UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

COMPARING PARAMETRIC AND SEMI-PARAMETRIC APPROACHES FOR BAYESIAN COST-EFFECTIVENESS ANALYSES IN HEALTH ECONOMICS

Caterina Conigliani and Andrea Tancredi

Working Paper n° 64, 2006

UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

Working Paper n° 64, 2006

- I “Working Papers” del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all’approvazione del Comitato Scientifico.

- Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall’art. 1 del D.L.L. 31.8.1945, n. 660 e successive modifiche.

- Copie della presente pubblicazione possono essere richieste alla Redazione.

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Ostiense, 139 - 00154 Roma
Tel. 0039-6-57374003 fax 0039-6-57374093
E-mail: dip_eco@uniroma3.it

UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

COMPARING PARAMETRIC AND SEMI-PARAMETRIC APPROACHES FOR BAYESIAN COST-EFFECTIVENESS ANALYSES IN HEALTH ECONOMICS

Caterina Conigliani* and Andrea Tancredi**

* *Università Roma Tre, Italy.*
** *Università di Roma "La Sapienza", Italy.*

CONTENTS

Abstract. *We consider the problem of assessing new and existing technologies for their cost-effectiveness in the case where data on both costs and effects are available from a clinical trial, and we address it by means of the cost-effectiveness acceptability curve. The main difficulty in these analyses is that cost data usually exhibit highly skew and heavy-tailed distributions, so that it can be extremely difficult to produce realistic probabilistic models for the underlying population distribution, and in particular to model accurately the tail of the distribution, which is highly influential in estimating the population mean. Here, in order to integrate the uncertainty about the model into the analysis of cost data and into cost-effectiveness analyses, we consider an approach based on Bayesian model averaging: instead of choosing a single parametric model, we specify a set of plausible models for costs and estimate the mean cost with its posterior expectation, that can be obtained as a weighted mean of the posterior expectations under each model, with weights given by the posterior model probabilities. The results are compared with those obtained with a semi-parametric approach that does not require any assumption about the distribution of costs.*

1 Introduction

The increasing burden on the budgets of health care providers has resulted in considerable interest in assessing new and existing technologies for their clinical effectiveness and cost-effectiveness.

Suppose that two health care technologies T_1 and T_2 are to be compared in a randomised controlled trial; data are direct measurements of effect and cost:

$$D = \left\{ x_{ij} = (e_{ij}, c_{ij})^T : \ i = 1, 2; \ j = 1, 2, ..., n_i \right\}$$

where e_{ij} and c_{ij} are the effect and the cost of treatment i on patient j respectively.

In order to assess if T_2 is more cost-effective than T_1, we need to compare expected effects γ_i and expected costs μ_i for each treatment. Let $\Delta_e = \gamma_2 - \gamma_1$ and $\Delta_c = \mu_2 - \mu_1$ be the effect and cost differentials. Moreover, let K be a decision-maker's *willingness to pay* coefficient, that is the units of money a decision maker is prepared to pay to obtain one unit of effectiveness.

The primary measure of cost-effectiveness of T_2 relative to T_1 is usually considered to be the *incremental cost-effectiveness ratio*, defined as $\rho = \Delta_c / \Delta_e$. However, as pointed out for instance in O'Hagan *et al.* (2000), cost-effectiveness of T_2 does not simply equate to ρ being less than K. It also depends on the sign of Δ_e, so

that it is the sign of the *net monetary benefit* $K\Delta_e - \Delta_c$ that is of interest: T_2 is cost-effective relative to T_1 if $K\Delta_e - \Delta_c > 0$, *i.e.* if in the plane of possible pairs of values of the population mean increments of effect and cost, (Δ_e, Δ_c) is below a sloping line of gradient K. This is usually referred as the *Net Benefit approach* (Stinnett and Mullahy, 1998), and inference about the net monetary benefit is generally presented by means of a Cost-Effectiveness Acceptability Curve (CEAC), that plots the probability $Q(K)$ that the net benefit is positive against the coefficient K (van Hout *et al.*, 1994), which is rarely unambiguously determined in practice. In this sense, a Bayesian approach is particularly natural, since no such probability exists or has any meaning in frequentist statistics (O'Hagan *et al.*, 2000). Thus, in the rest of the paper we assume that $Q(K)$ is the posterior probability

$$Q(K) = P(K\Delta_e - \Delta_c > 0 \mid D).$$

Clearly these cost-effectiveness analyses of clinical trial data rely on statistical models which describe the distribution of costs and effects and their interrelation across individual in the trial. The choice of models used in practice is often determined by convenience, and in particular it is often assumed that the data on costs and effects follow a bivariate normal distribution in each arm of the trial. However, although such assumption may be convenient for computational purposes, it is rarely realistic. In particular, cost data obtained for individual patients in health economic studies typically exhibit highly skew and heavy tailed distributions, and many problems arise with the various approaches currently available for analysing such data.

In fact, as discussed in O'Hagan and Stevens (2002, 2003), non-parametric methods, such as those based on the asymptotic normality of the sample mean or nonparametric bootstrapping, may be inefficient and their justification breaks down in small samples. See Dinh and Zhou (2006) for some recent developments on such methods. On the other hand, parametric modelling may lead to more efficient inference, but is dependent on the population distribution matching the model adequately. The main difficulty in this sense, as pointed out for instance in Nixon and Thompson (2004) and Thompson and Nixon (2005), is that the high skewness and kurtosis usually found in cost data imply that the population mean can be very sensitive to the tail of the distribution, that it is quite complicated to model accurately. One consequence of this is that parametric models that fit the data equally well can produce very different answers; conversely, in some cases models that fit badly can give

similar inferences to those that fit well. For these reasons, Thompson and Nixon (2005) recommend that the sensitivity of conclusions to the choice of the model is always investigated, so that model uncertainty becomes a crucial aspect of analysing cost data. They also suggest that, instead than allowing an arbitrary long tail for cost distributions, it might be more plausible to consider truncated distributions (for instance at twice the maximum observed cost), because costs for individual patients are bound to have some finite limit in practice. It is also interesting to note the paper by Briggs *et al.* (2005), where it is shown that (frequentist) inferences based on incorrect parametric assumptions, and in particular on the assumption of lognormality when the data come from a different distribution, can lead to totally misleading conclusions. Another problem related to the parametric modelling of costs concerns possible transformations of the data; in fact, as discussed in Thompson and Barber (2000) and Briggs and Gray (1998), mean values and confidence limits may be difficult to interpret on the transformed scales, and back-transformation onto the original scale is not always straightforward.

Here, in order to integrate the uncertainty about the model into the analysis of cost data and into cost-effectiveness analyses, we consider an approach based on Bayesian model averaging. This is presented in details in Section 2, and is compared with the semi-parametric approach of Conigliani and Tancredi (2005a, 2005b), that was introduced with the same aim of reporting model-based inference for mean costs without having to be too concerned about model misspecification problems. In Section 3 we consider the relative performance of the two approaches simulating cost data from a number of assumed parametric distributions. In Section 4 we repeat the comparison in an empirical context using a study on low back pain. A few concluding remarks are presented in the final section.

2 The model

In order to focus the attention on the distribution of costs, we find convenient to write the distribution for a single observation x_{ij} under treatment T_i as:

$$g_i\left(x_{ij}\,|\theta_i,\phi_i\right) = f_i\left(c_{ij}\,|\theta_i\right) h_i\left(e_{ij}\,|c_{ij},\phi_i\right)$$

where $f_i\left(c_{ij}\,|\theta_i\right)$ is the unconditional distribution for the cost of patient j under treatment T_i and $h_i\left(e_{ij}\,|c_{ij},\phi_i\right)$ is the conditional distribution for the effect on patient j under treatment T_i given the cost c_{ij}. In the remaining of this section we

will focus on the problem of specifying $f_i\left(c_{ij}\,|\theta_i\right)$; the choice of $h_i\left(e_{ij}\,|c_{ij},\phi_i\right)$ usually does not present too many problems, and will be addressed only in the example.

2.1 Bayesian model averaging

Suppose that under both treatment groups, instead of choosing a single parametric model for a cost observation c_{ij}, we specify a set of plausible models $\mathcal{M} = \{m = 1, 2, ..., M\}$. Also assume that the mean cost μ_i is an unknown parameter of all models in $\mathcal{M}$, and that aim of the analysis is obtaining the marginal posterior mean of μ_i. This is exactly a setting where Bayesian model averaging (BMA) can be applied to obtain the desired summary (Hoeting *et al.* 1999). In fact, the posterior marginal distribution of μ_i can be obtained as a mixture of its posterior marginal distributions under each of the models in $\mathcal{M}$:

$$\pi(\mu_i|c_{i1}, ..., c_{in_i}) = \sum_{m=1}^{M} \pi(\mu_i|c_{i1}, ..., c_{in_i}, m)\pi(m|c_{i1}, ..., c_{in_i}),$$

and the posterior expectation of μ_i can be expressed as a weighted mean of its posterior expectations under each model:

$$E(\mu_i|c_{i1}, ..., c_{in_i}) = \sum_{m=1}^{M} E(\mu_i|c_{i1}, ..., c_{in_i}, m)\pi(m|c_{i1}, ..., c_{in_i}),$$

with the mixing probabilities given by the posterior model probabilities $\pi(m|c_{i1}, ..., c_{in_i})$.

2.1.1 Prior assumptions and computational issues

Here we assume that the set $\mathcal{M}$ is made of the log-normal, the gamma, the Weibull, the log-logistic and the generalised Pareto distribution (GPD), not all necessarily having non-zero prior model probabilities; the probability density function, the mean and the coefficient of variation of these distributions are shown in Table 1. Moreover, we assume that the five distributions have finite mean and variance, so that for the GPD and the log-logistic we need to constrain the shape parameter to be less than 0.5. Then all models can be re-parametrized in terms of the mean cost μ_i and the coefficient of variation τ_i, and this is particularly useful when specifying the prior distributions for the parameters of the different models. In fact, it implies that the same prior distribution can be introduced under the various models in $\mathcal{M}$, and that

Table 1: Single parametric models included in the BMA procedure

	Pdf	**Mean**	**CV**
Log – normal	$p(x\,\vert\mu,\sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{[\log(x)-\mu]^2}{2\sigma^2}}$	$e^{\mu+\frac{\sigma^2}{2}}$	$\sqrt{e^{\sigma^2}-1}$
Gamma	$p(x\,\vert\nu,\lambda) = \frac{x^{\nu-1}\lambda^{\nu}}{\Gamma(\nu)} e^{-x\lambda}$	$\frac{\nu}{\lambda}$	$\frac{1}{\sqrt{\nu}}$
GPD	$p(x\,\vert\sigma,\xi) = \frac{1}{\sigma}\left(1+\frac{x\xi}{\sigma}\right)^{-1/\xi-1}$	$\frac{\sigma}{(1-\xi)}$ $(\xi<1)$	$\frac{1}{\sqrt{1-2\xi}}$ $(\xi<1/2)$
Weibull	$p(x\,\vert\beta,\delta) = \frac{\beta x^{\beta-1} e^{-(x/\delta)^{\beta}}}{\delta^{\beta}}$	$\frac{\delta}{\beta}\Gamma\left(\frac{1}{\beta}\right)$	$\sqrt{\frac{2\beta\Gamma\left(\frac{2}{\beta}\right)}{\Gamma^2\left(\frac{1}{\beta}\right)} - 1}$
Log – logistic	$p(x\,\vert\rho,\beta) = \frac{(x/\beta)^{1/\rho}}{x\rho\left[1+(x/\beta)^{1/\rho}\right]^2}$	$\pi\rho\beta\csc(\pi\rho)$ $(\rho<1)$	$\sqrt{\frac{\tan(\pi\rho)}{\pi\rho} - 1}$ $(\rho<1/2)$

the unknown parameters have a clear meaning, so that it is not difficult to elicit such prior from the experts' opinions. In particular, we assume that

$$\pi_i(\mu_i,\tau_i) = \pi_i(\mu_i\,\vert\tau_i)\,\pi_i(\tau_i)$$

where

$$\mu_i\,\vert\tau_i \sim N\left(\mu_i^0, (\tau_i\tau_i^0)^2\right)$$

$$\tau_i \sim \Gamma\left(\lambda_i^0, \nu_i^0\right)$$

and μ_i^0,τ_i^0, λ_i^0 and ν_i^0 are known constants.

Finally note that Bayesian inference for this model is possible using *Markov chain Monte Carlo* (MCMC) methods (see, for instance, Robert and Casella, 1999). In particular, one can simulate from the posterior distributions of the parameters under each of the models in $\mathcal{M}$, and then compute the marginal likelihoods required for obtaining the posterior model probabilities for instance with the method of Chib and Jeliazkov (2001). Alternatively, it is usually more efficient to use a single Markov Chain with Metropolis Hastings moves to explore both the parametric space within each model and to switch among models (see, for instance, O'Hagan and Forster, 2004, Chapter 10). Note that when all the models in $\mathcal{M}$ have the same number

of parameters, there is no need to apply the reversible jump methodology (Green, 2000), which anyway can be easily implemented to handle more general situations.

2.2 A semi-parametric approach

In recognising the extreme complexity of cost distributions (the construction of cost data as a weighted sum of different resource counts implies that cost distributions are really mixtures of many different distributions), Conigliani and Tancredi (2005a) suggested to model the bulk of the data and the tails separately. More specifically, for the problem of estimating the population mean cost differential, they introduce a distribution composed of a piecewise constant density up to an unknown endpoint, and a generalised Pareto distribution for the remaining tail data. The first component of the model, the step function, is very flexible, in the sense that it has the appealing property of catching all the relevant features of the data; if for instance the data exhibit multimodality, the corresponding model will be multimodal. However, the step function will hardly give any weight to values beyond the range of the data; for this reason, they introduced a different model for the upper tail of the distribution, the GPD, that is often used in extreme value theory to model tail data (Coles, 2001). Note this model has been applied to environmental data by Tancredi *et al.* (2002), and to cost-effectiveness analyses in the simple case where effects are measured as binary outcomes by Conigliani and Tancredi (2005b). Thus, for purposes of comparison with the approach of Section 2.1, here we assume this mixture model under both treatment groups, and write the density function for c_{ij} as

$$f_i(c_{ij}|\theta_i) = \begin{cases} (1-\omega)\sum_{i=1}^{s}\omega_i I_{[a_i,a_{i+1})}(c_{ij}) & 0 < c_{ij} < \alpha \\ \omega\,\dfrac{1}{\sigma}\left[1+\dfrac{\xi(c_{ij}-\alpha)}{\sigma}\right]^{-\frac{1}{\xi}-1} & \alpha \le c_{ij} < \infty \end{cases} \tag{1}$$

where ω is the probability that an observation c_{ij} is greater than α, s is the (unknown) number of steps of the piecewise constant density, $a^{(s)} = (a_2, ..., a_s)$ and $\omega^{(s)} = (\omega_1, ..., \omega_s)$ denote the vector of (unknown) step positions and the vector of (unknown) heights of the step function, α is both the (unknown) upper end point of the piecewise constant density and the threshold of the GPD, and σ and ξ are respectively the scale parameter and shape parameter of the GPD.

Note that, following Conigliani and Tancredi (2005a), we assume that the parameter vector $\theta_i = \left(s, \omega^{(s)}, a^{(s)}, \alpha, \sigma, \xi, \omega\right)$ varies with the treatment group i, and

that $\xi < 1$, so that the expected mean cost for treatment i is finite and can be written as

$$\mu_i = (1-\omega)\sum_{i=1}^{s} p_i \frac{a_{i+1}+a_i}{2} + \omega\left[\alpha + \frac{\sigma}{1-\xi}\right].$$

Moreover, in order to explore the use of truncated distributions for costs, as advocated by Thompson and Nixon (2005), here we consider the constraint $\xi > -1$; recall that the GPD density has an upper bound of $\alpha - \sigma/\xi$ on c_{ij} if $\xi \in (-1, 0)$, while it has no upper limit if $\xi \geq 0$.

Finally note that Bayesian inference for this model is possible using MCMC methods. Details of the algorithm can be found in Tancredi *et al.* (2002), while details about the prior distribution for the parameters can be found in Conigliani and Tancredi (2005a).

3 A simulation experiment

Before moving to cost-effectiveness analyses, we apply the two approaches presented in Section 2 to the simple problem of computing a 95% posterior credible interval for a mean cost, in order to compare their relative performance.

Thus, we generated 1000 samples of cost data and computed the credible intervals assuming either the model averaging procedure of Section 2.1 or the semi-parametric approach of Section 2.2. Under BMA we assumed prior model probabilities equal to zero for the Weibull and the log-logistic, and to 1/3 for the remaining models. Moreover, we completed the prior distributions for the parameters by letting $\mu_i^0 = 1000$, $\tau_i^0 = 500$, $\lambda_i^0 = 3$ and $\nu_i^0 = 3$, so that the prior mean of μ_i and τ_i were 1000 and 1 respectively, and the 90% of the prior distribution for τ_i was between 0.25 and 2. Finally, under the mixture model we assumed non-informative priors for most parameters, including ξ (Conigliani and Tancredi, 2005a). The results, presented in Table 2 and Table 3 in terms of coverage and average size of the intervals, are compared also with those obtained with the single parametric models considered by BMA, *i.e.* the log-normal, the gamma and the GPD. We also computed the proportion of the times that each of these distributions was the preferred model (for having the highest posterior probability); these are the numbers in parentheses in Table 2 and 3.

As data generating processes we considered a log-normal distribution with $\mu = 1000$ and $\tau = 2$, a Weibull distribution with $\mu = 1000$ and $\tau = 1.5$, a mixture

of a gamma distribution and a GPD, both with $\mu = 1000$ and $\tau = 2$, and with weights 1/8 and 7/8 respectively, and a mixture of three log-normal distributions with $\mu = 1000$, $\tau = 0.5, 2, 4$ and weights 0.61, 0.20, 0.19 respectively. The resulting distributions are plausible representations of cost data. In particular, the log-normal experiment is an example of the different performance of the two approaches when the set of models $\mathcal{M}$ specified for BMA includes the distribution that generated the data. Instead, in the remaining experiments none of the parametric models in $\mathcal{M}$ is *right*, and the BMA results are determined by all the models that have a non-zero posterior probability.

Samples of three different sizes ($n = 50, 200, 500$) were drawn from each distribution; this represented a total of 12 different simulation experiments. However we did not apply the mixture model (1) when $n = 50$, since we felt there were not enough observations to estimate its parameters. In fact, in the applications of extreme value theory it is common practice to use between the 5% and the 10% of the sample to estimate the upper tail of the model.

One point raises clearly from the results of Table 2 and Table 3, *i.e.* the posterior credible intervals obtained with the mixture model (1) are wider than the

Table 2: A simulation experiment: coverage of the posterior credible intervals (in parentheses the proportion of the times that each single parametric model was the preferred one)

Generating process	**n**	Log-normal		Gamma		GPD		BMA	Mixture
Log-normal	**50**	0.95	(0.660)	0.81	(0.008)	0.92	(0.332)	0.94	-
	200	0.95	(0.945)	0.75	(0.000)	0.92	(0.055)	0.95	0.94
	500	0.95	(0.998)	0.77	(0.000)	0.94	(0.002)	0.94	0.95
Weibull	**50**	0.93	(0.023)	0.92	(0.725)	0.92	(0.262)	0.93	-
	200	0.21	(0.001)	0.93	(0.914)	0.94	(0.085)	0.94	0.96
	500	0.01	(0.001)	0.91	(0.965)	0.95	(0.034)	0.92	0.95
Γ -GPD mixture	**50**	0.84	(0.019)	0.88	(0.703)	0.91	(0.278)	0.92	-
	200	0.09	(0.000)	0.86	(0.676)	0.93	(0.324)	0.90	0.95
	500	0.00	(0.000)	0.85	(0.710)	0.95	(0.290)	0.90	0.95
Log-normal mixture	**50**	0.94	(0.117)	0.86	(0.387)	0.93	(0.496)	0.89	-
	200	0.76	(0.034)	0.77	(0.529)	0.89	(0.437)	0.85	0.93
	500	0.39	(0.006)	0.75	(0.453)	0.89	(0.541)	0.85	0.94

Table 3: A simulation experiment: average size of the posterior credible intervals (in parentheses the proportion of the times that each single parametric model was the preferred one)

Generating process	**n**	Log-normal		Gamma		GPD		BMA	Mixture
Log-normal	**50**	889	(0.660)	643	(0.008)	696	(0.332)	801	-
	200	465	(0.945)	322	(0.000)	379	(0.055)	459	1841
	500	297	(0.998)	203	(0.000)	259	(0.002)	297	561
Weibull	**50**	1401	(0.023)	757	(0.725)	722	(0.262)	782	-
	200	1049	(0.001)	373	(0.914)	392	(0.085)	380	786
	500	823	(0.001)	234	(0.965)	265	(0.034)	238	371
Γ -GPD mixture	**50**	1657	(0.019)	778	(0.703)	697	(0.278)	791	-
	200	1388	(0.000)	375	(0.676)	382	(0.324)	379	1134
	500	1210	(0.000)	234	(0.710)	261	(0.290)	242	449
Log-normal mixture	**50**	877	(0.117)	508	(0.387)	595	(0.496)	569	-
	200	459	(0.034)	260	(0.529)	308	(0.437)	290	2174
	500	288	(0.006)	164	(0.453)	192	(0.541)	181	1227

corresponding ones obtained with Bayesian model averaging, but they are generally better in terms of coverage. This is not surprising, and is related to the way the two methods deal with model uncertainty. In fact, while the semi-parametric approach of Section 2.2 does not require any assumption about the distribution of costs, and allows inference on the mean cost to take account of the uncertainty about the tail, BMA implies the specification of a set of plausible models, which reduce the model uncertainty. It follows that BMA will generally lead to smaller intervals, but the characteristics of these intervals significantly depends on which models were included in $\mathcal{M}$. In fact, the results of BMA are strongly related to those obtained with the single models which have a non-zero posterior probability, which are not necessarily models that fit the data well. It follows that if we can specify a set of models that include the one that generated the data or at least a good approximation of it, as it is the case for instance for the log-normal data, BMA will provide more precise inference on mean cost than the mixture model. Otherwise we expect the mixture model to perform better. And in fact when the data are generated from the gamma-GPD mixture or the mixture of log-normal distributions, the coverage of the intervals produced with the semi-parametric approach is significantly better

Table 4: Low back pain trial: sample descriptive statistics of X-ray costs and rMRI costs

	X ray	**rMRI**
sample size	166	162
mean	1515	2187
standard deviation	1747	3378
median	926	871
minimum	44	49
maximum	9111	20664
skewness $\left(\overline{\mu}_3/\sigma^3\right)$	1.8	2.7
kurtosis $\left(\overline{\mu}_4/\sigma^4\right)$	6.4	10.9

than the one obtained with BMA. And if n is large enough (for instance $n = 500$), at least for the gamma-GPD mixture, also the difference in terms of average size of the intervals produced by the two approaches is less notable.

Finally note that the results of Table 2 and Table 3 confirm the finding of Briggs *et al.* (2005) that inferences based on the incorrect assumption of lognormality perform very poorly: in particular for the larger datasets, the coverage of the intervals obtained with the log-normal distribution when the true distribution is not log-normal are disastrous. In this sense, it is interesting to see the paper by Royall and Tsou (2003), where it is shown that while if we assume for instance a Gamma distribution the object of inference continues to be the mean of the true generating process also when the model fails, if we assume the log-normal working model then what the likelihood represents evidence about when the model fails is not $E_f(c)$ but the quantity $exp(E_f(log(c)) + \frac{1}{2}var_f(log(c)))$.

4 Analysis of the Low Back Pain Trial data

We present an example using a study on low back pain (Jarvik *et al.*, 2003). A total of 380 patients (out of which 328 were included in the health economic evaluation) were randomised in a 1:1 ratio to investigation by standard X-ray investigation and rapid magnetic resonance imaging (rMRI), and were followed for 12 months. Aim of the trial was to investigate whether rMRI would allow better diagnosis and treatment, or lead to unnecessary treatment without improvement in symptoms. The primary clinical endpoint was the change from baseline of the modified Roland

back pain score (Patrick *et al.*, 1995), while the primary economic endpoint was the total health care cost (in US$). This dataset has been analyzed by Thompson and Nixon (2005) in order to illustrate how sensitive inference about cost-effectiveness is to the choice of the model for costs.

The data are shown in Figure 1. Under both treatments the effects are apparently well represented by a normal distribution, while the distribution of costs is clearly highly skew and heavy-tailed; this fact is confirmed also by the sample summaries shown in Table 4. In particular, for both treatment groups the standard deviations are large, indicating that the data are spread quite far around the mean, and the median cost is smaller than the mean, indicating positively skew data; this fact is confirmed also by the standard skewness statistic $\overline{\mu}_3/\sigma^3$. Finally, the kurtosis statistic $\overline{\mu}_4/\sigma^4$ indicates that the two distribution of costs are significantly leptokurtic.

We now apply the approaches of Section 2.1 and Section 2.2 to this data set, and compare the results also with those that we obtain if we assume the single parametric models included in $\mathcal{M}$. Under BMA we assume that the prior model probabilities are zero for the Weibull and the log-logistic, and equal to 1/3 for the

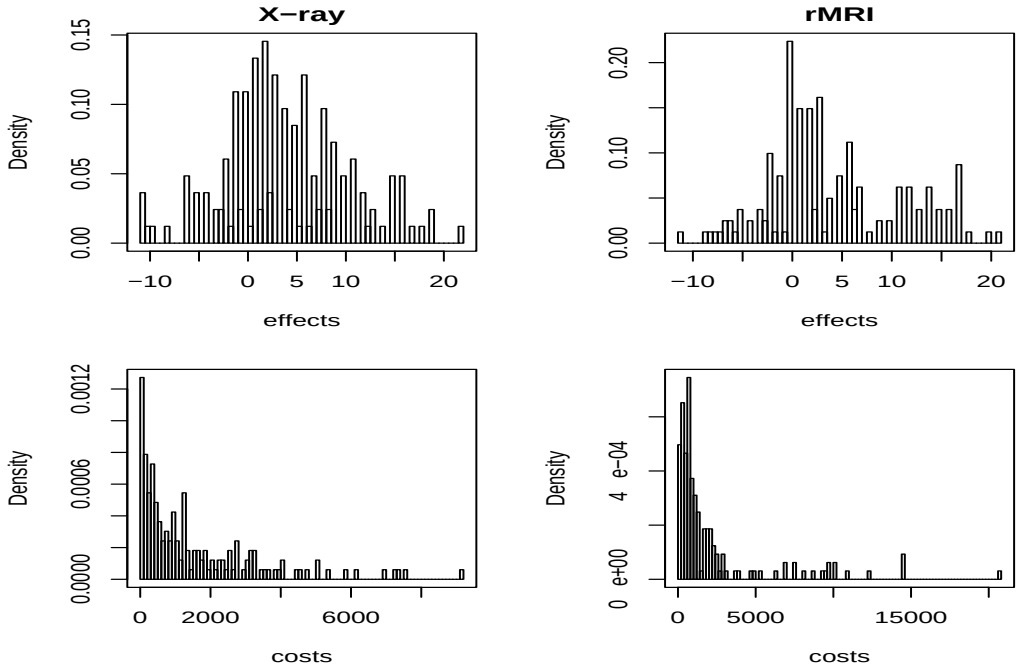

Figure 1: Data from the low back pain trial

Table 5: Low back pain trial: posterior summaries of mean costs

Model	$\mathbf{E}(\mu_1)$	$\mathbf{PCI_{0.95}}$	$\pi_1(\mathbf{m}\,|D)$	$\mathbf{E}(\mu_2)$	$\mathbf{PCI_{0.95}}$	$\pi_2(\mathbf{m}\,|D)$
Mixture	1555	1301; 1837	-	2377	1840; 3051	-
BMA	1542	1283; 1885	-	2047	1628; 2594	-
Log-normal	1852	1402; 2476	0.01	2051	1630; 2592	0.99
Gamma	1511	1275; 1785	0.49	2102	1784; 2479	0
GPD	1565	1285; 1930	0.50	2076	1677; 2580	0.01

remaining models. Moreover, the prior distributions for the unknown parameters are completed by letting $\mu_i^0 = 1000$, $\tau_i^0 = 500$, $\lambda_i^0 = 3$ and $\nu_i^0 = 3$. Under the mixture model we assume non-informative priors for most parameters, while we model the shape parameters under the two treatments as exchangeable by introducing the hierarchical prior:

$$\log\left(\frac{1+\xi}{1-\xi}\right) \sim N(\varepsilon, 0.5^2)$$

$$\varepsilon \sim N(0,1)$$

where ε is estimated using data from both treatment groups; note that this prior reflects the prior belief that the tails of the distributions of costs are not very different between the two treatments (Conigliani and Tancredi, 2005a; O'Hagan and Stevens, 2003). For the effects we suppose that

$$e_{ij}\,|c_{ij} \sim N\left(\gamma_{ij}, \sigma_i^2\right),$$

where the conditional mean γ_{ij} depends linearly on the cost c_{ij} of patient j under treatment i :

$$\gamma_{ij} = \beta_i + \delta_i c_{ij}.$$

The overall mean effect in the i-th arm of the trial can then be written as

$$\gamma_i = \beta_i + \delta_i \mu_i.$$

Table 5 shows a 95% posterior credible interval ($PCI_{0.95}$) and a point estimate for the mean costs μ_1 and μ_2 obtained with the different models assumed for costs, and it is interesting to note how these results are slightly different from those of Section 3. In fact, while for the rMRI group the posterior credible interval obtained with the mixture model (1) is wider than the one obtained with BMA (although

the difference in width is not as significant as in Section 3), and the estimated mean cost obtained with the mixture model is higher than the one obtained with BMA, for the X-ray group the approaches of Section 2.1 and of Section 2.2 produce very similar results.

Within the BMA approach it is also interesting to look at the posterior model probabilities $\pi_i(m\,|D)$ for the log-normal, the gamma and the GPD distributions, that are also presented in Table 5 for each arm of the trial. In particular, while for the rMRI group the data definitely support the log-normal distribution, for the X-ray group both the gamma and the GPD are plausible models, with posterior probabilities 0.49 and 0.50 respectively. Note that this is exactly a situation where it is appropriate to apply model averaging: instead of choosing between the gamma and the GPD, and then studying the sensitivity of the conclusions in terms of cost-effectiveness, model averaging takes into account both models.

Finally, in order to interpret the results of the mixture model, it is interesting to look at the posterior distributions of its parameters. In particular, Figure 2 show the posterior distribution of the shape parameter ξ, the threshold α and the number of steps s for the two treatment groups. Note that even with such small data set and with mostly non-informative priors, these posterior distributions are quite concentrated, so that they give clear indications for the estimation of these parameters. Moreover, it is interesting to notice that the two posterior distributions of s are concentrated on the value 3 (the posterior means are 3.44 and 3.05 respectively), so that the evidence is that only 3 steps are sufficient to model adequately the bulk of the data. Finally, it is worth noting that the posterior mean of ξ and the posterior probability that $\xi < 0$ are -0.05 and 0.70 respectively for the X-ray group, and -0.70 and 0.99 respectively for the rMRI group. It follows that for both treatment groups there is evidence that the distribution of costs has an upper end point, that can be estimated with

$$E\left(\alpha - \frac{\sigma}{\xi}\,|c_{i1},...,c_{in_i},\xi<0\right) = \begin{cases} 48851 & i=1 \\ 23257 & i=2. \end{cases}$$

In order to assess if rMRI is more cost-effective than standard X-ray, we now look at the joint posterior distribution of the cost differential Δ_c and the effect differential Δ_e. These are summarised as 95% contour plots in Figure 3 (obtained assuming the normality of the bivariate posterior distribution), while the corresponding posterior summaries and the corresponding cost-effectiveness acceptability curves are shown

in Table 6 and in Figure 4 respectively.

Several points emerge from these results. First, the estimated effect difference is quite close to zero, so that rMRI does not seem to allow better diagnosis and treatment. Only if we assume a log-normal distribution in the X-ray group the probability that Δ_e is greater than zero reaches 0.23, but this model is not supported by the data, and therefore the corresponding results should not be considered. Second, there is evidence of a higher mean cost in the rMRI group under all the models assumed for costs, and in particular under the mixture model, with an estimated mean cost difference of \$822, compared to the \$505 obtained with BMA (again, only if we assume a log-normal distribution in the X-ray group the estimated mean cost difference does not go above \$250, but this model is not supported by the data and therefore these results are of no interest). The discrepancy between the results of the approaches of Section 2.1 and Section 2.2 can be seen also when comparing the

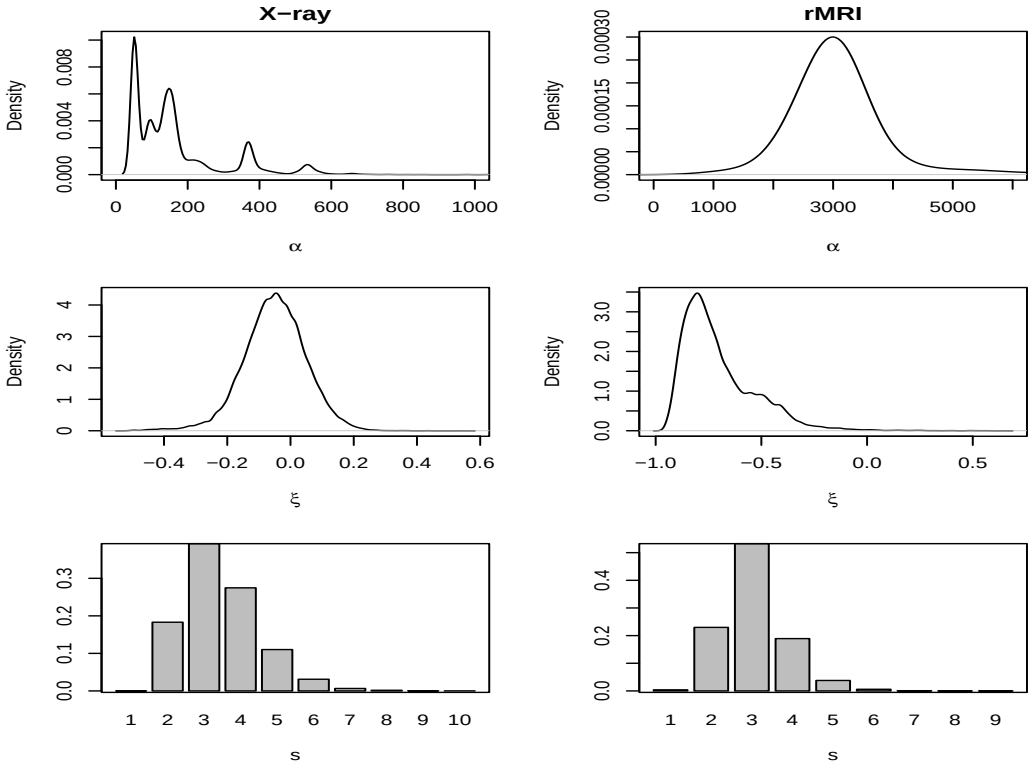

Figure 2: Low back pain trial: posterior distributions of the threshold, the shape parameter and the number of steps under the mixture model for X-ray costs and rMRI costs

Table 6: Low back pain trial: posterior summaries of cost differential and effect differential

X-ray	**rMRI**	$\mathbf{E}(\boldsymbol{\Delta_c})$	$\mathbf{PCI_{0.95}}$	$\mathbf{P}(\boldsymbol{\Delta_c}>\mathbf{0})$	$\mathbf{E}(\boldsymbol{\Delta_e})$	$\mathbf{PCI_{0.95}}$	$\mathbf{P}(\boldsymbol{\Delta_e}>\mathbf{0})$
Mixture	Mixture	822	212; 1538	1	0.03	-1.4;1.5	0.51
BMA	BMA	505	−36; 1116	0.97	0.07	-1.4;1.5	0.54
logN	logN	199	−554; 902	0.72	0.24	-1.2;1.7	0.62
logN	Gamma	250	−449; 842	0.79	0.23	-1.2;1.7	0.62
logN	GPD	224	−512; 911	0.75	0.24	-1.2;1.7	0.62
Gamma	logN	540	30; 1133	0.98	0.05	-1.4;1.5	0.53
Gamma	Gamma	591	176; 1030	1	0.04	-1.4;1.5	0.52
Gamma	GPD	565	79; 1123	0.99	0.05	-1.4;1.5	0.53
GPD	logN	486	−72; 1092	0.95	0.08	-1.3;1.5	0.54
GPD	Gamma	537	41; 1011	0.98	0.07	-1.4;1.5	0.54
GPD	GPD	511	−35; 1094	0.97	0.08	-1.3;1.5	0.54

contour plots in Figure 3, where the difference in the shape of the two joint distributions with respect to the vertical scale is quite obvious, and by looking at the CEACs in Figure 4. In fact, although the two curves have very similar behaviours, if we look at the values of K where the probability that rMRI is cost-effective is at least 0.5, we find $K \geq \$7400$ under BMA, and $K \geq \$30000$ under the mixture model. It follows that in terms of cost-effectiveness, here the semi-parametric ap-

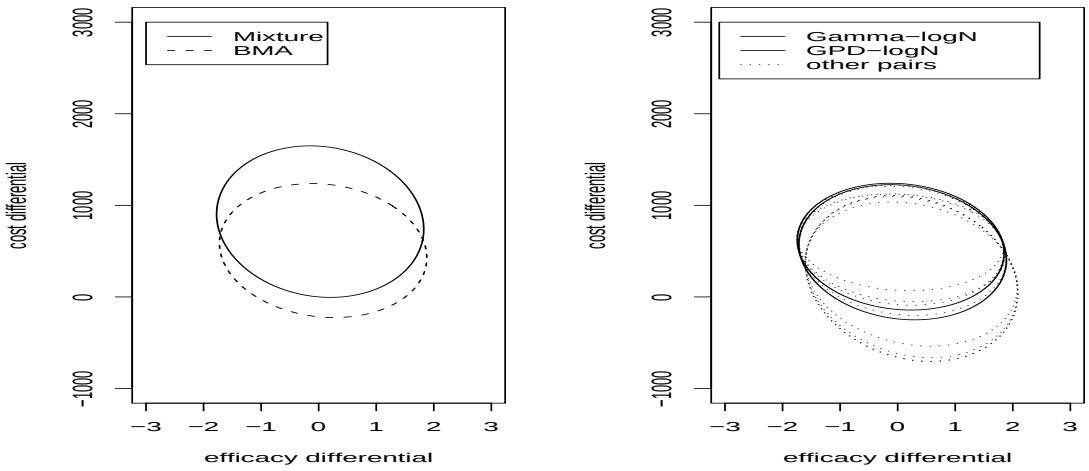

Figure 3: Low back pain trial: contour plots of cost-effectiveness density

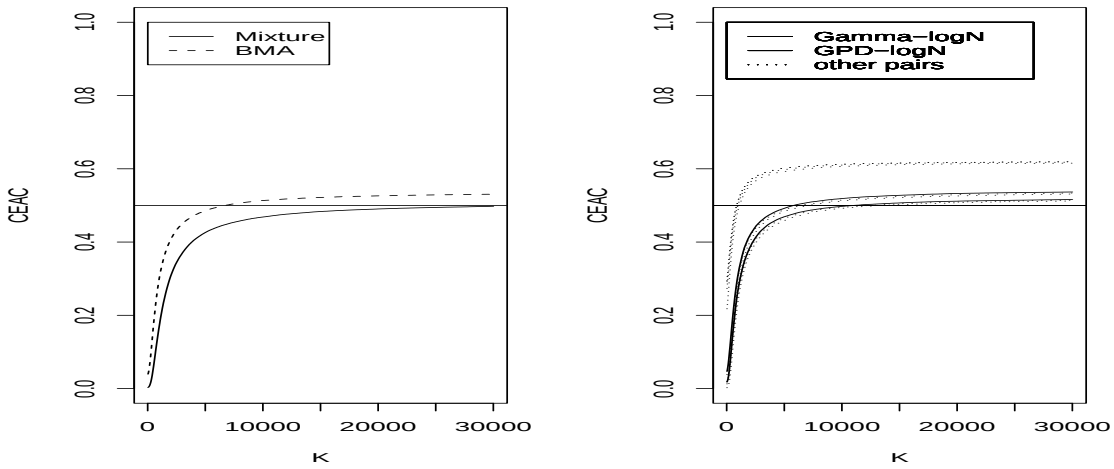

Figure 4: Low back pain trial: cost-effectiveness acceptability curves

proach and the Bayesian model averaging applied to the lognormal, gamma and GPD, produce rather different results. Note that from Table 5 we know that this is due to the fact that in the rMRI group the mixture model gives more weight to the upper tail of the cost distributions than any of the models included in $\mathcal{M}$, and in this sense it is particularly interesting to explore the behaviour of BMA when widening $\mathcal{M}$. If for instance we include also the Weibull and the log-logistic distributions in $\mathcal{M}$, plausible models for the X-ray group become the gamma, the GPD and the Weibull, with posterior probabilities 0.22, 0.23 and 0.55 respectively, but for the rMRI group the data still only support the log-normal distribution, so that in terms of cost-effectiveness the two analyses with BMA are quite similar, as it is shown by the contour plots and the CEACs in Figure 5.

5 Discussion

Most of the recent literature on cost-effectiveness analyses of clinical trial data agrees that inferences are significally sensitive to the choice of the model for costs, and in particular to how the upper tail of the cost distribution beyond the observed data is modelled. In particular it often happens that parametric models that fit the data equally well produce very different answers; conversely, models that fit badly can give similar inferences to those that fit well.

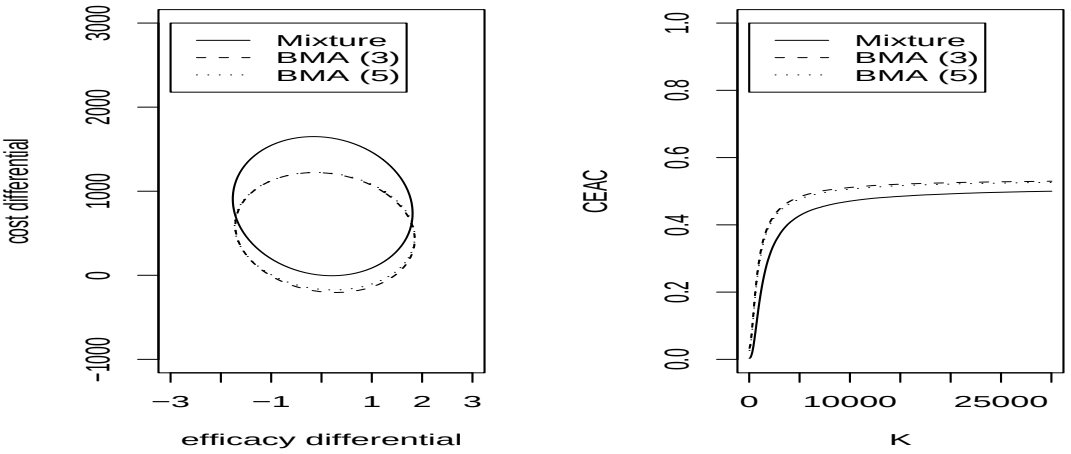

Figure 5: Low back pain trial: cost effectiveness analysis including the Weibull and the log-logistic distributions in the BMA procedure (in parentheses the number of models included in $\mathcal{M}$)

In this paper we have considered two different approaches to overcome this problem. The first one combines the semi-parametric approach to density estimation based on mixture models and the semi-parametric approach to tail estimation based on extreme value theory. The result is a very flexible model able to fit data set with very different shapes both in the bulk of data and in the tail. One drawback of this approach is that there is a price to pay for so much flexibility in terms of precision and efficiency of the corresponding inferences. Another problem is that for estimating the parameters of the the mixture model, and in particular the parameters of the tail, we need a large number of observations in each arm of the trial. The second approach is based on Bayesian model averaging performed on a sensible set of models for cost data, and is somehow in the spirit of the sensitivity analyses advocated by Thompson and Nixon (2005). It requires the specification of a set $\mathcal{M}$ of plausible models, but instead of studying how the conclusions change with the different models, it takes into account the inferences obtained with all the models in $\mathcal{M}$ that have a non-zero posterior probability. One drawback of this approach is that the models with a non-zero posterior probability are not necessarily models that fit the data well. It follows that particular care should be devoted to specify $\mathcal{M}$, in the sense that it should include all parametric models that might have generated

cost data. On the other hand, it is worth pointing out that the wider is $\mathcal{M}$, the more is difficult to assign and interpret the prior model probabilities. In fact, there is no requirement for the models in $\mathcal{M}$ to be distinct, so that for instance in our examples the exponential distribution belongs to the Weibull, the gamma and the GPD families. And in such cases, as pointed out for instance in O'Hagan and Foster (2004), a prior model probability may not reflect the total probability assigned to that model as a set of data-generating processes.

Acknowledgments We would like to thank Dr. Jerry Jarvik and Dr. Will Hollingworth for permission to use the low back pain trial data in our methodological study. Our retrospective analysis of the data is for expository purposes only. We are also grateful to Simon Thompson, Lara Jamieson and Tony O'Hagan for some usefull comments and ideas.

References

Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. *J. Health Serv. Res. Policy.* 1998; 3, 233-245.

Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence, *Health Economics* 2005; 14, 421-428.

Chib S, Jeliazkov I. Marginal likelihood from the Metropolis-Hastings output, *JASA* 2001; 96, 270-281.

Coles SG. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London, 2001.

Coles SG, Tawn JA. A Bayesian analysis of extreme rainfall data. *Appl. Stat.* 1996; 45, 463-478.

Conigliani C, Tancredi A. Semi-parametric modelling for costs of health care technologies, *Statist.Med.*, 2005.

Conigliani C, Tancredi A. A Bayesian semi-parametric approach for cost-effectiveness analysis in health economics, Working Paper del Dipartimento di Economia, Universit Roma Tre, n. 46, 2005.

Dinh P, Zhou X. Nonparametric Statistical Methods for Cost-Effectiveness Analyses, *Biometrics*, 2006.

Green PJ. A Primer on Markov Chain Monte Carlo. In *Complex Stochastic Systems*, edited by O.E. Barndorff-Nielsen, D.R. Cox, C. Kluppelberg, Chapman & Hall. 2000.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (with discussion), *Statistical Science* 14: 382-417, 1999.

Jarvik JG, Hollingworth W, Martin B, et al. Rapid magnetic resonance imaging vs. radiographs for patients with low back pain. *JAMA*, 2003; 289: 2810-18.

Nixon RM, Thompson SG. Parametric modelling of cost data in medical studies. *Statist.Med.* 2004; 23: 1311-1331.

O'Hagan A, Forster J. *Bayesian Inference*, Kendall's Advanced Theory of Statistics, II edn, vol. 2B, Arnold, London, 2004.

O'Hagan A, Stevens JW. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Statistical Methods in Medical Research* 2002; 11,469-490.

O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics* 2003; 12, 33-49.

O'Hagan A, Stevens JW., Montmartin J. Inference for the Cost-Effectiveness Acceptability Curve and Cost-Effectiveness Ratio, *Pharmacoeconomics* 2000; 17, 339-349.

Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. *Spine*, 1995, 20, 1899-908.

Robert CP, Casella G. *Monte Carlo Statistical Methods* Springer-Verlag, New York. 1999.

Royall R, Tsou TS. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions, *Journal of the Royal Statistical Society: Series B*, 2003, 65, 391-404.

Stinnett AA., Mulahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis, *Medical Decision Making*, 1998; 18: 68-80.

Tancredi A, Anderson CW, O'Hagan A. A Bayesian model for threshold uncertainty in extreme value theory. *Working paper 2002.12, Dipartimento di Scienze Statistiche, Università di Padova. Available at http://3w.eco.uniroma1.it/ utenti/tancredi/* . Submitted

Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? *Br. Med. J.* 2000; 329, 1197-1200.

Thompson SG, Nixon RM, How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making*, 2005; 416-423.

van Hout BA, Al MJ, Gordon GS, et al. Costs, effects and C/E ratios alonside a clinical trial *Health Economics*, 1994; 3:309-19.

Finito di stampare nel mese di giugno 2006, presso *Tipolitografia artigiana* **Colitti Armando** snc
00154 Roma • Via Giuseppe Libetta 15 a • Tel. 065745311/065740258
e-mail tcolitti@tin.it • www.colitti.it