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

On the basis of Monte-Carlo simulations of the literature, it was believed that both bootstrap and Fieller's methods had equivalent performances for calculating confidence regions for the incremental cost-effectiveness ratio. However, we have carried out Monte-Carlo simulations using real data issued from an economic evaluation carried out alongside a randomised clinical trial for women with breast cancer which show that these methods do not similarly perform in cases where the pair composed by the mean costs difference and the mean effects difference is close to the Y-axis or to the origin of the cost-effectiveness plane. Such situations will frequently occur in practice. In particular, we show that Fieller's method performed significantly better than bootstrap methods that became unstable or even inapplicable when the difference between average effects approaches statistically zero. Since Fieller's method seems to be the best method, we study it in detail in an analytical way for covering all the possibilities. We prove some theorems that permit to see that Fieller's method is "technically" applicable in all the situations, even near the CE plane origin. They also permit to see that the method provides an answer for the decision-making purpose in all the situations.

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, parametric and nonparametric 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). Current conclusions of the literature tend to suggest that both bootstrap and Fieller's methods are the most appropriate and it has been argued, on the basis of research using Monte-Carlo simulations, that both methods obtained similar results [8, 9, 10].

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 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 method, 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, we use Monte-Carlo simulations to compare the performances of bootstrap 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 show that mathematical limitations of the bootstrap methods should lead to consider Fieller's method, which is based on less restrictive conditions, as the method of choice.

Since Fieller's method seems to be the best method, we study it in detail in an analytical way for covering all the possibilities. We prove some theorems that permit to see that Fieller's method is "technically" applicable (i.e. it provides a confidence region) in all the situations, even near the CE plane origin. They also permit to see that the method provides an answer for the decision-making purpose in all the situations.

2 METHODS

In this section, after the definition of the ICER, we present bootstrap and 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 defined as follows:

$$R = \frac{\mu_{C1} - \mu_{C0}}{\mu_{E1} - \mu_{E0}} = \frac{\mu_{\Delta C}}{\mu_{\Delta E}} \ ,$$

where μ is the true mean value of costs and effects for treatments number 1 and number 0.

Since the true means corresponding to the theoretical population are not known, 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 Parametric and 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]. In our case, this method involves building up an empirical estimate of the sampling distribution of the ICER estimator.

The parametric bootstrap method involves assuming that the $(\Delta \bar{C}, \Delta \bar{E})$ pair follows a bivariate normal law with defined mean and variance-covariance matrix such that:

$$\begin{pmatrix} \Delta \bar{C} \\ \Delta \bar{E} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\Delta C} \\ \mu_{\Delta E} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\Delta C} & \sigma_{\Delta C \Delta E} \\ \sigma_{\Delta C \Delta E} & \sigma^2_{\Delta E} \end{pmatrix} \right),$$

$\mu_{\Delta C}$ and $\mu_{\Delta E}$ denote the expected (mean) values, $\sigma^2_{\Delta C}$ and $\sigma^2_{\Delta E}$ denote respectively the variances of the mean costs difference and of the mean effects difference, and $\sigma_{\Delta C \Delta E}$ denotes the covariance between the mean costs difference and the mean effects difference. All these parameters will be estimated by the empirical estimators. This bootstrap method consists in replicating many times (denoted B) bootstrap estimates of the $(\Delta \bar{C}, \Delta \bar{E})$ (denoted $(\Delta \bar{C}^*_b, \Delta \bar{E}^*_b)$, $b = 1, ..., B$) following the bivariate normal law defined above, for generating a vector of bootstrap estimates $(R^*_1, \ldots, R^*_B)$ which is an empirical estimate of the sampling distribution of the ICER estimator.

An alternative method has the substantial advantage of making no parametric assumptions about the sampling distribution of $\Delta \bar{C}$ and $\Delta \bar{E}$. This non parametric bootstrap 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.

In this paper, we have tested the six available modalities of bootstrap methods for calculating confidence intervals for the ICER: parametric bootstrap methods (associated with the percentile method denoted P0, the percentile-t method denoted Pt0 and the Bias Corrected and Accelerated (BCA) method denoted BCA0) as well as nonparametric bootstrap methods (associated with the percentile method denoted P1, the percentile-t method denoted Pt1 and the BCA method denoted BCA1).

All 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 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 $(0,0)$.

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 $\frac{\eta_1}{\eta_2}$. For this purpose, we draw up the (statistic) $Z = X_1 - \rho X_2$ and we note that:

$$Z \sim N(0, \omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}) \text{ under the assumption that ``}\rho = \frac{\eta_1}{\eta_2}\text{''}.$$

Therefore, we have:

$$\frac{Z^2}{\omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}} \sim \chi^2(1)$$
$$\Rightarrow P\left(\frac{(X_1 - \rho X_2)^2}{\omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}} \leq k_{1-\alpha}\right) = 1 - \alpha \;,$$

where $k_{1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom. Hence:

$$P\left((X_1 - \rho X_2)^2 - k_{1-\alpha}(\omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}) \leq 0\right) = 1 - \alpha \;.$$

If we denote:

$$Q(\rho) = (X_1 - \rho X_2)^2 - k_{1-\alpha}(\omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}),$$

we can write the function Q as a two degree polynomial function in the following way:

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

with $x = X_2^2 - k_{1-\alpha}\omega_2^2$, $y = 2(k_{1-\alpha}\omega_{12} - X_1X_2)$ and $z = X_1^2 - k_{1-\alpha}\omega_1^2$. To find the $(1-\alpha)$ confidence region for $\frac{\eta_1}{\eta_2}$, the following inequation must be solved:

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

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:

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

After solving the inequation 1, $(1-\alpha)$ confidence regions for the ICER can have different forms (see table 1). However, the precise forms of confidence regions obtained by Fieller's method has not been studied when the difference between average effects of the two treatments approaches statistically zero or when the (mean costs difference, mean effects difference) pair also approaches statistically zero.

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,

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

R^U denote the roots of the polynomial 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 $x < 0$ corresponds to the case where $\mu_{\Delta E}$ is statistically equal to zero for a test of size α. The table above was studied in details and each form of confidence region was rigorously interpreted and proved (for the interpretations, see appendix 2.3.4 and for the proofs see see appendix A). In particular, a theorem showed that when the discriminator is negative or null, this corresponds to the case where we cannot distinguish $(\mu_{\Delta C}, \mu_{\Delta E})$ from $(0, 0)$.

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 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 3.2).

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 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 ? In particular, Heitjan [19] has found that the confidence region is the whole real line but he did not explain how to use or 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 and the confidence region obtained is the whole real line. More precisely (see statement 2), $(\mu_{\Delta C}, \mu_{\Delta E})$ cannot statistically be distinguished from $(0, 0)$. This means that either both the treatments are equivalent on the two-fold basis of the cost and the medical effects of each treatment, either the sample size is not sufficiently large for distinguishing 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. Anyway, this result is informative. 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 whereas the confidence region can have no mathematical sense, *i.e.* when the ICER is not statistically defined (of the form "$\frac{0}{0}$") or when it is infinite (of the form "$\frac{1}{0}$"). Other approaches such as "the acceptability curves" or " the net benefits approach" do not give informative results too, since when $(\mu_{\Delta C}, \mu_{\Delta E})$ is unsignificant, neither the net benefit statistic cannot be statistically distinguished from zero and we cannot determine whether a treatment is cost-effective.

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

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

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 such that $]-\infty, R^U] \cup [R^L, +\infty[$. 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). 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. 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.

2.4 Empirical data used in Monte-Carlo simulations

In order to test the above theoretical hypothesis that Fieller's method is more appropriate than bootstrap methods when applied in 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 $(0, 0)$, we compare the application of these methods to empirical data issued from a real RCT. 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 ($N+ > 7$). 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.

Group variable	Sample mean	s.e.	c.v.	c.c.	Skewness	Kurtosis
Treatment: n=155				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=145				-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

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 whether the kurtosis was equal to 3 and performed a Jarque-Bera test [16] with both hypothesises combined. The version bootstrap of three tests programmed by Christian de Peretti (GREQAM, Université de la Méditerranée) were used. They showed that all the data were 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.

2.5 The Monte-Carlo simulations

Monte-Carlo simulations were carried out to assess the performances of the seven methods discussed above for calculating confidence region for the ICER with a nominal coverage equal to 0.95.

2.5.1 Methodology of the simulations

At the beginning, this study was performed by varying the following three components: the joint density function of the cost and effect for each treatment (which was either a standard theoretical law that we assumed to be Gaussian or an empirical distribution), the correlation between costs and effects for each treatment group, the distance between the $(\Delta\bar{C}, \Delta\bar{E})$ pair and the origin of the CE plane. Then, since we observed that the distribution of costs and effects as well as their correlation coefficient had little impact on the the performances of the methods (if the same parameters are used), we focus on the variation of the distance between the $(\Delta\bar{C}, \Delta\bar{E})$ pair and the origin of the CE plane and we only report the study based on the empirical distribution of costs and effects which is the most realistic case.

Three possible locations are considered for the $(\Delta\bar{C}, \Delta\bar{E})$ pair: 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

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 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

3.1 Various confidence regions obtained with the real data

Figure 1 gives an idea of the uncertainty associated with each of the methods studied. This figure shows 5000 replications of the incremental cost and incremental effect pairs, and the rays whose slopes correspond to the upper and lower limits of 0.95 confidence regions for the ICER obtained using Fieller's method and nonparametric bootstrap methods (the percentile, percentile-t, and BCA methods). Figure 1 is given as an illustration of the confidence regions obtained with each method but th results of Monte-Carlo simulations have to be examined for assessing the performances of the various methods.

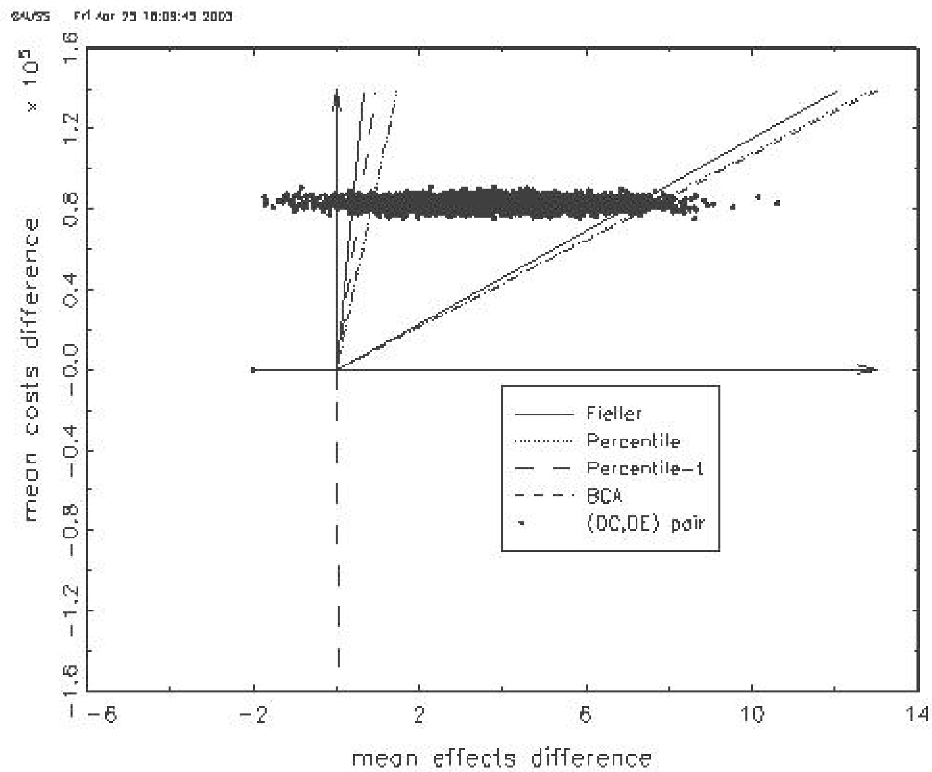

Figure 1: Confidence regions with Fieller's and nonparametric bootstrap methods

3.2 Results of Monte-Carlo simulations

With the location case 1, we observed that all the methods performed well. So, 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 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) are surprising because this methods has a better convergence rate than the percentile method, and should therefore theoretically work better than the percentile method. But locally, in finite samples (between 100 and 200) and when $\mu_{\Delta E}$ approaches statistically zero, the fact of "studentizing" the estimated ICER statistic makes it unstable and makes it farer from a pivotal statistic than the estimated ICER. This explains the substantial

Method	Coverage	Length ($\times 10^3$FF)	Angle (^{o})
Fieller	0.950(0.002)	∞	0.00467 (0.00021)
P0	0.973(0.002)	339047.33(437399.90)	74.72(88.69)
Pt0	0.779(0.007)	4845708.60(22490012.80)	169.38(42.42)
BCA0	0.915(0.002)	∞	55.78(83.24)
P1	0.973(0.002)	326054.57(428734.06)	72.87(88.36)
Pt1	0.743(0.007)	4670597.16(20931815.94)	172.98(34.84)
BCA1	0.920(0.002)	∞	42.81(76.63)

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

Table 4: Performances of the methods on location case 2

Method	Coverage	Length ($\times 10^3$FF)	Angle (^{o})
Fieller	0.952(0.002)	∞	63.26(85.40)
P0	0.971(0.002)	20470.65(30258.49)	113.57(86.48)
Pt0	0.873(0.007)	245663.69(1358599.53)	177.21(21.91)
BCA0	0.930(0.002)	∞	117.81(85.24)
P1	0.972(0.002)	19878.46(29634.71)	119.14(84.76)
Pt1	0.887(0.007)	234050.73(917526.56)	178.57(15.50)
BCA1	0.930(0.002)	∞	103.41(88.61)

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

Table 5: Performances of the methods on location case 3

under-coverage observed for confidence regions and why this method performed less efficiently than the percentile method. So, it is better not to use this method in the case of ratios.

As regards the 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 in case of the nonparametric bootstrap. We also observed that the percentile (respectively BCA) method systematically provided confidence intervals with over (respectively under) coverage, whatever the location of the data. It should be noted that the nonparametric bootstrap method perfomed slightly better than the parametric bootstrap method in case of the empirical distribution.

As regards Fieller's method, it performed 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 bootstrap methods in general, they became more unstable when the (mean costs difference, mean effects difference) pair approaches statistically zero. 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.

The question can be asked as to whether Fieller's method is stable when the (mean costs difference, mean effects difference) pair approaches statistically zero. The results showed that there is no problem here with this method and that the method performed quite well with a coverage almost equal to 0.95 and with around a twice smallest angle than with 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.

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 approximative and that Taylor's method is restrictive and constraining because of the normality hypothesis only valid in the asymptotic context.

In this paper, we have shown that contrary to previous recommendations of the literature, Fieller's and the classical bootstrap methods (without re-ordering) 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. This has been confirmed by Monte-Carlo simulations applied on empirical data. Our Monte-Carlo simulations clearly show that the classical 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.

Previous Monte-Carlo studies of the Literature [8, 9, 10] which compared the performances of Fieller's method and of classical bootstrap methods concluded that they had similar performances. Indeed, this seems to have been the case because the way the true miscoverage of confidence regions was calculated in these studies was not very accurate for distinguishing the methods. For example, the number of Monte-Carlo simulations used respectively in the study of Polsky et al. [8] and in the study of Briggs et al. [9] was equal respectively to 500 and 1000 for one population experimented, which corresponds to a standard error of the estimated overall miscoverage respectively equal to 0.01 and to 0.007 with a nominal miscoverage level chosen equal to 0.05 (i.e. the coefficient of variation of the estimated overall miscoverage was respectively equal to 0.2 and 0.14), which was not accurate enough for a reliable estimation with a nominal miscoverage level equal to 0.05.

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. For example, in [10], the coefficient of variation of $\Delta\bar{E}$ was equal to 0.2. Let us examine in details the coefficients of variation of the mean effects difference chosen in the studies of the literature. Before, it should be noted that the fact that $\mu_{\Delta E}$ is statistically equal to zero for a test of size 0.05 (respectively 0.01)) is equivalent to the following statement in terms of coefficient of variation: $|cv(\Delta\bar{E})| > 0.51$ (respectively $|cv(\Delta\bar{E})| > 0.38$). 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. With the study of Polsky et al. [8], if the correlation between costs and effects was taken equal to zero, the coefficients of variation of the mean effects difference were equal to 0.35 and 0.44 respectively with each distribution of effects chosen. With the first distribution having a coefficient of variation equal to 0.35, the $(\Delta\bar{C}, \Delta\bar{E})$ pair was not particularly close to the Y-axis of the CE plane since $\mu_{\Delta E}$ was statistically equal to zero only for a test of size 0.01. With the second distribution having a coefficient of variation equal to 0.44, the $(\Delta\bar{C}, \Delta\bar{E})$ pair could seem relatively close to the Y-axis, 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. In addition, even if the miscoverage of the worst case and that of the best case were presented, their standard errors were too large (because of the insufficient number of simulations) to provide reliable findings. Likewise, for the study of Briggs et al. [9], the coefficients of variation of the mean effects difference varied from 0.1 to 0.44, the latter value of the coefficients of variation is relevant in terms of neighbourhood of the $(\Delta\bar{C}, \Delta\bar{E})$ pair to the Y-axis of the CE plane and the same comments applied.

The last reason why equivalent performances have been Previously found with

Fieller's and bootstrap methods, is that the miscoverage was measured calculating the mean miscoverage on several DGPs: this hides the problematic cases and reduces artificially the standard error of the estimated miscoverage decrease. In addition, we can question the theoretical justification of using the mean of estimators across several populations which does not estimate the same parameter since the performances of the methods normally have to be calculated with a given Data Generating Process (DGP). Now, by calculating the mean on several DGPs, the bad performance obtained with a particular DGP is hidden in the mass of the other performances. In this context, the problematic cases cannot be detected and this gives the impression that all the methods have good and similar performances. 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.

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

Some of the recent "disappointment" in the literature towards the use of confidence regions for ICERs may have been partly due to a mistaken judgement of equivalence between bootstrap and Fieller's methods to calculate the ICERs, whereas, in fact, Fieller's method is quite robust even in the most 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.

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

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

where $k^{(1)}_{1-\alpha}$ 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

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

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

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

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

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

Proof of the theorem 1

We assume that $x > 0$. We have

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

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

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

Thus

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

Proof of the theorem 1 bis

By extending equation 3, we have

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

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

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

It results from equation 4 that

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

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).

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.