Uncertainty around the Mean Utility Assessment Accounting for Mapping Extrapolation: Application to Prostate Cancer *

Carole Siani [†a] **Christian de Peretti** [b] **Christel Castelli** [c]
Tu Phung [d] **Jean-Pierre Daurès** [c,f] **Gérard Duru** [e]

November 11, 2010

[a] Research Laboratory in Knowledge Engineering (ERIC, EA3083), Institut des Sciences Pharmaceutiques et Biologiques (ISPB), University Claude Bernard Lyon 1, University of Lyon, France.

[b] Laboratory of Actuarial and Financial Sciences (SAF, EA2429), Institute of Financial and Insurance Sciences (ISFA School), University Claude Bernard Lyon 1, University of Lyon, France.

[c] Département Biostatistique Epidemiologie clinique Santé Publique et Information Médicale (BESPIM), Centre Hospitalier Universitaire, Nîmes, France.

[d] Institut Universitaire de Recherche Clinique, laboratoire de biostatistique, Université de Montpellier 1, France.

[f] Laboratoire Epidémiologie et Biostatistique, Institut Universitaire de Recherche Clinique, Montpellier, France.

[e] Cyklad, France.

Abstract

In cost-effectiveness analysis, one or more medical treatment(s) were compared with a standard treatment on the two-fold basis of cost and utility. Since the health utility measure is not necessarily available for the entire sample, this utility measure is often extrapolated from a technical or medical questionnaire through a mapping function. In the literature this mapping is not accounted for when uncertainty is handled, leading to wrong decision-making with serious consequence on the patient's health. The purpose of this paper is then to build a confidence interval around the mean utility measure, accounting for the uncertainty coming from the

*The authors would like to thank IReSP and DGS, DREES-MiRe, InVS, HAS, CANAM, AFSSAPS, Inserm for funding this project. They also thank *Institut de Recherche en Santé Publique* (IReSP) and its director Alfred Spira for inviting them to the IReSP's seminar in Paris, 1st June 2010. Finally, the authors would like to give thanks to Mathea Orsini for her help with the data.

†Correspondence to: Carole Siani, Bâtiment d'odontologie, 11 Rue Guillaume Paradin, 69372 Lyon cedex 08, France. Tel: +33 (0)4 78 77 10 50. Fax: +33 (0)4 72 43 10 44 . Email: carole.siani@univ-lyon1.fr. Web: recherche.univ-lyon2.fr/eric/82-Carole-Siani.html.

questionnaire extrapolation. Analytic and nonparametric bootstrap procedures are proposed. An extension of the methodologies to the Incremental Cost-Utility Ratio is proposed in Appendix.

J.E.L. Classification: C13, C15, [C44]; I19.

Keywords: mapping, mean utility prediction, uncertainty, bootstrap.

1 Introduction

In cost-effectiveness analysis (CEA), one or more medical treatment(s) are compared with a standard treatment on the two-fold basis of cost and medical effectiveness by decision-makers. Only recently, the health utility has been taken into account instead of the sole effectiveness. Since collecting utility data is time consuming and human resource demanding, the utility data are often collected on a sub-sample, conversely to the other data such as cost data. Since the utility measure is not necessarily available for the entire sample, the utility measure is often extrapolated from a questionnaire that is not based on any utility principle such as clinical data of patient characteristics or self-reply questionnaires on quality of life. In practice, the utility is measured on a small sub-sample, whereas a technical questionnaire is collected from all the patients. A mapping is estimated between the utility measure and the questionnaire on the sub-sample. Then, the utility value is predicted for the other patients, extrapolating the questionnaire using the estimated mapping. For instance, the mapping techniques are often used in Rheumatoid Arthritis, where the EuroQol EQ-5D measure is extrapolated from the HAQ or DAS28 scores (see Ariza-Ariza et al. [2006], Nord et al. [1992], Longworth et al. [2005]). For other examples, see among others Torrance [1976], Krabbe et al. [1997], Dolan and Sutton [1997], O'Leary et al. [1995], Torrance et al. [1996].

We will show that decision-making based on utility values interpolated from mapping is not reliable if the mapping is not accounted for. On a medical point of view, when estimating the mean utility of a treatment, it is important to get an accurate confidence interval, since these classical methods can lead to wrong positive decisions.

The purpose of this paper is then to build a confidence interval around the mean utility measure, accounting for the uncertainty coming from the questionnaire extrapolation. We point out that if the extrapolated utility values are used to compute a confidence interval as if they were the observed values, this procedure dramatically decreases the confidence interval so that the conclusion is not reliable since the uncertainty is largely underestimated. This spurious decrease in uncertainty is not accounted for in the studies of the literature and in the CEAs conducted in the pharmaceutical industry. Furthermore, decision-making follows from these results.

In this paper, analytic and nonparametric bootstrap procedures are proposed to account for the questionnaire extrapolation procedure to build the confidence interval for the mean utility. The performance of the methods is assessed via Monte Carlo experiments for various sample sizes and models. Finally, an out of sample validation is carried out to check the performance of the various methods on prostate cancer data. In addition, an extension of the methodologies to the Incremental Cost-Utility Ratio is proposed in Appendix B.

All the programs are written and run with Gauss software ®: the confidence interval and Monte Carlo procedures, the regressions on data, as well as the out of sample val-

idation. The Gauss procedures are available at: http://recherche.univ-lyon2.fr/eric/82-Carole-Siani.html.

The remainder of this paper is organized as follows. section 2 describes the various mapping methods of the literature. section 3 proposes several methods to handle uncertainty around the mean utility. section 4 provides the Monte Carlo results for the performance of the methods. section 5 provides an application to prostate cancer as well as an out of sample validation. Finally, section 6 concludes. In addition, Appendix B proposes an extension of the methodologies to the Incremental Cost-Utility Ratio.

2 Mapping methods in the literature

In this section, mapping methods of the literature used in various medical situations are presented.

2.1 Notations

1. Preference-based single index measure:
   (a) U: overall score, which is rescaled to [0,1].
   (b) U_k, $k = 1, \ldots, K$: the level of the index dimensions.
2. Non-preference based condition-specific instrument:
   (a) X: overall score,
   (b) X_d, $d = 1, \ldots, n_D$: the score of the instrument domains,
   (c) Z_i, $i = 1, \ldots, n_I$: the level of the instrument items, which can be discrete variables,
   (d) $I_{i,z}$, $i = 1, \ldots, n_I$, $z \in E$: a dummy variable that $= 1$ when the level of the instrument item i is z. E is the set of values for z.
3. other independent variables:
   (a) W: row vector of patient characteristics such as age, age squared, sex or others.

The paper of Tsuchiya et al. [2002] is presented first. The authors convert Asthma Quality of Life Questionnaire (AQLQ) - a non-preference-based QOL instrument for asthma - into EQ-5D indices - a preference-based generic instrument -. They propose a simple transformation (linear), multi-linear regressions over the various domains or items. They are presented below.

2.2 Simple linear regression mapping

$$U = \alpha + \beta X + W\omega + u \tag{1}$$

α and β are parameters, and ω is a column vector of parameters.

2.3 Multi-linear regression mapping

$$
\begin{aligned}
U &= \delta_0 + \delta_1 X_1 + \ldots + \delta_{n_D} X_{n_D} + W\omega^{\star} + v && (2)\\
U &= \iota_0 + \iota_1 Z_1 + \ldots + \iota_{n_I} Z_{n_I} + W\omega^{\star\star} + w && (3)\\
U &= \iota'_0 + \sum_{i=1}^{n_I} \iota'_i Z_i + \sum_{i=1}^{n_I}\sum_{j=1}^{n_I} \iota'_{ij} Z_i Z_j + W\omega' + w' && (4)
\end{aligned}
$$

Models 1–4 are estimated using ordinary least squares (OLS), and the independent variables are treated as continuous.

$$
U = \iota''_0 + \sum_{i=1}^{n_I} \sum_{z \in E \setminus \{\text{last element of } E\}} \iota''_{i,z} I_{i,z} + W\omega'' + w'' \qquad (5)
$$

Model 5 continues to use OLS, but now the independent variables $I_{i,z}$ are treated as categorical variables (with possibly continuous variables W).

2.4 Multiple linear regressions

$$
U_k = \iota_{k,0} + \iota_{k,1} Z_1 + \ldots + \iota_{k,n_I} Z_{n_I} + W\omega_k + w_k, \qquad k = 1, \ldots, K \qquad (6)
$$

Again, OLS is used, and both the dependent U and the independent variables are treated as continuous. [1]

2.5 Power models

In Stevens et al. [2006], Shmueli [2007], the authors propose mapping between Visual Analogue Scale and Standard Gamble data. They assume a model of the following form:

$$
U_{is} = f(X_{is}) + \varepsilon_{is} \qquad (7)
$$

where i represents an individual and s represents a health state. [2] The authors assume that the transformation function $f(\ .\ )$ do not vary across individuals and that they are independence across observations. They assume the errors are normally distributed:

$$
\varepsilon_{is} \sim N(\mu_i, \sigma_i^2)
$$

Linear, quadratic, cubic and power functions are estimated. The linear and power models are estimated using the disvalue form, where disvalue =(1-value) and disutility =(1-utility). Since the utility measure belongs to [0,1]; the quadratic and cubic models use the value form and are constrained to pass through 0 and 1. To allow for this constraint, the quadratic model is estimated as:

$$
U - X^2 = \alpha(X - X^2)
$$

[1] In practice, if Z_i are discrete, the model is a mixture of discrete and continuous variables.

[2] s belongs to an (*a priori*) infinite continuous set, since for an analogical scale all the values are allowed. Otherwise, we are limited to some integer numbers.

and the cubic model is estimated as:

$$U - X^3 = \alpha(X - X^3) + \beta(X^2 - X^3)$$

where U is the utility, X is the overall score of the instrument, and α and β are parameters to be estimated.

The constraints to pass through 0 and 1 should also be applied to the previous models. However, they are not necessarily respected by the data in practice. In addition, the error terms u are out of this restriction. Consequently, we prefer not to impose them, and to estimate the following models:

$$\text{Quadratic: } U = \alpha + \beta X + \gamma X^2 + W\omega + u \quad (8)$$

$$\text{Cubic: } U = \alpha + \beta X + \gamma X^2 + \delta X^3 + W\omega + u \quad (9)$$

2.6 Generalised linear model (GLM)

The dependent variable U is transformed into an s-shaped non-linear variable that approaches 1, but does not reach it. The logit transformation can be applied. The obvious shortcoming of this in the context of Tsuchiya et al. [2002] is that there are many responses with observed EQ-5D index of 1.00, and the transformation will imply dropping these observations (because the transformed values approach infinity). Tsuchiya et al. [2002] accommodated this by standardising the raw EQ-5D indices to the range [0,1], based on an artificial range, say, [-0.5, +1.1], and then transforming this. In their paper, Tsuchiya et al. [2002], given the additional complication, the arbitrary nature of the standardisation and the transformation, and the fact that the maximum predicted EQ-5D indices of the simple linear models hardly exceed 1.00, the associated benefits of GLM do not seem to outweigh its costs. Therefore Tsuchiya et al. [2002] decided not to use GLM. Another way to overcome this problem would be to manage the 1's as $1 - \varepsilon$ in the logistic model, with ε converging to 0.

3 Methods for calculating the mean utility confidence interval

Beresniak et al. [2007] first criticised the use of mapping in cost-utility analysis since the uncertainty is not accounted for in usual methods.

In their papers, Stevens et al. [2006], Shmueli [2007], Salomon and Murray [2004], Longworth et al. [2005] report the prediction of their models for the mean utility. However, no confidence interval of this prediction is given. The absolute error of the model is provided, but again, it does not correspond to an estimate error of the mean.

In their paper, Rivero-Arias et al. [2009] also report the mean utility on the out of sample. They provide a confidence region for the mean utility. They argue that "In terms of predicting uncertainty around EQ-5D mean estimates, their model estimated tighter 95% CIs than the actual 95% CIs." However, these confidence intervals are underestimated (see Table 1).

The paper of Merkesdal et al. [2010] can also be cited. It deals with rheumatoid arthritis. The HAQ score changes are mapped to related health utility states expressed in QALYs gained. For this transformation, a standardised formula was used according to

the most published cost-effectiveness models in RA. In the base case scenario, the formula by Bansback et al. [2005] (female) was applied:

$$QoL = 0.76 - 0.28 * HAQ - 0.05 * \text{Female}. \tag{10}$$

The utility estimates of the model result are provided in terms of QALYs for each sequence of treatment with a confidence interval. It should be noted that it is impossible to get a confidence interval from Equation 10. Thus, the confidence interval provided by Merkesdal et al. [2010] should come from the mapped values, and should be undervaluated. In addition, the authors provide a cost-effectiveness acceptability curve, which should be wrong for the same reason.

Consequently, in this paper, we propose a method to provide the confidence interval of the mean utility.

3.1 Statistical properties of the utility measure

Assumption 1: Utility

It is assumed that the utility measures u_i follow independent random variables with $[0, 1]$ support and with distribution denoted D_U as follows:

$$U_i \sim \text{ independent } D_U(\mu_U, \sigma_U^2), \tag{11}$$

where i denotes the individuals, μ_U the mean, and σ_U^2 the variance.

Since the support of the random variable U_i is finite, all the moments of the distribution are finite. Since the true mean corresponding to the theoretical population is unknown, the mean utility can be estimated as follows, on the basis of data collected from groups of patients, each undergoing one of the forms of therapy (one group consisting of n individuals):

$$\bar{U} = \frac{1}{n} \sum_{i=1}^{n} U_i. \tag{12}$$

Under Assumption 1, and applying the Central Limit Theorem (CLT), we obtain:

$$\bar{U} \sim N\left(\mu_U, \frac{1}{n}\sigma_U^2\right). \tag{13}$$

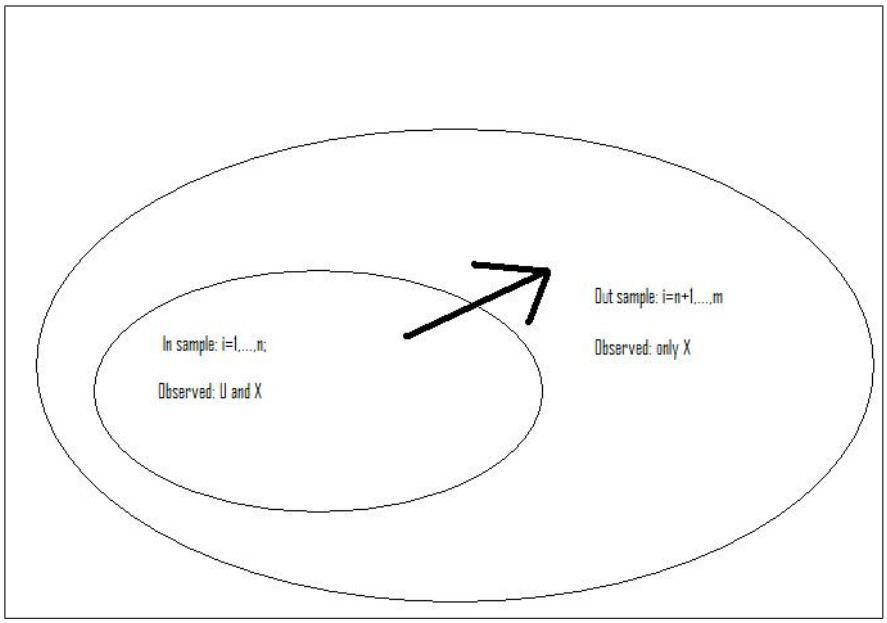
Assumption 2: Mapping

It is assumed that the utility measure can be explained by some variables X.

$$U_i = f(X_i; \beta) + \varepsilon_i, \tag{14}$$
$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{15}$$

where f is a function that depends on a parameter vector β, and ε_i is independent of X. For simplicitys sake, at a first stage, it can be assumed that f is linear: $U_i = X_i\beta + \varepsilon_i$. Thus, $\mu_U = E(X_i)\beta$, and $\sigma_U^2 = \beta' V(X)\beta + \sigma_\varepsilon^2$.

Let us consider an "in sample" with individuals $i = 1, \ldots, n$ where the utility measure is known, and an "out of sample" with individuals $i = n+1, \ldots, n+m$ where the utility value is unknown.



The aim of mapping is to assess the out of sample mean utility $\mu_{U_{\text{out}}} = E(U_{\text{out}})$. In the linear case, $\mu_{U_{\text{out}}} = E(X_{\text{out}})\beta$, where X_{out} is the matrix of independent variables for the out of sample. The estimator for the mean utility generally used is:

$$\hat{\mu}_{U_{\text{out}}} = \frac{1}{m}\sum_{i=n+1}^{n+m}\hat{U}_i = \frac{1}{m}\sum_{i=n+1}^{n+m}X_i\hat{\beta}_{\text{in}} = \bar{X}_{\text{out}}\hat{\beta}_{\text{in}} = \frac{1}{m}\iota_m' X_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'U_{\text{in}},$$

where $\hat{U}_i$ are the out of sample predicted values for the utility measure, $\hat{\beta}_{\text{in}}$ is the estimate of β using the mapping on the in sample, $\iota_m = (1, \ldots, 1)'$, X_{in} is the matrix of independent variables for the in sample, U_{in} are the utility measures for the in sample, and $'$ denote the transpose. It should be noted that:

$$\hat{\mu}_{U_{\text{out}}} = \frac{1}{m}\iota_m X_{\text{out}}\beta + \frac{1}{m}\iota_m X_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}.$$

Then, we have the following properties: $E(\hat{\mu}_{U_{\text{out}}}) = E(X_{\text{out}})\beta$. Thus, the estimator $\hat{\mu}_{U_{\text{out}}}$ is unbiased, and can be used to assess the mean utility. However, in order to take a decision, the uncertainty has to be accounted for.

3.2 "Naive" confidence interval

A naive, and wrong, way to compute the variance of $\hat{\mu}_{U_{\text{out}}}$ would be:

$$\widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U_{\text{out}}}\right) = \frac{1}{m}\left[\frac{1}{m-1}\sum_{i=n+1}^{n+m}\left(\hat{U}_i - \hat{\mu}_{U_{\text{out}}}\right)^2\right]. \tag{16}$$

We report here the confidence interval that is provided by Rivero-Arias et al. [2009]. A naive way to provide a $(1-\alpha)$ confidence interval is:

$$CI^{\text{Naive}}_{\mu_U}(1-\alpha) = \left[\hat{\mu}_{U_{\text{out}}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U_{\text{out}}}\right)}\right], \tag{17}$$

where Φ is the cumulative distribution function of a standard normal random variable.

3.3 Analytic confidence interval

In the case of a linear model, we derived the variance of $\hat{\mu}_{U_{\text{out}}}$, which can be estimated as follows:

$$\hat{V}\left(\hat{\mu}_{U_{\text{out}}}\right) = \frac{1}{m}\hat{\beta}'_{\text{in}}\hat{\Omega}_X\hat{\beta}_{\text{in}} + \hat{\sigma}^2_{\varepsilon_{\text{in}}}\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}}. \tag{18}$$

See proof in Appendix, section A. It should be noted that

$$\hat{V}\left(\hat{\mu}_{U_{\text{out}}}\right) = \widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U_{\text{out}}}\right) + \hat{\sigma}^2_{\varepsilon_{\text{in}}}\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}}. \tag{19}$$

Thus, the term $\hat{\sigma}^2_{\varepsilon_{\text{in}}}\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}}$ is missing in Equation 16. The asymptotic $(1-\alpha)$ confidence interval is:

$$CI^{\text{Analytic}}_{\mu_U}(1-\alpha) = \left[\hat{\mu}_{U_{\text{out}}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\hat{V}\left(\hat{\mu}_{U_{\text{out}}}\right)}\right]. \tag{20}$$

This analytic confidence interval is restricted to a linear framework with Gaussian error terms. It can easily be extended to a nonlinear framework, using an Edgeworth expansion, but it will be an approximation. Consequently, we prefer to develop a bootstrap methodology, which will account for any nonlinear specification (such as logistic specification for instance). In addition, if a nonparametric version of bootstrap method is used, non-Gaussian distributed error terms can also be accounted for.

3.4 Nonparametric bootstrap confidence interval

In this subsection, we propose a methodology for building a confidence interval based on the nonparametric bootstrap technique. For a general presentation of the percentile-t method, see Hall [1992], Davidson and MacKinnon [1993], Efron and Tibshirani [1993], Hjorth [1994], and Shao and Tu [1995].

First, a particular mapping model has to be chosen. Let us denote it:

$$U = f(X;\beta) + \varepsilon, \tag{21}$$

where X is a regressor matrix, and the function f is known and is parametric in the sense that it depends on a parameter vector β. The confidence interval is built as follows:

1. Equation 21 is estimated using appropriate methods such as OLS, Nonlinear Least Squares, Maximum Likelihood, ..., leading to $\hat{\beta}_{\text{in}}$, $\hat{\varepsilon}_{\text{in}}$, and $\hat{U}_{\text{out}}$.

2. $\hat{\mu}_{U_{\text{out}}}$ and $s = \sqrt{\hat{V}\left(\hat{\mu}_{U_{\text{out}}}\right)}$ are computed using the original dataset.

3. A bootstrap Data Generating Process (DGP) has to be defined. It may be either parametric or semi-parametric, characterized by $\hat{\beta}_{\text{in}}$ and by any other relevant estimates, such as the error variance $\hat{\sigma}^2_{\text{in}}$, that may be needed. In a general case, we propose:

$$U^b_i = f\left(X_i, \hat{\beta}_{\text{in}}\right) + \varepsilon^b_i, \tag{22}$$

for $i = 1, \ldots, n$. In the case of linear modelling, the DGP is restricted to:

$$U_i^b = X_i \hat{\beta}_{\text{in}} + \varepsilon_i^b, \tag{23}$$

The distribution of ε_i^b will be discussed in the last paragraph of this section.

4. B bootstrap samples are generated, $(U_i^b)_{i=1}^n$, $b = 1, \ldots, B$.

5. For each of these, an estimate $\hat{\mu}^b_{U_{\text{out}}}$ and its standard error s^b are computed in exactly the same as for $\hat{\mu}_{U_{\text{out}}}$ and s from the original data. Then the bootstrap "t-statistic" is computed:

$$\tau^b = \frac{\hat{\mu}^b_{U_{\text{out}}} - \hat{\mu}_{U_{\text{out}}}}{s^b}.$$

6. The asymmetric equal-tail bootstrap confidence interval can be written as:

$$CI^{\text{Bootstrap}}_{\mu_U}(1-\alpha) = \left[\hat{\mu}_{U_{\text{out}}} - s \cdot \hat{F}^{-1}(1 - \frac{\alpha}{2}), \hat{\mu}_{U_{\text{out}}} - s \cdot \hat{F}^{-1}(\frac{\alpha}{2})\right], \tag{24}$$

where $\hat{F}$ is the Empirical Distribution Function of the B bootstrap statistics τ^b.

We consider the following way of generating the bootstrap residuals ε_i^b (see Weber [1984]). The ε_i^b are generated by independent uniform draws with replacement among the vector with the typical element $\tilde{\varepsilon}_i$ constructed as follows:

1. Calculate $(P_X)_{i,i}$, $i = 1, \ldots, n$, the diagonal elements of the projection matrix on X.

2. Calculate $\frac{\hat{\varepsilon}_i}{\sqrt{1-(P_X)_{i,i}}}$, $\forall i = 1, \ldots, n$.

3. Re-centre the vector that results.

4. Rescale it so that it has the variance $\hat{\sigma}^2_{\varepsilon}$.

This makes it possible to correct the heteroskedasticity in the residuals due to the regressors.

4 Performance of the methods: Monte Carlo experiments

In this section, the performance of the various methods is examined using Monte Carlo experiments.

4.1 Monte Carlo methodology

Data Generating Processes (DGP) are used to generate simulated data samples. The various methods are applied to each simulated sample $j = 1, \ldots, S$, where S is the number of Monte Carlo replications in an experiment. It is examined if each confidence interval number j contains or not the true value μ_U of the mean utility (which is known,

since the DGP is known, conversely to real data). The coverage c of the confidence intervals can be estimated as follows:

$$\hat{c} = \frac{1}{S} \sum_{j=1}^{S} I\left(\mu_U \in Interval_j\right). \tag{25}$$

The standard deviation of this Monte Carlo estimate of the coverage is equal to $\sqrt{\frac{1}{S}c(1-c)}$, where c is the coverage.

In our Monte Carlo experiments, we choose the confidence level $1-\alpha$ is equal to 0.95. The number of bootstrap replications is $B = 999$. The number of Monte Carlo replications is $S = 10,000$. If the true coverage $c = 0.95$, the standard deviation of the Monte Carlo estimate of the coverage is 0.002179. At worst (when $c = 0.5$) the standard deviation is 0.005. Several values for the in sample size n and the out of sample size m are chosen. Small values for n and large values for m are chosen to reflect the case where the utility is assessed only on a small sub-sample, and then extrapolated to the other patients.

4.2 Data Generating Process

A variety of DGP are proposed to check the robustness of the methods: linear, nonlinear, with Gaussian and non-Gaussian error terms.

4.2.1 Linear Case

$$\begin{aligned}
U_i &= \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \varepsilon_i, && (26)\\
X_{1i} &\sim i.i.d. U([0,1]), && (27)\\
X_{2i} &\sim i.i.d. B(0.5, 3), && (28)\\
\varepsilon_i &\sim i.i.d. N(0, \sigma^2). && (29)
\end{aligned}$$

The parameters values are set to: $\beta_0 = 0.1$, $\beta_1 = 0.6$, $\beta_2 = 0.2$, $\sigma = 0.3$. We have $E(U_i) = 0.7$.

4.2.2 Random Linear Case

The model is the same, but the parameters vary randomly across the Monte Carlo replications:

$$\begin{aligned}
\beta_0 &\sim i.i.d. N(0.1, 0.1^2), && (30)\\
\beta_1 &\sim i.i.d. N(0.6, 0.6^2), && (31)\\
\beta_2 &\sim i.i.d. N(0.2, 0.2^2), && (32)\\
\sigma &\sim i.i.d. U([0.3 \cdot 0.5, 0.3 \cdot 1.5]). && (33)
\end{aligned}$$

4.2.3 Non-Gaussian Random Linear Case

The model is the same as in Equation 26–Equation 28, but the error terms follow the uniform distribution:

$$\varepsilon_i \sim i.i.d. U([-0.6, 0.6]). \tag{34}$$

4.2.4 Nonlinear Case

$$
\begin{aligned}
U_i &= F\left[\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \varepsilon_i\right] && (35)\\
X_{1i} &\sim i.i.d.U([0,1]) && (36)\\
X_{2i} &\sim i.i.d.B(0.5,3) && (37)\\
\varepsilon_i &\sim i.i.d.N(0,\sigma^2) && (38)
\end{aligned}
$$

The parameters values are set to: $\beta_0 = -0.6$, $\beta_1 = 0.6$, $\beta_2 = 0.2$, $\sigma = 0.3$. We have $E(U_i) = 0.5$.

4.3 Performance

The performance of the various confidence intervals is assessed in terms of coverage and interval length for a 95% confidence level. These coverages are computed using Monte Carlo experiments for various sample sizes, and are presented in Table 1. The sample sizes are chosen to correspond to a mapping assessment on a small sub-sample ($n \leq 40$), and then an extrapolation of the utility values to the remaining sample having a size that can be encountered in practice ($m \geq 400$). If a method performs correctly, its coverage has to be close to the confidence level.

Table 1: Coverage and mean length of the 95% confidence intervals

		Coverage			Mean length		
n	m	Naive	Analytic	Bootstrap	Naive	Analytic	Bootstrap
Linear Data Generating Process							
40	400	0.3930	0.9465	0.9419	0.05149	0.1975	0.1962
30	600	0.2834	0.9458	0.9477	0.04145	0.2251	0.2318
20	800	0.2017	0.9333	0.9472	0.03580	0.2767	0.2984
Random Linear Data Generating Process							
40	400	0.4709	0.9466	0.9253	0.06616	0.2045	0.1952
30	600	0.3681	0.9400	0.9366	0.05268	0.2294	0.2312
20	800	0.2624	0.9394	0.9463	0.04545	0.2789	0.2966
Non-Gaussian Random Linear DGP							
40	400	0.4308	0.9440	0.9378	0.06662	0.2311	0.2248
30	600	0.3202	0.9411	0.9509	0.05318	0.2630	0.2679
20	800	0.2348	0.9366	0.9573	0.04615	0.3241	0.3468
Nonlinear Data Generating Process							
40	400	0.3991	0.9443	0.9421	0.01928	0.07376	0.07323
30	600	0.2794	0.9387	0.9426	0.01558	0.08414	0.08661
20	800	0.1987	0.9369	0.9509	0.01352	0.1039	0.1119

n is the in sample size used to assess the mapping. m is the out of sample size where the utility values are predicted.

The results show that the coverage of the "naive" confidence interval is very small (around 30%) with respect to the confidence level (95%), whereas both the analytic and bootstrap confidence intervals perform correctly.

5 Application to Prostate Cancer

The above methodologies are then applied to data collected from a longitudinal prospective cohort study evaluating patients' health-related quality of life after treatment for prostate cancer. Several centres in the Herault and Gard regions participated in the study. It has been financed by the French National Institute of Cancer (INCa).

5.1 Data description

Data on patients' health-related quality of life were repeatedly collected 4 times. The first time was done after diagnosis of prostate cancer and before treatment. The second, third and fourth times were respectively 2 months, 6 months and 12 months after treatment, respectively. The treatment under consideration here is laparoscopic radical prostatectomy for which there are 116 patients. The utility of patients' own health state was evaluated directly by Standard Gamble method. The health related quality of life information was evaluated by several multidimensional auto-administered quality of life questionnaires. There were two generic questionnaires (EQ-5D and SF-36), one cancer specific quality of life questionnaire (EORTC-QLQ C30), one prostate cancer specific questionnaire (EORTC-PR25) and 3 prostate cancer – related symptoms questionnaire (ICS, IIEF, and IPSS).

The dependent variable of the model is the SG score. The independent variables included in the study are the following:

- The QLQ-C30 questionnaire, including
  - the Global health status (GS),
  - the Functional scales (FS) with the Physical Functioning (CF), the Role functioning (CT), the emotional functioning (EM), the Cognitive functioning (CC) and the social functioning (CS) scores,
  - and the Symptom scales (SyS) with the Fatigue (FA), Nausea and vomiting (NV), Pain (DOU), Dyspnoea (DY), Insomnia (IN), Appetite loss (MA), Constipation (CO), Diarrhoea (DI) and Financial difficulties (DF) scores.
- IPSS and IIEF-5 questionnaires.
- The SF-36 questionnaire, including
  - the Physical Composite Score (PCS),
  - and the Mental Composite Score (MCS).
- The Visual Analogic Scales (VAS).
- Some individual characteristics: age and Prostate Specific Antigen (PSA) of the patients. [3]

[3] Since the gender is necessarily "male", thus, the sex variable is not included.

5.2 Link between the explanatory variables and the utility measure

All the explanatory variables and square age are considered in a linear regression including the patients undergoing laparoscopic radical prostatectomy treatment. First, only visits 2, 3, and 4 are considered (since visit 1 corresponds to a visit before intervention). After recursively removing insignificant variables up to the 10% level [4], the results for the remaining variables are presented in Table 2. It should be noted that all the variables are significant at the 5% level.

Table 2: Laparoscopic radical prostatectomy regression
Visits 2, 3, and 4 (after intervention)

Dependent variable: SG

Valid cases:	32	Missing cases:	55
Total SS:	0.712	Degrees of freedom:	25
R-squared:	0.733	Rbar-squared:	0.669
Residual SS:	0.190	Std error of est:	0.087
F(6,25):	11.432	Probability of F:	0.000

Variable	Estimate	Standard Error	t-value	Prob $> \|t\|$	Standardized Estimate	Cor with Dep Var
Constant	2.374438	0.963147	2.465292	0.021	—	—
PCS	0.038420	0.005605	6.854077	0.000	1.511375	0.328354
MCS	0.009238	0.002438	3.789367	0.001	0.655397	-0.059723
CT	-0.009654	0.001611	-5.993522	0.000	-1.593518	-0.190286
DI	0.004355	0.001747	2.493401	0.020	0.402254	-0.370489
Age	0.000377	0.000117	3.206646	0.004	5.481131	0.132206
Age^2	-0.000000	0.000000	-3.299575	0.003	-5.631635	0.150259

The utility after prostatectomy intervention is explained by the Physical Composite Score (PCS) and the Mental Composite Score (MCS) of the SF-36 questionnaire, Role Functioning and Diarrhoea scores of the QLQC-30 questionnaire as well as age. It should be noted that the SF-36 generic questionnaire and some scores of the QLQ-C30 are sufficient to explain the utility, whereas the prostate cancer – related symptoms questionnaires do not provide any additional explanation.

For information, the results for regression for visit 1 (before the intervention) are provided bellow. The regression is also run for visit 1 (before the intervention). The results are provided in Table 3.

The utility is only partially explained (R-squared = 0.578) by the Visual Analogic Scales (VAS) and by the Fatigue (FA) and the Appetite loss (MA) of the QLQ-C30 questionnaire. The SF-36 does not explain the utility before intervention.

[4] All the variables are first included in the regression equation. Then the most insignificant variable is removed (the variable that has the largest P value). The regression equation is re-estimated. The most insignificant variable is removed from the remaining variables. The process is iterated until all the variables are significant at a certain significance level, say 10 or 5%.

Table 3: Laparoscopic radical prostatectomy regression
Visit 1 (before intervention)

Dependent variable: SG

Valid cases:	21	Missing cases:	8
Total SS:	0.586	Degrees of freedom:	17
R-squared:	0.578	Rbar-squared:	0.503
Residual SS:	0.247	Std error of est:	0.121
F(3,17):	7.759	Probability of F:	0.002

Variable	Estimate	Standard Error	t-value	Prob $>\|t\|$	Standardized Estimate	Cor with Dep Var
Constant	0.377337	0.156540	2.410489	0.028	—	—
FA	0.003813	0.001326	2.875257	0.010	0.543055	0.218644
MA	-0.004309	0.001459	-2.953608	0.009	-0.567466	-0.356843
VAS	0.005371	0.001862	2.883979	0.010	0.464202	0.552962

5.3 Cross validation of the confidence intervals

The purpose of this subsection is to apply the confidence intervals on the data sample (prostate cancer in the case of laparoscopic radical prostatectomy). The confidence intervals are computed as if the utility was observed only on a sub-sample, and it is examined whether the intervals contain or not the empirical average utility on the remaining sub-sample.

It should be noted that the coverage of the confidence intervals cannot be computed using out of sample validation, since the true value for the expected utility is not known. Since the empirical average utility is a random variable, the probability of containing it is smaller than the coverage. However, examining whether the empirical average utility is in each confidence interval allows checking the behaviour of the intervals in the case of real data. This is the reason why this procedure is run only few times, just for information, and not a very large number of times as for Monte Carlo experiments, which have to numerically compute the coverage with the largest possible precision.

The 10% significant level explanatory variables of Table 2 are considered: PCS, MCS, CT, DI, Age, Age^2, as well as the SG score. Removing the missing values, 32 observations are still valid. The mapping is then carried out on a sub-sample constituted from the 11 observations (about a third of the size of the full sample), which are randomly drawn (following a uniform distribution with no replacement). Then, the various confidence intervals are computed and reported in Table 4 and are compared to the empirical mean utility of the remaining sample. This procedure is run ten times.

The results show that the naive interval contains the empirical mean utility of the remaining sample only 6 times out of 10, whereas the analytic interval contains it 7 times out of 10, and the bootstrap interval contains it 6 times out of 10. If this procedure is run 1000 times, the naive interval contains the mean utility about 74%, the analytic one 85% (which is not bad, since the coverage is greater), and the bootstrap one 70%. The nonparametric bootstrap suffers from the small sample size. In this case, the analytic method is more robust.

Table 4: 95% confidence intervals

Gauss RndSeed †	Limit	95% Confidence interval Naive	Analytic	Bootstrap	Mean utility
1	lower	0.8799	0.8478	0.8695	0.9119
	upper	0.9850	1.0171	0.9946	
	Contains ‡	yes	yes	yes	
2	lower	0.9430	0.5876	0.5136	0.9405
	upper	1.1494	1.5049	1.6310	
	Contains	no	yes	yes	
3	lower	0.9586	0.9464	0.9460	0.8833
	upper	0.9969	1.0090	1.0129	
	Contains	no	no	no	
4	lower	0.8325	0.7924	0.8561	0.9362
	upper	0.9845	1.0247	0.9658	
	Contains	yes	yes	yes	
5	lower	0.8884	0.9076	0.9465	0.9167
	upper	1.0572	1.0379	1.0035	
	Contains	yes	yes	no	
6	lower	0.8460	0.7474	0.7464	0.9219
	upper	0.9662	1.0648	1.0566	
	Contains	yes	yes	yes	
7	lower	0.9001	0.9047	0.9241	0.8952
	upper	0.9835	0.9789	0.9602	
	Contains	no	no	no	
8	lower	0.8794	0.8122	0.8482	0.9267
	upper	0.9917	1.0590	1.0264	
	Contains	yes	yes	yes	
9	lower	0.8427	0.7851	0.8189	0.9167
	upper	0.9491	1.0067	0.9819	
	Contains	yes	yes	yes	
10	lower	0.9449	0.9466	0.9518	0.8857
	upper	1.0090	1.0073	0.9992	
	Contains	no	no	no	

† The RndSeed number specifies the seed of the random number generator in Gauss software ®.
Each different seed leads to a different (and independent) set of random numbers.
‡ This row indicates whether the confidence interval contains or not the empirical mean utility of the remaining sample.
The number of bootstrap replications is $B = 999$.

6 Discussion and conclusion

On a medical point of view, when estimating the mean utility of a treatment, it is important to get an accurate confidence interval. Indeed, the final purpose is to carry out a cost-utility analysis (using for instance the incremental cost-utility ratio -ICUR-).

In CUA, the classical methods do not account for the mapping phenomenon. We have shown that decision-making based on utility values interpolated from mapping is not reliable if the mapping is not accounted for : a "naive interval" (which does not account for mapping) would lead to a serious mistake: the coverage is between 20% and 40% for a 95% confidence level! Then, the utility can be greatly under or over evaluated. These methods can lead to wrong positive decisions. In addition, it is often possible to determine any sub-groups of patients for which a new treatment is more effective (in terms of utility) than the standard treatment (for instance a sub-group for which the disease if more serious). Thus, in these cases, it is essential to account for the uncertainty due to mapping to get an accurate confidence interval with respect to the risk of coverage α.

Our analytic and bootstrap procedures, integrating the mapping, provide very accurate results. Monte Carlo experiments show that the analytic and bootstrap 95% confidence intervals display coverage between 94% and 96% for various sample sizes, with a reasonable interval length. In addition, the cross validation from the laparoscopic radical prostatectomy data shows similar results in terms of coverage.

In the example of the prostate cancer, the utility after prostatectomy intervention is explained by the Physical Composite Score (PCS) and the Mental Composite Score (MCS) of the SF-36 questionnaire, Role Functioning and Diarrhoea scores of the QLQC-30 questionnaire as well as age. These results have to be interpreted with caution because of the small number of patients for which we have all the information. In addition, in our model, to increase the number of observations, we had to consider that the visits were independent, which is not necessarily true in practice. However, the large place taken up by the SF36, and completed by any scores of the QLQC-30, is coherent. Nevertheless, any items, characteristic of the consequences of an intervention on the prostate, do not appear.

These methodologies can b extended to the Incremental Cost-Utility Ratio. This allows to directly do the cost-utility analysis on the basis of cost and utility simultaneously. Appendix B proposes an extension of the methodologies to the Incremental Cost-Utility Ratio.

Appendix

A Proof of the analytic confidence interval

Assume the following linear model:

$$U_i = X_i\beta + \varepsilon_i.$$

Let the in sample be denoted: $i = 1, \ldots, n$, and the out of sample be denoted: $i = n+1, \ldots, n+m$.

The aim is to assess the out of sample mean utility:

$$\mu_{U_{\text{out}}} = E(U_{\text{out},i}) = E(X_{\text{out},i})\beta.$$

The estimator of $\mu_{U_{\text{out}}}$ is:

$$\hat{\mu}_{U_{\text{out}}} = \bar{X}_{\text{out}}\hat{\beta}_{\text{in}} = \frac{1}{m}\iota_m' X_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'U_{\text{in}},$$

where $\iota_m = (1, \ldots, 1)'$. It can be noted that:

$$\hat{\mu}_{U_{\text{out}}} = \frac{1}{m}\iota_m X_{\text{out}}\beta + \frac{1}{m}\iota_m X_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}. \tag{39}$$

Then, the estimator has the following properties.

A.1 Bias

$E(\hat{\mu}_{U_{\text{out}}}) = E(X_{\text{out}})\beta$, the estimator is unbiased.

A.2 Variance

$$V(\hat{\mu}_{U_{\text{out}}}) = V(\bar{X}_{\text{out}}\hat{\beta}_{\text{in}}).$$

From Equation 39, we have:

$$\begin{aligned}
V\left(\hat{\mu}_{U_{\text{out}}}\right) &= V\left[\bar{X}_{\text{out}}\beta + \bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\right] && (40)\\
&= V\left[\bar{X}_{\text{out}}\beta\right] + V\left[\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\right] \\
&\quad +2\cdot cov\left[\bar{X}_{\text{out}}\beta, \bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\right] && (41)\\
&= A + B + 2\cdot C && (42)\\
A &= \frac{1}{m}\beta' V(X_{\text{out},i}\beta) = \frac{1}{m}\beta'\Omega_X\beta && (43)\\
B &= E\left[\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\varepsilon_{\text{in}}'X_{\text{in}}(X_{\text{in}}'X_{\text{in}})^{-1}\bar{X}_{\text{out}}'\right] \\
&\quad -\left\{E\left[\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\right]\right\}^2 && (44)\\
&= E\left\{E\left[\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}\varepsilon_{\text{in}}'X_{\text{in}}(X_{\text{in}}'X_{\text{in}})^{-1}\bar{X}_{\text{out}}'|X\right]\right\} \\
&\quad -\left\{E\left\{E\left[\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'\varepsilon_{\text{in}}|X\right]\right\}\right\}^2 && (45)\\
&= E\left\{\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'E\left[\varepsilon_{\text{in}}\varepsilon_{\text{in}}'|X\right]X_{\text{in}}(X_{\text{in}}'X_{\text{in}})^{-1}\bar{X}_{\text{out}}'\right\} \\
&\quad -\left\{E\left\{\bar{X}_{\text{out}}(X_{\text{in}}'X_{\text{in}})^{-1}X_{\text{in}}'E\left[\varepsilon_{\text{in}}|X\right]\right\}\right\}^2 && (46)
\end{aligned}$$

Since $E\left[\varepsilon_{\text{in}}\varepsilon'_{\text{in}}|X\right] = \sigma_\varepsilon^2 I_n$ and $E\left[\varepsilon_{\text{in}}|X\right] = 0$, then

$$\begin{aligned} B &= \sigma_\varepsilon^2 E\left\{\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}}\right\} && (47)\\ C &= E\left[\bar{X}_{\text{out}}\beta\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}X'_{\text{in}}\varepsilon_{\text{in}}\right] \\ &\qquad -E\left[\bar{X}_{\text{out}}\beta\right]\cdot E\left[\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}X'_{\text{in}}\varepsilon_{\text{in}}\right] && (48)\\ &= 0 && (49)\\ V\left(\hat{\mu}_{U_{\text{out}}}\right) &= \frac{1}{m}\beta'\Omega_X\beta + \sigma_\varepsilon^2 E\left\{\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}}\right\}. && (50)\end{aligned}$$

It should be noted that $\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}} = O\left(\frac{1}{n}\right)$. Then, as $n \longrightarrow \infty$ and $m \longrightarrow \infty$, $V\left(\hat{\mu}_{U_{\text{out}}}\right) \longrightarrow 0$.

A.3 Confidence interval

In practice, this variance can be estimated as follows:

$$\hat{V}\left(\hat{\mu}_{U_{\text{out}}}\right) = \frac{1}{m}\hat{\beta}'_{\text{in}}\hat{\Omega}_X\hat{\beta}_{\text{in}} + \hat{\sigma}^2_{\varepsilon_{\text{in}}}\bar{X}_{\text{out}}(X'_{\text{in}}X_{\text{in}})^{-1}\bar{X}'_{\text{out}},$$

where $\hat{\Omega}_X$ is estimated on the whole sample rather than only on X_{out} to increase the precision.

B Extension of the methodologies to the Incremental Cost-Utility Ratio

B.1 Background: the incremental cost-utility ratio

B.1.1 Definition and estimation

In economic evaluations, an ICUR statistic in which a new therapy ($T = 1$) is compared with a standard therapy ($T = 0$) is defined by:

$$ICUR = \frac{\mu_C^1 - \mu_C^0}{\mu_U^1 - \mu_U^0}, \tag{51}$$

where μ is the true mean value of (subscripts) costs (C) and utility (U) for treatments number 1 and number 0. Since the true means corresponding to the theoretical population are not known, the ICUR can be estimated as follows, on the basis of data collected from two groups of patients, each undergoing one of the forms of therapy (group number 1, consisting of n^1 individuals, underwent treatment ($T = 1$) and group number 0, consisting of n^0 individuals [5], underwent treatment ($T = 0$)):

$$\widehat{ICUR} = \frac{\overline{C^1} - \overline{C^0}}{\overline{U^1} - \overline{U^0}} = \frac{\Delta\bar{C}}{\Delta\bar{U}}, \tag{52}$$

where $\overline{C^1}$, $\overline{C^0}$ are the sample mean of the costs and $\overline{U^1}$, $\overline{U^0}$ in the two treatment arms are the sample mean of utility.

B.1.2 Assumptions and statistical properties

Assumption 3: Utility distribution

It is assumed that the utilities of treatment $T = 0, 1$ follow independent random variables with $[0, 1]$ support and with distribution D_U^T as follows:

$$U_i^T \sim D_U^T(\mu_U^T, {\sigma_U^T}^2), \tag{53}$$

where i denotes the individuals.

Since the support is finite, all the moments of the distribution are finite.

Assumption 4: Mapping

It is assumed that the utility measure can be explained by some variables X.

$$U_i^T = f(X_i^T, \varepsilon_i^T; \beta^T), \tag{54}$$
$$\varepsilon_i^T \sim N(0, \sigma_\varepsilon^2), \tag{55}$$

where f is a function that depends on a parameter vector β^T, and ε_i^T independent of X^T. It should be noted that β^T is specific to the treatment.

Let us consider an "in-sample" where the utility measure is known be denoted: $i = 1, \ldots, n^T$, and an "out-of-sample" where the utility is unknown be denoted: $i = n^T + 1, \ldots, n^T + m^T$. The aim of mapping is to assess the out-of-sample mean utility:

$$\mu_{U^T_{\text{out}}} = E(U^T_{\text{out}}) = E(f(X^T_{i,\text{out}}, \varepsilon^T_{i,\text{out}}; \beta^T)).$$

[5] n^1 is generally different from n^0.

Assumption 5: Cost distribution

It is assumed that the costs of treatment $T = 0, 1$ follow independent random variables with $(0, +\infty)$ support and with distribution D_C^T as follows:

$$C_i^T \sim D_C^T(\mu_C^T, {\sigma_C^T}^2), \tag{56}$$

where ${\sigma_C^T}^2 < \infty$. i denotes the individuals.

Assumption 6: Cost-Utility link

It is assumed that the utility measure and the cost are correlated for individual i and treatment T:

$$Cov(U_i^T, C_i^T) = \sigma_{UC}^T \tag{57}$$

Although the data in question do not follow normal distributions, we can generally apply the Central Limit Theorem (CLT), partly thanks to the fact that each sequence of pairs of random variables $(C_i^1, U_i^1)_{i=1,\cdots,n^1}$, $(C_i^0, U_i^0)_{i=1,\cdots,n^0}$ is independent and identically distributed (because the data were obtained in a randomised trial). Therefore, $\overline{C^1}$, $\overline{C^0}$, $\overline{U^1}$ and $\overline{U^0}$ are asymptotically normally distributed, and the same applies for $\Delta\bar{C}$ and $\Delta\bar{E}$ as the difference between normally distributed variables. [6]

B.1.3 Linear approximation of the mapping

Utility modeling and estimate

If a first order approximation of the function f in Equation 54 is computed, it can be assumed that U^T is approximated by:

$$U_i^T = X_i^T \beta^T + \varepsilon_i^T, \tag{58}$$

where the constant term is assumed to be contained in X^T. Thus,

$$\mu_{U^T_{\text{out}}} = E(X^T_{i,\text{out}})\beta^T.$$

The estimator usually used is:

$$\hat{\mu}_{U^T_{\text{out}}} = \frac{1}{m^T} \sum_{i=n^T+1}^{n^T+m^T} \hat{U}_i^T = \frac{1}{m^T} \sum_{i=n^T+1}^{n^T+m^T} X_i^T \hat{\beta}^T_{\text{in}} = \bar{X}^T_{\text{out}} \hat{\beta}^T_{\text{in}} \tag{59}$$

where $\hat{U}_i$ are the out-of-sample predicted values for the utility measure, $\hat{\beta}^T_{\text{in}} = (X^{T\prime}_{\text{in}} X^T_{\text{in}})^{-1} X^{T\prime}_{\text{in}} U^T_{\text{in}}$ the in-sample estimate of β^T, and $\iota_m = (1, \ldots, 1)'$. It should be noted that:

$$\hat{\mu}_{U^T_{\text{out}}} = \frac{1}{m^T} \iota'_m X^T_{\text{out}} \beta^T + \frac{1}{m^T} \iota'_m X^T_{\text{out}} (X^{T\prime}_{\text{in}} X^T_{\text{in}})^{-1} X^{T\prime}_{\text{in}} \varepsilon^T_{\text{in}}. \tag{60}$$

Then, we have the following properties: $E(\hat{\mu}_{U^T_{\text{out}}}) = E(X^T_{\text{out}})\beta^T$. The estimator is unbiased, and can be used to assess the mean utility. However, to take a decision, the uncertainty has to be accounted for. The problem, which will be handle in subsection B.2, is to estimate the variance of $\hat{\mu}_{U^T_{\text{out}}}$.

[6] The estimated ICUR does not necessarily have a defined mean or a defined variance, mainly owing to the fact that the denominator of the ratio can be statistically close to zero. In this case, the estimated ICUR will be very large (so that it is statistically close to infinite) or indeterminate, depending on whether $\Delta\bar{C}$ is also statistically close to zero, and the distribution of the estimated ICUR will be close to a Cauchy distribution (whose moments are infinite).

Cost estimate

The out-of-sample mean cost can be estimated as follows:

$$\bar{C}^T_{\text{out}} = \frac{1}{m^T} \sum_{i=n^T+1}^{n^T+m^T} C^T_i. \tag{61}$$

Under Assumption 6, and applying the Central Limit Theorem (CLT) as $n^T \longrightarrow \infty$, we obtain:

$$\bar{C}^T_{\text{out}} \sim N\left(\mu^T_C, \frac{1}{n^T}\sigma^{T2}_C\right). \tag{62}$$

The variance of C^T_i can be estimated as follows:

$$\left(\hat{\sigma}^T_{C_{\text{out}}}\right)^2 = \frac{1}{m^T} \sum_{i=n^T+1}^{n^T+m^T} \left(C^T_i - \bar{C}^T_{\text{out}}\right). \tag{63}$$

Utility-Cost link modeling and estimate

$$\sigma^T_{UC} = \text{Cov}(X^T_{i1}, C^T_i)\beta_1 + \cdots + \text{Cov}(X^T_{iK}, C^T_i)\beta_K + \text{Cov}(\varepsilon^T_i, C^T_i), \tag{64}$$

where K is the number of explanatory variables, X_0 corresponds to the constant term. Let us denote $\text{Cov}(X^T_{ik}, C^T_i) = \gamma^T_k$, $(\gamma^T_1, \ldots, \gamma^T_K) = \gamma^T$, and $\text{Cov}(\varepsilon^T_i, C^T_i) = \gamma_\varepsilon$. It should be noted that $\text{Cov}(\varepsilon^T_i, C^T_i)$ is not necessarily equal to 0, since among two individuals that have the same characteristics, if one has a random problem that causes its health status, it will also causes the corresponding cost. The covariance vector can be estimated as follows:

$$\hat{\gamma}^T = \frac{1}{m^T} \sum_{i=n^T+1}^{m^T} (X^T_i - \iota_{m^T}\bar{X}^T)'(C_i - \bar{C}), \tag{65}$$

where $\iota_{m^T} = (1, \ldots, 1)$, $\bar{X}^T = \frac{1}{m^T}\sum_{i=n^T+1}^{m^T} X^T_i$, and $\bar{C} = \frac{1}{m^T}\sum_{i=n^T+1}^{m^T} C_i$.

B.2 Confidence region for the ICUR

B.2.1 Case of no mapping: Fieller's method

Why Fieller's method?

Siani and Moatti [2003] analyzed all the method of the literature for calculating a confidence region for the incremental cost effectiveness ration (ICER): box method, delta method, ellipse method, Fieller's method, classical bootstrap method, and "re-ordered" bootstrap method. They found that the only two methods that are reliable are Fieller's method and the "re-ordered" bootstrap method. Siani et al. [2004], Siani and de Peretti [2010] then focused on the performance of Fieller's and "reordered" 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. Their Monte Carlo simulations show that the non-reordered bootstrap method performs worse than Fieller's method in these problematic cases. They also confirm that the reordered bootstrap method and Fieller's method have similar performance

most of the time. Nevertheless, their Monte Carlo simulations show that in any extreme cases Fieller's method performs significantly better than reordered bootstrap method. [7]

When no mapping is used to assess the utility measure, the statistical properties of the ICUR are similar to the ones of the ICER. Thus Fieller's method should also over-perform the other methods for the ICUR.

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. The method involves calculating confidence regions using the pivotal function technique, which consists of solving a second degree equation for the ICER. We briefly recall the general context of Fieller's theorem Fieller [1954] (see also Heitjan [2000]). Fieller's theorem has been considered in the univariate case Fieller [1940b,a], Finney [1978] and in the multivariate case Volund [1980], Zerbe et al. [1982], Laska et al. [1985]. It is assumed here that X_1 and X_2 are two random normally distributed variables such that:

$$X \sim N(\eta, \Omega) \text{ with } X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} , \tag{66}$$

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:

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

Therefore, we have:

$$\frac{Z^2}{\omega_1^2 + \rho^2\omega_2^2 - 2\rho\,\omega_{12}} \sim \chi^2(1), \tag{67}$$

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

where $k_{1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom. To find the $(1-\alpha)$ confidence region for $\frac{\eta_1}{\eta_2}$, the following inequation must be solved:

$$Q(\rho) \leq 0, \tag{68}$$

where

$$Q(\rho) = x\rho^2 + y\rho + z, \tag{69}$$

with $x = X_2^2 - k_{1-\alpha}\omega_2^2$, $y = 2(k_{1-\alpha}\omega_{12} - X_1X_2)$ and $z = X_1^2 - k_{1-\alpha}\omega_1^2$. If the variances and covariances are unknown (for example in the case of small sample size), they can be replaced by their estimators, in which case $k_{1-\alpha}$ is interpreted as the $(1-\alpha)$ quantile of an F distribution with the appropriate degrees of freedom.

[7] 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, these studies do not clearly mention whether they deal with reordered or non-reordered bootstrap methods and we will clarify this point. In addition, in these studies, the number of simulations used was insufficient to provide a great accuracy, the simulated data dealt with were often configured in such a way that they were sufficiently far from problematic cases, the miscoverage was measured calculating the mean miscoverage on several Data Generating Process, giving the impression that all the methods have similar performance, and consequently, no reliable conclusion can be drawn from these results.

Application to the ICER

We assume that (C^T, E^T) are vector random variables with mean (μ_C^T, μ_E^T), variance $\left((\sigma_C^T)^2, (\sigma_E^T)^2\right)$ and covariance σ_{CE}^T for $T = 0, 1$. The variables used in Fieller's method correspond to the following values:

$$\begin{aligned} X_1 &= \Delta\bar{C}, \\ X_2 &= \Delta\bar{E}, \\ \omega_1^2 &= {\sigma_C^0}^2/n^0 + {\sigma_C^1}^2/n^1, \\ \omega_2^2 &= {\sigma_E^0}^2/n^0 + {\sigma_E^1}^2/n^1, \\ \omega_{12} &= \sigma_{CE}^0/n^0 + \sigma_{CE}^1/n^1. \end{aligned}$$

After solving the inequation 68, $(1-\alpha)$ confidence regions for the ICER can have different forms. Depending on the sign of x, defined in Equation 69, and depending on the sign of the discriminant Δ of the polynomial function Q, the various forms of the confidence region obtained with Fieller's method are shown in Table 5. R^L and R^U are the roots of

Table 5: Form of the confidence region

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

the polynomial function Q, given by the following formulas:

$$R^L = \frac{X_1X_2 - k_{1-\alpha}\omega_{12} - \sqrt{(k_{1-\alpha}\omega_{12} - X_1X_2)^2 - (X_2^2 - k_{1-\alpha}\omega_2^2)(X_1^2 - k_{1-\alpha}\omega_1^2)}}{X_2^2 - k_{1-\alpha}\omega_2^2}, \quad (70)$$

$$R^U = \frac{X_1X_2 - k_{1-\alpha}\omega_{12} + \sqrt{(k_{1-\alpha}\omega_{12} - X_1X_2)^2 - (X_2^2 - k_{1-\alpha}\omega_2^2)(X_1^2 - k_{1-\alpha}\omega_1^2)}}{X_2^2 - k_{1-\alpha}\omega_2^2}. \quad (71)$$

If $x > 0$, 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 in which $\mu_{\Delta E}$ 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). A detailed analyze of all the cases can be get under request to the authors.

B.2.2 Case of mapping: various proposals

"Naive" confidence region

A naive, and wrong, way to compute the variance of $\hat{\mu}_{U^T_{\text{out}}}$ would be:

$$\widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U^T_{\text{out}}}\right) = \frac{1}{m^T}\left[\frac{1}{m^T-1}\sum_{i=n^T+1}^{n^T+m^T}\left(\hat{U}_i^T - \hat{\mu}_{U^T_{\text{out}}}\right)^2\right], \tag{72}$$

and the covariance between $\hat{\mu}_{U^T_{\text{out}}}$ and $\bar{C}^T_{\text{out}}$:

$$\widehat{Cov}^{\text{Naive}}\left(\hat{\mu}_{U^T_{\text{out}}}, \bar{C}^T_{\text{out}}\right) = \frac{1}{m^T}\left[\frac{1}{m^T-2}\sum_{i=n^T+1}^{n^T+m^T}\left(\hat{U}_i^T - \hat{\mu}_{U^T_{\text{out}}}\right)\left(C_i^T - \bar{C}^T\right)\right]. \tag{73}$$

The standard deviation in Equation 72 seems to be used in Rivero-Arias et al. [2009] to compute the confidence interval for the mean utility. A naive way to provide a $(1-\alpha)$-confidence region for the ICUR is to provide the following moments to the Fieller's method:

$$\begin{aligned}
X_1 &= \Delta\bar{C} = \bar{C}^1_{\text{out}} - \bar{C}^0_{\text{out}}, && (74)\\
X_2 &= \Delta\hat{U} = \hat{\mu}_{U^1_{\text{out}}} - \hat{\mu}_{U^0_{\text{out}}}, && (75)\\
\hat{\omega}_1^2 &= \left(\hat{\sigma}^1_{C_{\text{out}}}\right)^2/n^1 + \left(\hat{\sigma}^0_{C_{\text{out}}}\right)^2/n^0, && (76)\\
\hat{\omega}_2^2 &= \widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U^1_{\text{out}}}\right) + \widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U^0_{\text{out}}}\right), && (77)\\
\hat{\omega}_{12} &= \widehat{Cov}^{\text{Naive}}\left(\hat{\mu}_{U^1_{\text{out}}}, \bar{C}^1_{\text{out}}\right) + \widehat{Cov}^{\text{Naive}}\left(\hat{\mu}_{U^0_{\text{out}}}, \bar{C}^0_{\text{out}}\right). && (78)
\end{aligned}$$

The confidence region for the ICUR is then given by Table 5, Equation 70, and Equation 71 when using the moments above.

Analytic confidence region

In the case of a linear approximation, the variance of $\hat{\mu}_{U_{\text{out}}}$ can be estimated as follows:

$$\hat{V}\left(\hat{\mu}_{U^T_{\text{out}}}\right) = \frac{1}{m^T}\hat{\beta}^{T\prime}_{\text{in}}\hat{\Omega}_{X^T_{\text{out}}}\hat{\beta}^T_{\text{in}} + \hat{\sigma}^2_{\varepsilon^T_{\text{in}}}\bar{X}^T_{\text{out}}(X^{T\prime}_{\text{in}}X^T_{\text{in}})^{-1}\bar{X}^{T\prime}_{\text{out}}. \tag{79}$$

The proof can be get under request to the authors. It should be noted that

$$\hat{V}\left(\hat{\mu}_{U^T_{\text{out}}}\right) = \widehat{V}^{\text{Naive}}\left(\hat{\mu}_{U^T_{\text{out}}}\right) + \hat{\sigma}^2_{\varepsilon^T_{\text{in}}}\bar{X}^T_{\text{out}}(X^{T\prime}_{\text{in}}X^T_{\text{in}})^{-1}\bar{X}^{T\prime}_{\text{out}}. \tag{80}$$

In Equation 72, the term $\hat{\sigma}^2_{\varepsilon^T_{\text{in}}}\bar{X}^T_{\text{out}}(X^{T\prime}_{\text{in}}X^T_{\text{in}})^{-1}\bar{X}^{T\prime}_{\text{out}}$ is missing. The covariance between $\hat{\mu}_{U^T_{\text{out}}}$ and $\bar{C}^T_{\text{out}}$ can be estimated as follows:

$$\widehat{Cov}\left(\hat{\mu}_{U^T_{\text{out}}}, \bar{C}^T_{\text{out}}\right) = \frac{1}{m^T}\hat{\gamma}^{T\prime}_{\text{out}}\hat{\beta}^T_{\text{in}}. \tag{81}$$

The proof can be get under request to the authors. The confidence region for the ICUR is then given by Table 5, Equation 70, and Equation 71 when using the following moments:

$$X_1 = \Delta\bar{C} = \bar{C}^1_{\text{out}} - \bar{C}^0_{\text{out}}, \tag{82}$$

$$X_2 = \Delta\hat{U} = \hat{\mu}_{U^1_{\text{out}}} - \hat{\mu}_{U^0_{\text{out}}}, \tag{83}$$

$$\hat{\omega}_1^2 = \left(\hat{\sigma}^1_{C_{\text{out}}}\right)^2 / n^1 + \left(\hat{\sigma}^0_{C_{\text{out}}}\right)^2 / n^0, \tag{84}$$

$$\hat{\omega}_2^2 = \widehat{V}\left(\hat{\mu}_{U^1_{\text{out}}}\right) + \widehat{V}\left(\hat{\mu}_{U^0_{\text{out}}}\right), \tag{85}$$

$$\hat{\omega}_{12} = \widehat{Cov}\left(\hat{\mu}_{U^1_{\text{out}}}, \bar{C}^1_{\text{out}}\right) + \widehat{Cov}\left(\hat{\mu}_{U^0_{\text{out}}}, \bar{C}^0_{\text{out}}\right). \tag{86}$$

These analytic confidence region is restricted to a linear framework with Gaussian error terms. It can easily be extended to a nonlinear framework, using an Edgeworth expansion, but it will be an approximation. Consequently, we prefer to propose a bootstrap methodology, which will account for nonlinear specification, which may be used (such as logistic specification for instance). In addition, using nonparametric bootstrap, error terms with non-Gaussian distribution can also be accounted for.

Nonparametric bootstrap confidence region

In this subsection, we propose a methodology for building a confidence region based on the nonparametric bootstrap technique to compute the moments of the estimators. For a general presentation of the percentile-t method, see Hall [1992], Davidson and MacKinnon [1993], Efron and Tibshirani [1993], Hjorth [1994], and Shao and Tu [1995]. A mapping model is chosen:

$$U_i^T = f(X_i^T, \varepsilon_i^T; \beta^T), \tag{87}$$

$$C_i^T = g(U_i^T, \nu_i^T; \beta_C^T), \tag{88}$$

where X^T is a regressor matrix, and the functions f and g are known and are parametric in the sense that they depend on a parameter vector β^T or β_C^T. ε_i^T and ν_i^T are not assumed to be Gaussian. $V(\varepsilon_i^T) = (\sigma_\varepsilon^T)^2$, and $V(\nu_i^T) = (\sigma_\nu^T)^2$. The confidence region is built as follows:

1. Equation 87 and Equation 88 are estimated, providing $\hat{\beta}^T_{\text{in}}$, $\hat{\varepsilon}^T_{\text{in}}$, $(\hat{\sigma}_\varepsilon^T)^2$, $\hat{\beta}^T_{C,\text{in}}$, $\hat{\nu}^T_{\text{in}}$, and $(\hat{\sigma}_\nu^T)^2$.

2. A bootstrap Data Generating Process (DGP) has to be defined. It may be either parametric or semiparametric, characterized by $\hat{\beta}_{\text{in}}$, $\hat{\beta}^T_{C,\text{in}}$, and by any other relevant estimates that may be needed. In a general case, we propose:

$$U_i^{T,b} = f(X_i^{T,b}, \varepsilon_i^{T,b}; \hat{\beta}^T_{\text{in}}), \tag{89}$$

$$C_i^{T,b} = g(U_i^{T,b}, \nu_i^{T,b}; \hat{\beta}^T_{C,\text{in}}), \tag{90}$$

$$X_i^{T,b} \sim \text{ i.i.d. uniform distribution on } X_i^T, \tag{91}$$

for $T = 1, 0$ and $i = 1, \ldots, n^T$. The distribution of ε_i^b will be discussed later. f and g have to be chosen. For simplicitys sake, a linear model is chosen here:

$$U_i^{T,b} = X_i^{T,b}\hat{\beta}^T_{\text{in}} + \varepsilon_i^{T,b}, \tag{92}$$

$$C_i^{T,b} = U_i^{T,b}\hat{\beta}^T_{C,\text{in}} + \nu_i^{T,b}, \tag{93}$$

where X^T is assumed to contain the constant term, but a more specific nonlinear model can be chosen in practice in accordance with the data.

3. B bootstrap samples are generated:

$$(X_i^{T,b})_{i=1}^{n^T+m^T}, (U_i^{T,b})_{i=1}^{n^T}, (C_i^{T,b})_{i=1}^{n^T+m^T},$$

for $T = 1, 0$ and $b = 1, \ldots, B$.

4. For each of these samples, we compute $\hat{\beta}_{\text{in}}^T$ -denoted $\hat{\beta}_{\text{in}}^{T,b}$-, $\Delta\hat{U}^b = \hat{\mu}_{U^{1,b}_{\text{out}}} - \hat{\mu}_{U^{0,b}_{\text{out}}} = \bar{X}_{\text{out}}^{1,b}\hat{\beta}_{\text{in}}^{1,b} - \bar{X}_{\text{out}}^{0,b}\hat{\beta}_{\text{in}}^{0,b}$ and $\Delta\bar{C}^b = \bar{C}_{\text{out}}^{1,b} - \bar{C}_{\text{out}}^{0,b}$.

5. The variance-covariance matrix of $X_1 = \Delta\bar{C} = \bar{C}_{\text{out}}^1 - \bar{C}_{\text{out}}^0$ and $X_2 = \Delta\hat{U} = \hat{\mu}_{U^1_{\text{out}}} - \hat{\mu}_{U^0_{\text{out}}}$ is then computed as follows:

$$\hat{\omega}_1^2 = \frac{1}{B}\sum_{b=1}^{B}\left(\Delta\bar{C}^b - \Delta\bar{C}\right)^2, \tag{94}$$

$$\hat{\omega}_2^2 = \frac{1}{B}\sum_{b=1}^{B}\left(\Delta\hat{U}^b - \Delta\hat{U}\right)^2, \tag{95}$$

$$\hat{\omega}_{12} = \frac{1}{B}\sum_{b=1}^{B}\left(\Delta\bar{C}^b - \Delta\bar{C}\right)\left(\Delta\hat{U}^b - \Delta\hat{U}\right). \tag{96}$$

6. The confidence region is obtained by applying Fieller's method to the moments computed using bootstrap techniques.

We consider the following way of generating the bootstrap residuals $\varepsilon_i^{T,b}$ and $\nu_i^{T,b}$ (see Weber Weber [1984]). The $\varepsilon_i^{T,b}$ are generated by independent uniform draws with replacement among the vector with the typical element $\tilde{\varepsilon}_i^T$ constructed as follows:

1. Calculate $(P_{X^T})_{i,i}$, $i = 1, \ldots, n^T$, the diagonal elements of the projection matrix on X^T.
2. Calculate $\frac{\hat{\varepsilon}_i^T}{\sqrt{1-(P_{X^T})_{i,i}}}$, $\forall i = 1, \ldots, n^T$.
3. Recenter the vector that results.
4. Rescale it so that it has the variance $(\hat{\sigma}_\varepsilon^T)^2$.

This permits to correct the heteroskedasticity in the residuals due to the regressors. The same procedure is applied for $\nu_i^{T,b}$.

B.3 Performance of the methods: Monte Carlo experiments

Data Generating Process (DGP) are use to generate simulated data samples. The methods are applied to each simulated sample $j = 1, \ldots, S$, and it is examined if each confidence region j contains or not the true value $ICUR$ of the ratio (which is known,

since the DGP is known, conversely to real data). The coverage c of the confidence regions can be estimated as follow:

$$\hat{c} = \frac{1}{S} \sum_{j=1}^{S} I\left(\mu_U \in Interval_j\right). \tag{97}$$

The standard deviation of this Monte Carlo estimate of the coverage is $\sqrt{\frac{1}{S}c(1-c)}$, where c is the true coverage.

In our Monte Carlo experiments, we choose the confidence level $1-\alpha = 0.95$. The number of bootstrap replications is $B = 999$. The number of Monte Carlo replications is $S = 10,000$. If the true coverage $c = 0.95$, the standard deviation of the Monte Carlo estimate of the coverage is 0.002179. At most (where $c = 0.5$) the standard deviation is 0.005. Several values for n^T and m^T are chosen. Small values for n^T and large values for m^T are first allow to reflect the case where the utility is assessed only on a small subsample, and then extrapolated to the other patients.

B.3.1 Data Generating Process

A variety of DGP are proposed to check the robustness of the methods: linear, nonlinear, with non-Gaussian error terms.

Linear Case

$$\begin{aligned}
U_i^T &= \beta_0^T + \beta_1^T \cdot X_{1i}^T + \beta_2 \cdot X_{2i}^T + \varepsilon_i^T, && (98)\\
C_i^T &= \beta_{C,0}^T + \beta_{C,1}^T \cdot U_i^T + \nu_i^T, && (99)\\
X_{1i}^T &\sim i.i.d.U([0,1]), && (100)\\
X_{2i}^T &\sim i.i.d.B(0.5,3), && (101)\\
\varepsilon_i^T &\sim i.i.d.N(0,(\sigma_\varepsilon^T)^2), && (102)\\
\nu_i^T &\sim i.i.d.N(0,(\sigma_\nu^T)^2). && (103)
\end{aligned}$$

The parameters values are set to:
$\beta_0^1 = 0.2$, $\beta_1^1 = 0.7$, $\beta_2^1 = 0.3$, $\sigma_\varepsilon^1 = 0.15$, $\beta_{C,0}^1 = 0.35$, $\beta_{C,1}^1 = 0.5$, $\sigma_\nu^1 = 0.15$.
$\beta_0^0 = 0$, $\beta_1^0 = 0.5$, $\beta_2^0 = 0.1$, $\sigma_\varepsilon^0 = 0.15$, $\beta_{C,0}^0 = 0.35$, $\beta_{C,1}^0 = 0.5$, $\sigma_\nu^0 = 0.15$.
We have $E(U_i^1) = 1$, $E(C_i^1) = 0.85$, $E(U_i^0) = 0.4$, $E(C_i^0) = 0.55$, $ICUR = 0.5$.

Random Linear Case

The model is the same, but the parameters vary randomly across the Monte Carlo replications:

$$\begin{aligned}
\beta_0^1 \sim i.i.d.N(0.2,0.2^2) \quad &, \quad \beta_0^0 \sim i.i.d.N(0,0^2), && (104)\\
\beta_1^1 \sim i.i.d.N(0.7,0.7^2) \quad &, \quad \beta_1^0 \sim i.i.d.N(0.5,0.5^2), && (105)\\
\beta_2^1 \sim i.i.d.N(0.3,0.3^2) \quad &, \quad \beta_2^0 \sim i.i.d.N(0.1,0.1^2), && (106)\\
\sigma_\varepsilon^1 \sim i.i.d.U([0.15\cdot 0.5, 0.15\cdot 1.5]) \quad &, \quad \sigma_\varepsilon^0 \sim i.i.d.U([0.15\cdot 0.5, 0.15\cdot 1.5]), && (107)\\
\beta_{C,0}^1 \sim i.i.d.N(0.35,0.35^2) \quad &, \quad \beta_{C,0}^0 \sim i.i.d.N(0.35,0.35^2), && (108)\\
\beta_{C,1}^1 \sim i.i.d.N(0.5,0.5^2) \quad &, \quad \beta_{C,1}^0 \sim i.i.d.N(0.5,0.5^2), && (109)\\
\sigma_\nu^1 \sim i.i.d.U([0.15\cdot 0.5, 0.15\cdot 1.5]) \quad &, \quad \sigma_\nu^0 \sim i.i.d.U([0.15\cdot 0.5, 0.15\cdot 1.5]). && (110)
\end{aligned}$$

We have $E(U_i^T|\beta^T) = (1, 0.5, 1.5)\beta^T$, $E(C_i^T|\beta^T, \beta_C^T) = \left(1, E(U_i^T|\beta^T)\right)\beta_C^T$.

Non-Gaussian Random Linear Case

The model is the same as in Equation 98–Equation 101, but the error terms follow the uniform distribution:

$$\begin{aligned} \varepsilon_i^T &\sim i.i.d.U([-2, 2]) * \sigma_\varepsilon^T, && (111) \\ \nu_i^T &\sim i.i.d.U([-2, 2]) * \sigma_\nu^T. && (112) \end{aligned}$$

The parameters follow the same distributions as in Equation 104–Equation 110. Then, we also have $E(U_i^T|\beta^T) = (1, 0.5, 1.5)\beta^T$, $E(C_i^T|\beta^T, \beta_C^T) = \left(1, E(U_i^T|\beta^T)\right)\beta_C^T$.

Nonlinear Case

$$\begin{aligned} U_i^T &= F\left[\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \varepsilon_i\right] && (113) \\ C_i^T &= \left(\beta_{C,0}^T + U_i^T \beta_{C,1}^T\right) \chi^2(1) && (114) \end{aligned}$$

F is the cumulative distribution function of a normal variable. The parameters values are set to:
$\beta_0^1 = -0.5$, $\beta_1^1 = 0.7$, $\beta_2^1 = 0.3$, $\sigma_\varepsilon^1 = 0.3$, $\beta_{C,0}^1 = 0.25$, $\beta_{C,1}^1 = 0.5$.
$\beta_0^0 = -0.7$, $\beta_1^0 = 0.5$, $\beta_2^0 = 0.1$, $\sigma_\varepsilon^0 = 0.3$, $\beta_{C,0}^0 = 0.25$, $\beta_{C,1}^0 = 0.5$.
We have $E(U_i^1) = 0.61791142$, $E(C_i^1) = 0.55895571$, $E(U_i^0) = 0.38208858$, $E(C_i^0) = 0.44104429$, $ICUR = 0.5$.

B.3.2 Performance

The coverage of the various confidence regions, computed via Monte Carlo experiments for various samples sizes, are presented in Table 6. The sizes are chosen to correspond to a mapping assessment on a small subsample, and then an extrapolation of the utility values to the remaining sample of sizes that can be encountered in practice. The results show that the coverage of the "naive" confidence region is small (down to 60% depending on the sample sizes) with respect to the confidence level (95%), whereas both the analytic and bootstrap confidence regions perform correctly.

Table 6: Coverage and mean length of the 95% confidence intervals

		Coverage			Mean angle		
n^T	m^T	Naive	Analytic	Bootstrap	Naive	Analytic	Bootstrap
Linear Data Generating Process							
40	400	0.9239	0.9818	0.9818	0.0929	0.1309	0.1275
30	600	0.7892	0.9620	0.9567	0.0761	0.1322	0.1273
20	800	0.6659	0.9598	0.9471	0.0659	0.1501	0.1403
Random Linear Data Generating Process							
80	40	0.9593	0.9679	0.9685	0.4240	0.4538	0.4518
100	100	0.9417	0.9616	0.9616	0.2848	0.3104	0.3102
40	400	0.8348	0.9667	0.9605	0.1324	0.2337	0.2265
30	600	0.7495	0.9620	0.9523	0.1101	0.2520	0.2406
20	800	0.6349	0.9540	0.9437	0.0826	0.2519	0.2319
Non-Gaussian Random Linear DGP							
40	400	0.8000	0.9538	0.9495	0.1457	0.2666	0.2591
30	600	0.7057	0.9577	0.9498	0.1010	0.2624	0.2511
20	800	0.6112	0.9461	0.9381	0.0872	0.3170	0.2886
Nonlinear Data Generating Process							
40	400	0.9512	0.9580	0.9321	0.7086	0.7383	0.7164
30	600	0.9382	0.9559	0.9171	0.5827	0.6292	0.5989
20	800	0.9230	0.9483	0.9207	0.5119	0.5967	0.5608

n^T is the in sample size used to assess the mapping. m^T is the out of sample size where the utility values are predicted.

References

R. Ariza-Ariza, B. Hernández-Cruz, B. Carmona, M. D. Ruiz-Montesinos, J. Ballina, F. Navarro-Sarabia, The Costs, and Quality of Life in Rheumatoid Arthritis Study Group. Assessing utility values in rheumatoid arthritis: A comparison between time trade-off and the euroqol. *Arthritis and Rheumatism (Arthritis Care and Research)*, 55(5):751–756, October 15 2006. DOI 10.1002/art.22226, American College of Rheumatology. 1

N. Bansback, A. Brennan, and O. Ghatnekar. Cost-effectiveness of adalimumab in the treatment of patients with moderate to severe rheumatoid arthritis in sweden. *Ann. Rheum. Dis.*, 64:995–1002, 2005. 3

A. Beresniak, A.S. Russell, B. Haraoui, L. Bessette, C. Bombardier, and G. Duru. Advantages and limitations of utility assessment methods in ra. *Journal of Rheumatology*, 34(11):2193–2200, 2007. 3

R. Davidson and J. G. MacKinnon. *Estimation and inference in economics*. Oxford University Press, New York, 1993. 3.4, B.2.2

P. Dolan and M. Sutton. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Soc Sci Med*, 44(10):1519–1530, 1997. 1

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1993. 3.4, B.2.2

E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16:175–183, 1954. 7

E.C. Fieller. The biological standardization of insulin. *Journal of the Royal Statistical Society*, 137:1–53, 1940a. 7

E.C. Fieller. A fundamental formula in the statistics of the biological assay, and some applications. *Quarterly Journal of Pharmacy and Pharmacology*, 17:117–123, 1940b. 7

D.J. Finney. *Statistical methods in biological assay*. Macmillan, 1978. New York. 7

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York, 1992. 3.4, B.2.2

D.F. Heitjan. Fieller's method and net health benefits. *Health Economics*, 9:327–335, 2000. 7

J. S. U. Hjorth. *Computer Intensive Statistical Methods*. Chapman and Hall, London, 1994. 3.4, B.2.2

P. F. Krabbe, M. L. Essink-Bot, and G. J. Bonsel. The comparability and reliability of five health-state valuation methods. *Soc Sci Med*, 45(11):1641–1652, 1997. 1

E.M. Laska, H.B. Kushner, and M. Meisner. Reader reaction: multivariate bioassay. *Biometrics*, 41:547–554, 1985. 7

L. Longworth, M. Buxton, M. Sculpher, and D. H. Smith. Estimating utility data from clinical indicators for patients with stable angina. *Eur J Health Econom*, 6:347–353, 2005. 1, 3

S. Merkesdal, T Kirchhoff, D. Wolka, G. Ladinek, A. Kielhorn, and A. Rubbert-Roth. Cost-effectiveness analysis of rituximab treatment in patients in germany with rheumatoid arthritis after etanercept-failure. *Eur J Health Econ*, 11:95–104, 2010. DOI 10.1007/s10198-009-0205-y. 3, 3

E. Nord, J. Richardson, and K. Macarounas-Kirchmann. Social evaluation of health care versus personal evaluation of health states: Evidence on the validity of four health state scaling instruments using norvegian and australian survey data. Centre for Health Program Evaluation (CHPE), Working Paper 23, September 1992. 1

J. F. O'Leary, D. L. Fairclough, M. K. Jankowski, and J. C. Weeks. Comparison of time trade-off utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Med Decis Making*, 15(2):132–137, 1995. 1

O. Rivero-Arias, M. Ouellet, A. Gray, J. Wolstenholme, P. M. Rothwell, and R. Luengo-Fernandez. Mapping the modified rankin scale (mrs) measurement into the generic euroqol (eq-5d) health outcome. In *7th World Congress on Health Economics*, volume Harmonizing Health and Economics, Beijing, China, July 12–15 2009. International Health Economics Association. Poster presentation. 3, 3.2, B.2.2

J. A. Salomon and C. J. L. Murray. A multi-method approach to measuring health-state valuations. *Health Economics*, 13:281–290, 2004. Published online 20 June 2003 in Wiley InterScience (www.interscience.wiley.com). DOI:10.1002/hec.834. 3

J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995. 3.4, B.2.2

A. Shmueli. It might be premature to reject the assumption of a power curve relationship between vas and sg data: Three comments on stevens, mccabe and braziers mapping between vas and sg data; results from the uk hui index 2 valuation survey. *Health Economics*, 16(in press):755–758, 2007. Published online 18 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hec.1188. 2.5, 3

C. Siani and C. de Peretti. Are fieller's and bootstrap methods really equivalent for calculating confidence regions for ratios. *Health, Decision and Management*, forthcoming, 2010. B.2.1

C. Siani, C. de Peretti, and J. P. Moatti. Rehabilitating the confidence region for icers and which method to use: Fieller or re-ordered bootstrap? *Applied Health Economics and Health Policy*, 3(1):S62–S63, 2004. supplement. B.2.1

C. Siani and J.-P. Moatti. Quelles méthodes de calcul des régions de confiance du ratio coût-efficacité incrémental choisir ? *Revue d'Epidémiologie et de Santé Publique*, 51: 255–276, 2003. B.2.1

K. J. Stevens, C. J. McCabe, and J. E. Brazier. Mapping between visual analogue scale and standard gamble data; results from the uk health utilities index 2 valuation survey. *Health Econ., Health Economics Letters*, 15:527–533, 2006. Published online 3 January

2006 in Wiley InterScience (www.interscience.wiley.com). DOI:10.1002/hec.1076. 2.5, 3

G. W. Torrance. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Planning Sci*, 10(3):129–136, 1976. 1

G. W. Torrance, D. H. Feeny, W. J. Furlong, R. D. Barr, Y. Zhang, and Q. Wang. Multiattribute utility function for a comprehensive health status classification system. health utilities index mark 2. *Med Care*, 34(7):702–722, 1996. 1

A. Tsuchiya, J. Brazier, E. McColl, and D. Parkin. Deriving preference-based single indices from non-preference based condition-specific instruments: Converting aqlq into eq5d indices. Sheffield Health Economics Group, Discussion Paper Series Ref: 02/1; The University of Sheffield, ScHARR, School of Health and Related Research, May 2002. 2.1, 2.6

A. Volund. Multivariate bioassay. *Biometrics*, 36:225–236, 1980. 7

N.C. Weber. On resampling techniques for regression models. *Statistics and Probability Letters*, 2:275–278, 1984. 3.4, B.2.2

E.M. Zerbe, G.O.and Laska, M. Meisner, and H.B. Kushner. On multivariate confidence regions and simultaneous confidence limits for ratios. *Communications in Statistics, Theory and Methods*, 11:2401–2425, 1982. 7