the distribution of the ICER estimator, which raises problems, in particular when the ICER is not statistically defined. Again, our results show that this problem can be avoided with Fieller's method.

A general criticism often made about confidence regions for ICERs, for example by Briggs et al. [7], is that these regions do not directly address the question of whether

a new intervention is cost-effective, in particular when the ceiling ratio belongs to the confidence region. But this is not what they are intended to do, and Fieller's method does not per se solve this problem. However, inference can easily be done with the ICER: acceptance regions or the p-value can be calculated with the same methods as those used for calculating confidence regions. Besides, we must insist on the fact that confidence regions constitute only a descriptive statistical tool which can be used by decision-makers in a first step to quantify uncertainty approximately, rather than using the ICER alone. In a second step, to be able to make a decision, the only way is to use inference. Another objection of the latter authors' is that the nominal level of the confidence region is often fixed equal to 0.05, this assumes the convention that 0.05 significance is the appropriate level, whereas

this level can vary depending on the intervention under consideration, in particular because of the number of patients to treat. In fact, instead of computing only a confidence region with Fieller's method associated with a particular level, we can easily plot the confidence bounds according to various nominal levels and the decision-maker will therefore be able to choose a suitable nominal level ant to have the associated confidence region.

5 CONCLUSION

The Monre Carlo experiments studied the performances of all bootstrap and Fieller's method focus ing on the problematic cases of the difference between average effects of the two treatments approaching statistically zero or of the (mean costs difference, mean effects difference) pair also approaching statistically zero using Monte-Carlo simulations. The simulations show that the non reordered bootstrap method perform worse than Fieller's method in these latter cases, and that it should be better to use "reordered" bootstrap methods. In addition, they confirm that the re-ordered bootstrap method and Fieller's method have similar performance most of the time. Nevertheless, our Monte Carlo experiments show that Fieller's method performs significantly better than re-ordered bootstrap method in case of (mean costs difference, mean effects difference) pair approaching statistically zero and/or strongly skewed/leptokurtic data. Consequently, since Fieller's method seems to be the best method, we study it in detail and we proved some theorems that permit to show that Fieller's method is mathematically applicable in all the situations and is always usable for decision-making, even in the problematic cases. A wider use of Fieller's method in future empirical CE studies may help rehabilitating the use of ICERs as a useful tool to inform decision-making.

References

[1] Drummond MF. Experimental versus observational data in the economic evaluation of pharmaceuticals. *Medical Decision Making: An International Journal Of The Society For Medical Decision Making* 1998; **18(2)**: S12–S18.

[2] Coyle D, Drummond MF. Analyzing differences in the costs of treatment across centers within economic evaluations. *International Journal of Technology Assessment in Health Care* 2001; **17(2)**: 155–163.

[3] Fieller EC. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* 1954; **16**: 175–183.

[4] van Hout BA, Al MJ, Gordon, GS *et al.* Costs, effects and C/E ratios alongside a clinical trial. *Health Economics* 1994; **3**: 309–319.

[5] O'Brien BJ, Drummond MF, Llabelle BJ *et al.* In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* 1994; **32**: 150–63.

[6] Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstrap: a non-parametric approach to confidence interval estimation. *Health Economics* 1997; **6**: 327–340.

[7] Briggs AH, Fenn P. Confidence intervals or surfaces ? Uncertainty on the cost-effectiveness plane. *Health Economics* 1998; **7**: 723–740.

[8] Polsky D, Glick HA, Willke R *et al.* Confidence intervals for cost-effectiveness ratios: a comparison or four methods. *Health Economics* 1997; **6**: 243–252.

[9] Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and nonparametric techniques using Monte-Carlo simulation. *Statistics in Medicine* 1999; **18**: 3245–3262.

[10] Briggs AH. *Economic evaluation in health care*. Oxford University Press: Oxford, 1993; 172–214.

[11] Davidson R, MacKinnon JG. The size distortion of bootstrap tests. *Working paper GREQAM* 1996; **96a15**.

[12] Davidson R, MacKinnon JG. *Estimation and inference in econometrics*. Oxford University Press: New York, 1993;175–208.

[13] Efron B, Tibshirani RJ. *An introduction to the bootstrap*. Chapman and Hall: New York, 1993.

[14] Willan AR, O'Brien BJ. Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Economics* 1996; **5**: 297–305.

[15] Chaudhary MA, Stearns S. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine* 1996; **15**: 1447–1458.

[16] Bera AK, Jarque CM. An efficient large sample test for normality of observations and regression residuals. *Working Paper in Econometrics* 1981; **40**.

[17] Willan AR, Lin DY. Incremental net benefit in randomised clinical trials. *Statistics in Medicine* 2001; **20**: 1563–1574.

[18] Heitjan DF, Moskowitz AJ, Whang W. Problems with interval estimates of the incremental cost-effectiveness ratio. *Medical Decision Making* 1999; **19**: 9–15.

[19] Heitjan DF. Fieller's method and net health benefits. *Health Economics* 2000; **9**: 327–335.

[20] Glick HA, Briggs AH, Polsky D. Quantifying stochastic uncertainty and presenting results of cost-effectiveness analysis. *Expert Revue of Pharmacoeconomics Outcomes Res* 2001; **1(1)**: 89–99.

[21] Siani C, de Peretti C. Is Fieller's method applicable in all the situations ?. *Working paper GREQAM* 2003; **03a33**: .

A Proofs of theorems in Fieller's method

A.1 Proof of theorems 1 and 1 bis

Preliminary remark
We keep the same notations as for the section 2.3.1. Particularly, η_1 (respectively η_2) corresponds to $\mu_{\Delta C}$ (respectively $\mu_{\Delta E}$) and X_1 (respectively X_2) corresponds to $\Delta\overline{C}$ (respectively $\Delta\overline{E}$). The discriminator of the polynomial function Q is the following

$$\begin{aligned}\Delta &= y^2 - 4xz,\\ \Delta &= 4k_{1-\alpha}^{(1)}\left[\omega_2^2X_1^2 - 2\omega_{12}X_1X_2 - \omega_1^2X_2^2) - k_{1-\alpha}^{(1)}\right],\end{aligned}$$

where $k_{1-\alpha}^{(1)}$ denotes the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom. Let $\gamma = \omega_1^2\omega_2^2 - \omega_{12}^2$ and $c = corr(X_1, X_2)$. We assume that $\gamma \neq 0$, otherwise X_1 and X_2 are perfectly correlated and Ω is not invertible. So, Δ can be writen as follows

$$\Delta = 4k_{1-\alpha}^{(1)}\gamma\left[\frac{\omega_2^2}{\gamma}X_1^2 - 2\frac{\omega_{12}}{\gamma}X_1X_2 + \frac{\omega_1^2}{\gamma}X_2^2 - k_{1-\alpha}^{(1)}\right].$$

If we set

$$\Gamma = \begin{pmatrix} \frac{\omega_2}{\sqrt{\gamma}} & -\frac{\omega_{12}}{\sqrt{\gamma}\omega_2} \\ 0 & \frac{1}{\omega_2} \end{pmatrix} \tag{2}$$

Δ can be written as follows

$$\Delta = 4k_{1-\alpha}^{(1)}\gamma\left[\|\Gamma X\|_2^2 - k_{1-\alpha}^{(1)}\right] ,$$

where $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $\|.\|_2$ denotes the Euclidian norm. The matrix γ is strictly positive: $\gamma = \omega_1^2\omega_2^2(1-c^2)$ with $\omega_1^2 > 0,\ \omega_2^2 > 0$ et $1 - c^2 > 0$ because $c \in]-1, 1[$. Thus

$$\Delta > 0 \Leftrightarrow \|\Gamma X\|_2^2 > k_{1-\alpha}^{(1)}. \tag{3}$$

Proof of the theorem 1
We assume that $x > 0$. We have

$$\begin{aligned} x > 0 &\Leftrightarrow X_2^2 - k_{1-\alpha}^{(1)}\omega_2^2 > 0 ,\\ &\Leftrightarrow \left(\frac{X_2}{\omega_2}\right)^2 > k_{1-\alpha}^{(1)},\end{aligned}$$

this means that $\eta_2 \neq 0$ statistically. It follows from equation 2 that

$$\|\Gamma X\|_2^2 = \left(\frac{\omega_2}{\gamma}X_1 - \frac{\omega_{12}}{\omega_2\gamma}X_2\right)^2 + \left(\frac{X_2}{\omega_2}\right)^2. \tag{4}$$

However, we have

$$\left(\frac{\omega_2}{\gamma}X_1 - \frac{\omega_{12}}{\omega_2\gamma}X_2\right)^2 \geq 0 \text{ and } \left(\frac{X_2}{\omega_2}\right)^2 > k_{1-\alpha}^{(1)},$$

Thus

$$\|\Gamma X\|_2^2 > k_{1-\alpha}^{(1)} \text{ and } \Delta > 0 \text{ , see equation 3.}$$

Proof of the theorem 1 bis
By extending equation 3, we have

$$\Delta \geq 0 \Leftrightarrow \|\Gamma X\|_2^2 \geq k_{1-\alpha}^{(1)}. \tag{5}$$

We assume that $x \geq 0$. We have

$$x \geq 0 \Leftrightarrow \left(\frac{X_2}{\omega_2}\right)^2 \geq k_{1-\alpha}^{(1)}.$$

It results from equation 4 that

$$\|\Gamma X\|_2^2 \geq k_{1-\alpha}^{(1)}.$$

Lastly, we have $\Delta \geq 0$ (see equation 5) and thus $x \geq 0 \Leftrightarrow \Delta \geq 0$.

A.2 Proof of the theorem 2

- Proof of statement 1): It should be noted that $(\Gamma^T\Gamma)^{-1} = \Omega$ and we have

$$\Delta > 0 \Leftrightarrow \|\Gamma X\|_2^2 > k_{1-\alpha}^{(1)}, \text{ with } \Gamma = \Omega^{-1/2}, \text{ see equation 3.}$$

- Proof of statement 2): We have $(X_1, X_2) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\ \Omega\right)$ under (H_0). Let Γ such that $(\Gamma^T\Gamma)^{-1} = \Omega$, then $\Gamma X \sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\ I_2\right)$ under (H_0), where I_2 denotes the matrix identity of $\mathbb{R}^2$. Thus $\|\Gamma X\|_2^2 \sim \chi^2(2)$ under (H_0).
  A test of size α' is done having as null hypothesis $(H_0) :\ (\eta_1, \eta_2) = (0, 0)$. The optimal test yields a rejection region with a size of α' having the following form

$$\|\Gamma X\|_2^2 > k_{1-\alpha'}^{(2)},$$

  where $k_{1-\alpha'}^{(2)}$ denotes the $(1-\alpha')$ quantile of the chi-squared distribution with two degrees of freedom. If $k_{1-\alpha'}^{(2)}$ and $k_{1-\alpha}^{(1)}$ are identified in the previous inequality, it results from equation 3 that Δ is strictly positive.
  Lastly, $\Delta > 0$ is équivalent to reject the null hypothesis $(H_0) :\ (\eta_1, \eta_2) = (0, 0)$ for the test of size $\alpha' > \alpha$ such that $k_{1-\alpha'}^{(2)} = k_{1-\alpha}^{(1)}$.

- Proof of statement 3): This proof results immediately from statement 1).