Bayesian Statistical Economic Evaluation Methods for Health Technology Assessment

ANDREA GABRIO, GIANLUCA BAIO AND ANDREA MANCA

Summary

The evidence produced by healthcare economic evaluation studies is a key component of any health technology assessment (HTA) process designed to inform resource allocation decisions in a budget limited context. To improve the quality (and harmonize the generation process) of such evidence, many HTA agencies have established methodological guidelines describing the normative framework inspiring their decision-making process. The information requirements that economic evaluation analyses for HTA must satisfy typically involve the use of complex quantitative syntheses of multiple available datasets, handling mixtures of aggregate and patient-level information, and the use of sophisticated statistical models for the analysis of non-Normal data (e.g. time-to-event, quality of life and costs). Much of the recent methodological research in economic evaluation for healthcare has developed in response to these needs, in terms of sound statistical decision-theoretic foundations, and is increasingly being formulated within a Bayesian paradigm. The rationale for this preference lies in the fact that by taking a probabilistic approach, based on decision rules and available information, a Bayesian economic evaluation study can explicitly account for relevant sources of uncertainty in the decision process and produce information to identify an "optimal" course of actions. Moreover, the Bayesian approach naturally allows the incorporation of an element of judgement or evidence from different sources (e.g. expert opinion or multiple studies) into the analysis. This is particularly important when, as often occurs in economic evaluation for HTA, the evidence base is sparse and requires some inevitable mathematical modelling to bridge the gaps in the available data. The availability of free and open source software in the last two decades has greatly reduced the computational costs and facilitated the application of Bayesian methods and has the potential to improve the work of modellers and regulators alike, thus advancing the fields of economic evaluation of health care interventions. This chapter provides an overview of the areas where Bayesian methods have contributed to the address the methodological needs that stem from the normative framework adopted by a number of HTA agencies.

Keywords: Health Economics; Bayesian Statistics; Cost-Effectiveness Analysis ; Health Technology Assessment ; Probabilistic Sensitivity Analysis ; Decision Analytic Models ; Individual Level Data ; Aggregated Level Data

Decision-Making in Health Technology Assessment

Concerns about the rising healthcare costs and the desire to secure access to high quality and affordable medical care, have prompted the interest of the public towards the need to establish the most appropriate way to identify how to best invest limited healthcare resources for the benefit of Society. To this extent, many stakeholders have expressed interest in, and devoted efforts to use *Health Technology Assessment*

(HTA) processes to guide these decisions. HTA has been defined as “a multidisciplinary field of policy analysis, studying the medical, economic, social and ethical implications of development, diffusion and use of health technology” (International Network of Agencies for Health Technology Assessment, 2018).

In many countries, the organizations that perform HTAs are public sector agencies, reflecting the financing and/or provision of healthcare. For example, in the United Kingdom (UK), the *National Institute for Health and Care Excellence* (NICE) relies on HTAs to formulate guidance on the use of health technologies in the National Health Service for England and Wales. Similar agencies exist in Europe, the North American continent, South America and Australasia. The extent to which HTA activities are linked to a particular decision about the reimbursement, coverage, and use of a health technology influences the extent to which firm recommendations are made on the basis of the assessment and “appraisal” processes (NICE, 2013). HTA inherently requires consideration of the integration of medical interventions into clinical care and, as such, involves balancing a number of factors, including societal values, clinical and organizational context in which the technology will be used. An important element that informs the decision-making process in HTA is the *economic evaluation*, which has been defined as … *the comparative analysis of alternative courses of action in terms of both their costs and consequences* (Drummond et al., 2005b).

The type of data used in economic evaluations typically come from a range of sources, whose evidence is combined to inform HTA decision-making. Traditionally, relative effectiveness data are derived from *randomised controlled clinical trials* (RCTs), while healthcare resource utilisation, costs and preference-based quality of life data may come from the same study that estimated the clinical effectiveness or not. A number of HTA agencies have developed their own methodological guidelines to support the generation of the evidence required to inform their decisions. For example, the NICE guidelines on the analytical methods to be used (NICE, 2013) have been derived from the normative framework the Institute has adopted to increase consistency in analysis and decision making by defining a set of standardised methods for HTA. In this context, the primary role of economic evaluation for HTA is not the estimation of the quantities of interest (e.g. the computation of point or interval estimation, or hypothesis testing), but to aid decision making. The implication of this is that the standard frequentist analyses that rely on power calculations and *P*-values to estimate statistical and clinical significance, typically used in RCTs, are not well-suited for addressing these HTA requirements. It has been argued that, to be consistent with its intended role in HTA, economic evaluation should embrace a decision-theoretic paradigm (Claxton, 1999; Spiegelhalter et al., 2004; Briggs et al., 2006) and develop ideally within a Bayesian statistical framework (O’Hagan and Stevens, 2001; Spiegelhalter et al., 2004; Baio, 2012; Baio et al., 2017) to inform two decisions (a) whether the treatments under evaluation are cost-effective given the available evidence and (b) whether the level of uncertainty surrounding the decision is acceptable (i.e. the potential benefits are worth the costs of making the wrong decision). This corresponds to quantify the impact of the uncertainty in the evidence on the entire decision-making process (e.g. to what extent the uncertainty in the estimation of the effectiveness of a new intervention affects the decision about whether it is paid for by the public provider).

There are several reasons that make the use of Bayesian methods in economic evaluations particularly appealing. First, Bayesian modelling is naturally embedded in the wider scheme of decision theory; by taking a probabilistic approach, based on decision rules and available information, it is possible to explicitly account for relevant sources of uncertainty in the decision process and obtain an “optimal” course of action. Second, Bayesian methods allow extreme flexibility in modelling using computational algorithms such as *Markov Chain Monte Carlo* (MCMC) methods; this allows to handle in a relatively easy way the generally sophisticated structure of the relationships and complexities that characterise effectiveness, quality of life and cost data. Third, through the use of prior distributions, the Bayesian approach naturally allows the incorporation of evidence from different sources in the analysis (e.g. expert opinion or multiple studies), which may improve the estimation of the quantities of interest; the process is generally referred to as evidence synthesis and finds its most common application in the use of meta-analytic tools (Spiegelhalter et al., 2004). This may be extremely important when, as it often happens, there is only some partial (imperfect) information to identify the model parameters. In this case analysts are required to develop chain-of-evidence models (Ades, 2003). When required by the limitations in the evidence base, subjective prior distributions can be specified based on the synthesis and elicitation of expert opinion to identify the model, and their impact on the results can be assessed by presenting or combining the results across a range of plausible alternatives. Finally, under a Bayesian approach, it is straightforward to conduct *sensitivity analysis* to properly account for the impact of uncertainty in all inputs of the decision process; this is a required component in the approval or reimbursement of a new intervention for many decision-making bodies, such as NICE in the UK (NICE, 2013).

The general process of conducting a Bayesian analysis (with a view of using the results of the model to perform an economic evaluation) can be broken down in several steps, which are graphically summarized in Figure 1.1.

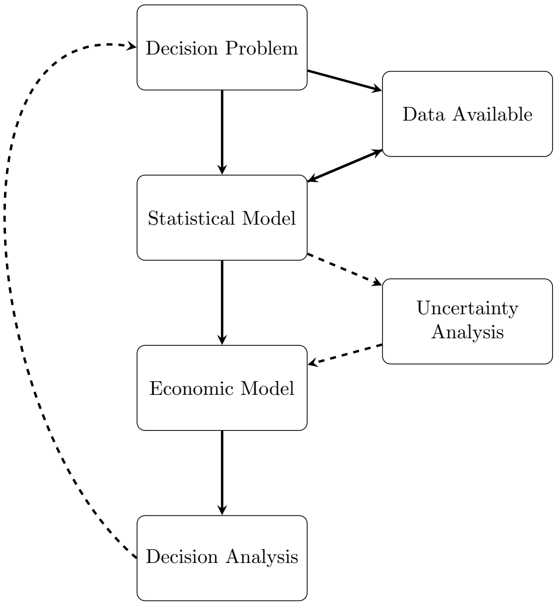


Figure 1.1: A graphical representation of the process of health economic evaluation. Once the decision problem has been defined (e.g. which interventions should be assessed) and the data have been identified and acquired, a statistical model is fitted to the available data to estimate the relevant model inputs, which are then fed to the economic model, where suitable population averages for the costs and benefits of each intervention are computed and compared. These are in turn used in the decision analysis, which dictates the best course of action given the available evidence and whose results can be used to update the decision problem. The uncertainty analysis assesses the impact of uncertainty in the statistical inputs throughout the whole process. Adapted with permission from Baio et al. (2017).

The starting point is the identification of the decision problem, which defines the objective of the economic evaluation (e.g. the interventions being compared, the target population, the relevant time horizon).

In line with the decision problem, a statistical model is constructed to describe the (by necessity, limited) knowledge of the underlying clinical pathways. This implies, for example, the definition of suitable models to describe variability in potentially observed data (e.g. the number of patients recovering from the disease because of a given treatment), as well as the epistemic uncertainty in the population parameters (e.g. the underlying probability that a random individual in the target population is cured, if given the treatment under study). At this point, all the relevant data are identified, collected and quantitatively sytnthesised to derive the estimates of the input parameters of interest for the model.

These parameter estimates (and associated uncertainties) are then fed to the economic model, with the objective of obtaining some relevant summaries indicating the benefits and costs for each intervention under evaluation.

Uncertainty analysis represents some sort of "detour" from the straight path going from the statistical model to the decision analysis: if the output of the statistical model allowed us to know with perfect certainty the "true" value of the model parameters, then it would be possible to simply run the decision analysis and make the decision. Of course, even if the statistical model were the "true" representation of the underlying data generating process (which it most certainly is not), because the data may be limited in terms of length of follow up, or sample size, the uncertainty in the value of the model parameters would still remain.

This "parameter" (and "structural") uncertainty is propagated throughout the whole process to evaluate its impact on the decision-making. In some cases, although there might be substantial uncertainty in the model inputs, this may not turn out to modify substantially the output of the decision analysis, i.e. the new treatment would be deemed as optimal irrespectively. In other cases, however, even a small amount of uncertainty in the inputs could be associated with very serious consequences. In such circumstances, the decision-maker may conclude that the availbale evidence is not sufficient to decide on which intervention to select and require more information before a decision can be made.

The results of the above analysis can be used to inform policy makers about two related decisions: (a) whether the new intervention is to be considered (on average) "value for money", given the evidence base available at the time of decision, and (b) whether the consequences (in terms of net health loss) of making the wrong decision would warrant further research to reduce this "decision uncertaint". While the type and specification of the statistical and economic models vary with the nature of the underlying data (e.g. individual level versus aggregated data – see Section 1.0.2), the decision and uncertainty analyses have a more standardised set up.

Decision Analytic Framework

Unlike in a standard clinical study, where the objective is to analyse a single primary outcome (e.g. survival, or the chance of experiencing some event), in economic evaluation the interest is in the analysis of a multivariate outcome $\boldsymbol{y} = (e,c)$, composed of a suitable measure of benefits e and the corresponding costs c. Consider $t = 0$ as the standard intervention currently available for the treatment of a specific condition and $t = 1,\ldots,T$ as a (set of) new option(s) being assessed. For a set of alternatives, optimality can be determined by framing the problem in decision theoretic terms (O'Hagan and Stevens, 2001; Spiegelhalter et al., 2004; Briggs et al., 2006; Baio, 2012).

More specifically, the decision analytic framework of economic evaluations can be summarised in the following steps. First, the sampling variability in the economic outcome (e,c) is characterised using a probability distribution $p(e,c \mid \boldsymbol{\theta})$, indexed by a set of parameters $\boldsymbol{\theta}$. Within the Bayesian framework, uncertainty in the parameters is also modelled using a prior probability distribution $p(\boldsymbol{\theta})$. Secondly, a summary measure (also called "utility function" in Bayesian language) is chosen to quantify the value associated with the uncertain consequences of a possible intervention. Thirdly, the optimal treatment option is determined by computing for each intervention the expectation of the chosen summary measure, with respect to both population (parameters) and individual (sampling) uncertainty/variability. Given the available evidence, the best intervention is the one associated with the maximum expected utility, which is equivalent to maximizing the probability of obtaining the outcome associated with the highest value for the decision-maker (Bernardo and Smith, 1999; Briggs et al., 2006; Baio, 2012).

Although in theory there are many possible choices for the utility function to be associated with each intervention, it is standard practice in economic evaluations to adopt the monetary *Net Benefit* (NB; Stinnett and Mullahy, 1998)

$$nb_t = ke_t - c_t. \tag{1}$$

Here, for each option t, (e_t,c_t) is the multivariate response, subject to individual variability expressed by a joint probability distribution $p(e,c \mid \boldsymbol{\theta})$. The parameter k in Equation 1 is a *threshold value* used by the decision maker to decide whether the new intervention represents "value for money". The NB is linear in (e_t,c_t), which facilitates interpretation and calculations. Notice that in this formulation the use of the NB implies that the decision maker is risk neutral, which is by no means always appropriate in health policy problems (Koerkamp et al., 2007; Baio, 2012).

In a full Bayesian setting, a complete ranking of the alternatives is obtained by computing the overall expectation of the NB over both individual variability and parameter uncertainty:

$$\mathcal{NB}_t = k\text{E}[e_t] - \text{E}[c_t],$$

that is, the expectation here is taken with respect to the full joint distribution $p(e,c,\boldsymbol{\theta}) = p(e,c \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. The option t associated with the maximum overall expected NB, i.e. $\mathcal{NB}^\star = \max_t \mathcal{NB}_t$, is deemed to be the most cost-effective, given current evidence. In the simple case where only two interventions $t = (0,1)$ are considered, the decision problem reduces to the assessment of the *Expected Incremental Benefit* (EIB),

$$\text{EIB} = \mathcal{NB}_1 - \mathcal{NB}_0. \tag{2}$$

If EIB > 0, then $t = 1$ is the most cost-effective treatment – notice that this analysis also applies in the case of pairwise comparisons, that is contrasting a generic intervention t against any of the others, e.g. treatment 1 vs treatment 2. Equation 2 can also be re-expressed as:

$$\text{EIB} = \text{E}\left[k\Delta_e - \Delta_c\right] = k\text{E}\left[\Delta_e\right] - \text{E}\left[\Delta_c\right],$$

where

$$\Delta_e = \text{E}[e \mid \boldsymbol{\theta}_1] - \text{E}[e \mid \boldsymbol{\theta}_0] = \mu_1^{(e)} - \mu_0^{(e)}$$

and

$$\Delta_c = \text{E}[c \mid \boldsymbol{\theta}_1] - \text{E}[c \mid \boldsymbol{\theta}_0] = \mu_1^{(c)} - \mu_0^{(c)}$$

are the the average increment in the benefits (from using $t = 1$ instead of $t = 0$) and in the costs, respectively. It is also possible to define the *Incremental Cost-Effectiveness Ratio* (ICER) as

$$\text{ICER} = \frac{\text{E}[\Delta_c]}{\text{E}[\Delta_e]} \tag{3}$$

so that, when EIB > 0 (or equivalently ICER $< k$), then $t = 1$ is the optimal treatment (associated with the highest expected net benefit). Thus, decision-making can be effected by comparing the ICER (defined as in Equation 3) to the threshold k. Notice that, in the Bayesian framework, the quantities (Δ_e, Δ_c) are random variables, because while sampling variability is being averaged out, these are defined as functions of the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. The second layer of uncertainty (i.e. about the parameters) can be further averaged out. Consequently, $\text{E}\left[\Delta_e\right]$ and $\text{E}\left[\Delta_c\right]$ are actually deterministic quantities and so is the ICER.[1] The uncertainty underlying the joint distribution of (Δ_e, Δ_c) can be represented in the *Cost Effectiveness Plane* (CEP; Black, 1990). Some relevant examples are shown in Figure 1.2. Intuitively, the CEP characterizes the uncertainty in the parameters $\boldsymbol{\theta}$ (and thus their functions Δ_e and Δ_c) represented by the dots in Figure 1.2a, typically obtained using simulation methods. This approach allows to assess the uncertainty surrounding the ICER, depicted as the red dot in Figure 1.2b, where the simulated distribution of Δ_e and Δ_c has been shaded out.

Another useful visual aid in economic evaluation for HTA decisions is represented by the "acceptance region" (also termed "sustainability area" in Baio, 2012). This is the portion of the CEP lying below the line $\text{E}\left[\Delta_c\right] = k\text{E}\left[\Delta_e\right]$ for a set value of k, indicated with a shaded area in Figure 1.2c. Because of the relationship between the EIB and the ICER highlighted in Equation 3, it is possible to see that an intervention associated with an ICER that lies in the acceptance region is a cost-effective strategy. Upon varying the value for the threshold, it is possible to assess (a) the extent to which the optimal decision changes and (b) the decision uncertainty associated with this new threshold value. For example, Figure 1.2d shows the acceptance region for a different choice of k. In this case, because the ICER lies outside the sustainability area, the new intervention $t = 1$ cannot be considered as cost-effective.

[1] Since both sampling variability *and* parameter (epistemic) uncertainty about both random variables Δ_e and Δ_c are marginalised out, then there is no uncertainty left about the ICER (which is just a number), given the model assumptions.

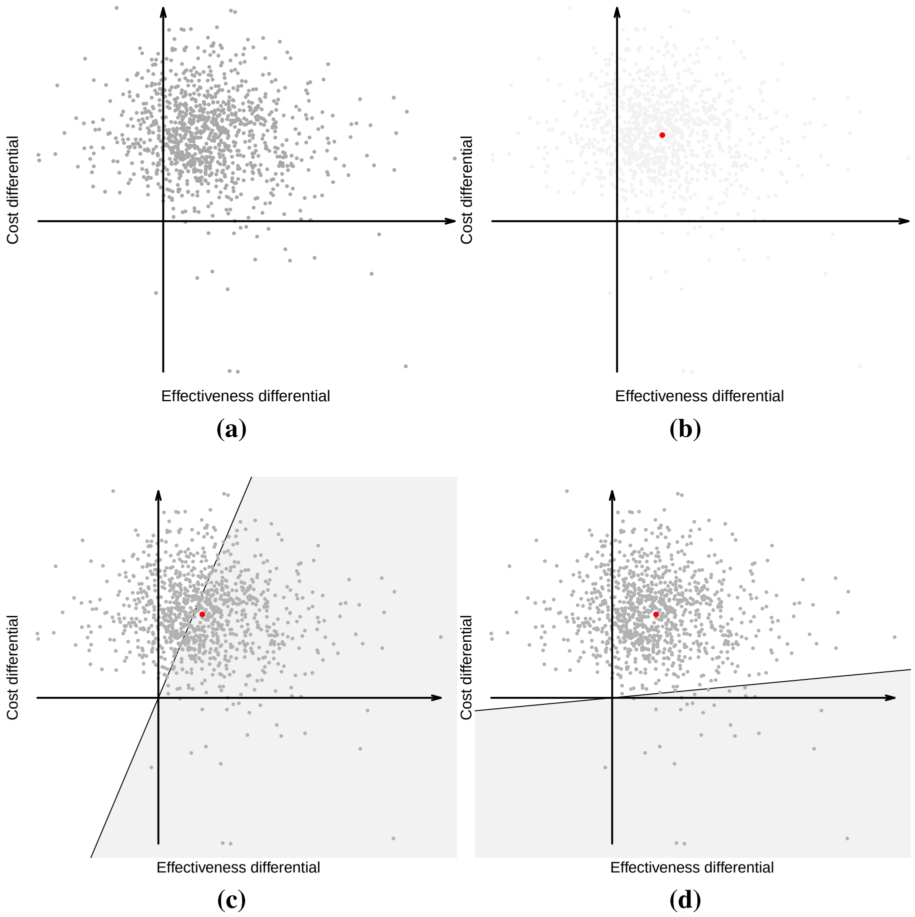


Figure 1.2: Graphical representation of the CEP under different scenarios. Panels (a) and (b) show the joint distribution for (Δ_e, Δ_c) and the position of the ICER, respectively. Panels (c) and (d) represent the CEP and corresponding acceptance region (shaded area) under two alternative choices for k. Source: Baio (2012).

Uncertainty Analysis

Health economic decision models are subject to various forms of uncertainty (Bilcke et al., 2011), including the uncertainty about the parameters of the model, typically referred to as *parameter uncertainty*, and the uncertainty about the model structure or *structural uncertainty*.

Parameter Uncertainty

Parameter uncertainty refers to our limited knowledge of the true value of the input parameter. We carry out *Probabilistic Sensitivity Analysis* (PSA) to propagate parameter(s) uncertainty throughout the model and to assess the level of confidence in the resulting output, in relation to the variability in the model inputs and the extent to which this is translated into decision uncertainty. In practice, in the decision process, parameter uncertainty is assessed for a large number of simulations from the joint (posterior) distribution of the model parameters (Claxton et al., 2005; Briggs et al., 2006; Baio and Dawid, 2015). Each of these simulations represents a potential future realisation of the state of the world in terms of the true, underlying value of the parameters, which in turn may affect the cost-effectiveness profile of the interventions being assessed.

PSA can be implemented by repeatedly sampling from the joint distribution $p(e,c,\boldsymbol{\theta})$ and propagating each realisation through the economic model. This allows to obtain a full distribution of decision-making processes, which is induced by the current level of uncertainty in the parameters. It is then possible to assess the expected net benefit taken with respect to individual variability only – that is, as a function of the parameters $\boldsymbol{\theta}$. Using the NB framework, this leads to

$$\text{NB}_t(\boldsymbol{\theta}) = k\text{E}[e_t \mid \boldsymbol{\theta}] - \text{E}[c_t \mid \boldsymbol{\theta}], \tag{4}$$

where the expectation is taken with respect to the conditional distribution $p(e,c \mid \boldsymbol{\theta})$. Thus, $\text{NB}_t(\boldsymbol{\theta})$ is a random quantity (randomness being induced by uncertainty in the model parameters $\boldsymbol{\theta}$). This is then used to compute the *incremental benefit* (IB)

$$\text{IB}(\boldsymbol{\theta}) = \text{NB}_1(\boldsymbol{\theta}) - \text{NB}_0(\boldsymbol{\theta}), \tag{5}$$

which corresponds to $\text{IB}(\boldsymbol{\theta}) = k\Delta_e - \Delta_c$.

Given the rationale underlying the computation of the quantities in Equation 4 and Equation 5, PSA effectively captures the uncertainty in the model parameters using probability distributions. Notably, because it is likely that the model parameters are characterized by some level of correlation, it is important to consider a full joint distribution. Although the Bayesian approach offers an intuitive framework to perform PSA, alternative approaches are available in the literature. For example, under a frequentist approach, PSA is typically performed using a two-stage approach. As part of the first stage the model parameters are estimated through standard techniques, e.g. maximum likelihood estimates $\hat{\boldsymbol{\theta}}$, based on the observed data. These estimates are then used to define a probability distribution describing the uncertainty in the parameters, based on a function $g(\hat{\boldsymbol{\theta}})$. For example, the method of moments can be used to determine a suitable form for a given parametric family for $g()$ (e.g. Normal, Gamma, Binomial, Beta, etc.) to match the observed mean and, perhaps, standard deviation or quantiles (Dias et al., 2013). Random samples from these "distributions" are then obtained using simple Monte Carlo simulations to represent parameter uncertainty. The most relevant implication of the distinction between the two approaches is that a frequentist analysis typically either assumes joint normality or ignores the potential correlation among the parameters when sampling from univariate distributions for each of the model parameters. Conversely, a full Bayesian model can account for this correlation automatically in the joint posterior distribution and therefore can generally provide a more realistic assessment of the impact of model uncertainty on the decision-making process (Ades et al., 2006; Baio et al., 2017).

Perhaps, the most popular tool used to represent the results of the PSA and its impact on decision uncertainty is the *Cost-Effectiveness Acceptability Curve* (CEAC; Van Hout et al., 1994). The CEAC can be derived by calculating the probability that that the intervention is cost-effective, estimated for a range of threshold values, typically ranging from 0 to a very high amount. In other words, the CEAC represents, for a range of possible cost-effectiveness threshold values, the probability that the intervention is value for money (to notice that this has a natural interpretation within a Bayesian approach as a posterior probability, given the observed data). This can be linked to the CEP: for a given value of k, the CEAC is the proportion of the joint distribution (Δ_e, Δ_c) that falls in the acceptance region.

The main advantage of the CEAC is that it allows a simple summarisation of the probability of cost-effectiveness upon varying the threshold parameter, effectively performing a sensitivity analysis on k. However, the CEAC only provides a partial picture of the overall uncertainty in the decision process. This is because it only assesses the probability of "making the wrong decision", while it does not consider the resulting expected consequences. More specifically, the CEAC can only address the problem of how likely it is that resolving parameters' uncertainty will change the optimal decision, without making any reference to the possible change in the payoffs.

For example, let us imagine an intervention that – for a particular value of the threshold k – is associated with a probability to be cost-effective that is just above 50%. Based on this information alone, the HTA policy-maker may prefer to request that further research is carried out to gather additional evidence and reduce decision uncertainty before making its funding decision. However, this decision would be warranted only in the case in which the consequences (in terms of net health losses) of this uncertainty justify the cost of funding the additional research, else the decision-maker may well use the available evidence and accept the risk associated with the error probability. Notably there may even be extreme scenarios where an intervention has a probability to be cost-effective of 99%, but the dramatic consequences for the 1% chance that it is not in fact cost-effective (e.g. all the relevant population is killed at a huge cost in terms of population health losses) may suggest that the tiny uncertainty is worth addressing, possibly delaying the decision-making process until better data can be collected to reduce it.

Another potential issue is that it has been suggested that very different distributions for the IB can produce the same value of the CEAC, which makes it difficult to interpret and might lead to incorrect conclusions (Koerkamp et al., 2007). As it will be shown later, the possible change in the payoffs can be taken into account using methods to analyse the *value of information*.

When more than two treatments are available to choose from, the CEAC can still be used to represent decision uncertainty, but should not be used to determine the optimal decision. Instead, analysts should use the *Cost-Effectiveness Acceptability Frontier* (CEAF; Fenwick et al., 2001) which shows the decision uncertainty surrounding the optimal choice. This is becasue the option that has the highest probability of being cost-effective need not have the highest expected net benefit (Barton et al., 2008). The CEAF represents the probability of the optimal treatment being cost-effective at different threshold values. As the threshold increases the preferred treatment changes, the switch point being where the threshold value increases beyond the relevant ICER reported for the treatment of interest. This type of presentation is particularly useful if there are three or more alternatives being compared, in which case there may be two or more switch points at different threshold values.

A fully decision theoretic approach to PSA that overcomes the shortcomings of the CEAC is based on the analysis of the *Value of Information* (VoI; Howard, 1966), which has increasingly been used in economic evaluation (Claxton, 1999, 2001; Briggs et al., 2006; Welton and Thom, 2015).

A VoI analysis quantifies the expected value of obtaining additional information about the underlying model parameters (e.g. by conducting a new study). Specifically, VoI assesses whether the potential value of additional information exceeds the cost of collecting this information. This assessment is based on two components: the probability of giving patients the incorrect treatment if the decision is based on current evidence (summarised by the CEAC) and the potential consequences (in terms of net benefits) of doing so.

One measure to quantify the value of additional information is the *Expected Value of Perfect Information* (EVPI), which translates the uncertainty associated with the cost-effectiveness evaluation in the model into an economic quantity. This quantification is based on the *Opportunity Loss* (OL), which is a measure of the potential consequences of choosing the most cost-effective intervention on average when it does not result in the intervention with the highest payoff in a “possible future”. A future can be thought of as obtaining enough data to know the exact value of the payoffs for the different interventions, i.e. to allow the decision-makers to known the optimal treatment with certainty. In a Bayesian setting, the “possible futures” are represented by the samples obtained from the posterior distribution of the quantities of interest, conditional on the model used to be true. Thus, the OL occurs when the optimal treatment on average is non-optimal for a specific point in the distribution of these quantities.

To calculate the EVPI, the values in each simulation are assumed to be known, corresponding to a possible future, which could happen with a probability based on the available knowledge included in and represented by the model. Under the NB framework, the OL is the difference between the known distribution net benefit associated with the most cost-effective intervention under the available evidence $\text{NB}^\star(\boldsymbol\theta) = \max_t \text{NB}_t(\boldsymbol\theta)$ and the maximum known-distribution net benefit given the parameters' simulated value $\text{NB}_\tau(\boldsymbol\theta)$, that is

$$\text{OL}(\boldsymbol\theta) = \text{NB}^\star(\boldsymbol\theta) - \text{NB}_\tau(\boldsymbol\theta),$$

where $t = \tau$ is the intervention associated with the optimal intervention overall. Taking the average over the distribution of $\text{OL}(\boldsymbol\theta)$ produces the EVPI,

$$\text{EVPI} = \text{E}_{\boldsymbol\theta}[\text{OL}(\boldsymbol\theta)] = \text{E}_{\boldsymbol\theta}[\text{NB}^\star(\boldsymbol\theta)] - \mathcal{NB}^\star. \tag{6}$$

The EVPI compares the ideal decision-making process, made under perfect information and represented by $\text{NB}^\star(\boldsymbol\theta)$, with the actual one made under current evidence and described by $\mathcal{NB}^\star$. The EVPI, as defined in

Equation 6, places an upper limit on the total amount that the decision maker should be willing to invest to collect further evidence and reduce completely the decision-uncertainty.

In practice, however, decision-makers are not interested in resolving the uncertainty for all model parameters, but only for those that drive the decision uncertainty. Indeed, some parameters may be already well understood, whereas for some others, it may not be possible to gather more evidence. These considerations lead to the definition of a further measure of VoI: the *Expected Value of Partial Perfect Information* (EVPPI), which is essentially a conditional version of the EVPI.

The basic principle is that the vector of parameters can in general be split into two components $\boldsymbol\theta = (\boldsymbol\phi, \boldsymbol\psi)$, where $\boldsymbol\phi$ is the subvector of parameters of interest (i.e. those that could be investigated further) and $\boldsymbol\psi$ are the remaining nuisance parameters. The EVPPI is calculated as a weighted average of the net benefit for the optimal decision at every point in the support of $\boldsymbol\phi$ after having marginalized out the uncertainty due to $\boldsymbol\psi$, that is

$$\text{EVPPI} = E_{\boldsymbol\phi}\left[\max_t \text{E}_{\boldsymbol\psi|\boldsymbol\phi}[\text{NB}_t(\boldsymbol\theta)]\right] - \mathcal{NB}^{\star}. \tag{7}$$

The EVPPI corresponds to the value of learning $\boldsymbol\phi$ with no uncertainty (approximated by averaging over its probability distribution as shown in Equation 7), while maintaining the current level of uncertainty on $\boldsymbol\psi$.

An additional measure to quantity the value of information is the *Expected Value of Individualized Care* (EVIC), which represents the potential value of collecting further information to inform individualised treatment decisions (Basu and Meltzer, 2007). The EVIC quantifies the gain in the decision-making from the incorporation of individual-level values of heterogeneous parameters, such as for instance patient preferences as measured by quality of life weights for various health states (in a decision making context that does not use societal utility values) or very detailed genetic information about the patient.

For example, assume that the heterogeneity in patients' preferences in the target population is reflected by a vector of patient-level attributes $\boldsymbol\theta = (\theta_1, \ldots, \theta_S)$ which determines the outcomes from any treatment t considered, and use the prefix "i" to indicate the individual patient outcomes expressed in terms of net benefits (i.e. iNB). Then, the EVIC is calculated as the average of the maximum net benefits of the treatments in each patient minus the maximum of the average net benefits of the treatments in all patients

$$\text{EVIC} = \int \max_t i\text{NB}(\boldsymbol\theta)p(\boldsymbol\theta)d\boldsymbol\theta - \max_t \int i\text{NB}(\boldsymbol\theta)p(\boldsymbol\theta)d\boldsymbol\theta, \tag{8}$$

where $p(\boldsymbol\theta)$ is the joint distribution of the patient-level attributes $\boldsymbol\theta$. Calculation of EVIC requires data on the outcomes of each treatment option in each individual patient, and so data from most (randomised) clinical studies are not suitable for EVIC analysis as they divide the study population into separate arms. Patient data must therefore be retrieved from studies with special designs that allow for individualized comparative effectiveness research, or be generated in decision-analytic models based on individual patient simulation (Basu, 2009).

The EVIC is obtained as the difference between the two quantities on the right-hand side of Equation 8, which can be interpreted as the average per patient societal values obtained under the "individualised care" and "paternalistic"' model, respectively (Basu and Meltzer, 2007). The first assumes that physicians are able to choose the optimal treatment for each individual patient based on the patient's true value of $\boldsymbol\theta$, while the second assumes that physicians are unaware of the values of $\boldsymbol\theta$ for individual patients but base their decisions on the distribution of $p(\boldsymbol\theta)$.

The approach used for the calculation of the EVIC is similar to that of the EVPI but the interpretation of the two quantities is different. The EVPI is the expected cost of uncertainty in parameters that are unknown to both the physician and the patient. It represents the maximum value of research to acquire additional information on those uncertain parameters in the population to inform "population-level decisions". Conversely, the EVIC is the expected cost of ignorance of patient-level information that may help explain heterogeneity and represents the potential value of research that helps to elicit individualized information on heterogeneous parameters that can be used "to make individualized decisions". The EVIC has also been applied as an informative metric to implement a subgroup-based policy (Van Gestel et al., 2012) to provide an estimate of the health gained due to the understanding of heterogeneity (i.e. observable characteristics that explain differences between subgroups) for specific parameters.

The EVIC falls in the so-called *Value of Heterogeneity* (VoH) framework (Espinoza et al., 2014), which indicates the additional health gains obtained by explicitly accounting for heterogeneity in decision-making. Specifically, VoH recognises that by taking into account the parameters that may determine heterogeneity between patients (i.e. based on some treatment effect moderators or baseline characteristics), different recommendations could be made for different subgroups. This results in a greater expected NB compared with decisions based on the average across the patient population.

When $s = 1,\ldots,S$ mutually exclusive subgroups are considered in the target population, it is possible to define the EVPI for the subgroup s as

$$\text{EVPI}_s = \text{E}_{\boldsymbol{\theta}}[\text{NB}^\star_s(\boldsymbol{\theta})] - \mathcal{NB}^\star_s, \tag{9}$$

where $\text{E}_{\boldsymbol{\theta}}[\text{NB}^\star_s(\boldsymbol{\theta})]$ and $\mathcal{NB}^\star_s$ are the expected value of the decision for subgroup s under perfect and partial (current) information, respectively. Equation 9 provides an upper bound for further research considering the overall uncertainty in the target population, which includes the uncertainty given by both exchangeable and non-exchangeable parameters (where exchangeable parameters are those whose estimate in a subgroup can be used to inform the cost-effectiveness in another mutually exclusive subgroup).

The total EVPI when considering S subgroups is obtained as the weighted average of each subgroup-specific EVPI weighted by the proportion of each subgroup in the population

$$\text{EVPI}_{(S)} = \sum_{s=1}^{S} \text{EVPI}_s w_s, \tag{10}$$

where w_s is is a weight indicating the proportion of the total population represented by subgroup s, with $\sum_{s=1}^{S} w_s = 1$. The population EVPI can be estimated by multiplying Equation 10 by the future population of patients expected to benefit from the new information. The $EVPI_{(S)}$ quantifies the value of heterogeneity both from the existing evidence and from the collection of new evidence to reduce the sampling uncertainty associated with subgroup-specific parameter estimates. It addresses the question of whether further research should be conducted, considering that different decisions can be made in different subgroups with future information.

Espinoza et al. (2014) extended the VoI concept to encompass the value of resolving systematic between-patient variability (as opposed to uncertainty) that can be understood as heterogeneity. They term this value of heterogeneity to emphasise that this quantity represents the health gain that can be derived by understanding heterogeneity for decision making. The authors decompose the VoH into 2 components. The first, called *static* VoH results from further exploration of the existing evidence (without collection of additional data) to identify, characterize, and quantify heterogeneity. The second component, coined *dynamic* VoH represents the value of collecting new evidence to reduce the sampling uncertainty associated with subgroup-specific parameter estimates.

Structural Uncertainty

Among all sources of uncertainty in health economic evaluations, structural uncertainty is the one that has received less attention in the literature, although many guidelines for good practice in decision modelling recognize the need to explore structural assumptions and the evidence supporting the chosen model structure (Hay et al., 1999; Weinstein et al., 2003). Structural uncertainty can be broadly defined as to the choice of the appropriate model structure in terms of assumptions or parameters for which there are only subjective data inputs. Examples of structural uncertainties are the choice of data used to inform a particular parameter (e.g. treatment effect), the extrapolations of parameter values estimated from short-term data, or the choice of the statistical model used to estimate specific parameters.

A possible approach to assess structural uncertainty is to present a series of results under alternative scenarios, which represent the different assumptions or model structures explored. The alternative models and their results are then presented to the decision-maker who uses this information to generate an implicit weight for each of the models. The model with the highest weight will be regarded as the "true" model and all other models discarded. Although this method can be useful, there are a number of potential problems. The weights applied to each model are subject to the interpretation process of the decision-maker which is difficult to replicate given an alternative set of decision-makers. Most importantly, by removing the uncertainty associated with choosing between multiple alternative models from the actual modelling process, structural uncertainty is not formally quantified (Snowling and Kramer, 2001). Given the limitations of presenting alternative scenario analyses, other strategies have been proposed to explore structural uncertainties in a more quantifiable and explicit way, such as model averaging (Bojke et al., 2006, 2009; Jackson et al., 2009).

Model averaging requires the analyst to build alternative models, based on different structural assumptions, and then average across these models weighting each by the plausibility (prior) of their assumptions, with the weights that are commonly obtained from experts' opinion. Model averaging more explicitly quantifies structural uncertainty, compared with simply presenting the results under alternative scenarios, and is therefore a better approach from a decision-making perspective. The problem of averaging across models can be viewed in a Bayesian sense as one in which a decision-maker needs to make the best possible use of information on an available model structure, and is typically referred to as Bayesian model averaging (Hoeting et al., 1999; Conigliani and Tancredi, 2009; Jackson et al., 2010).

A commonly used type of Bayesian model averaging assumes that, given $k = 1, \ldots, K$ alternative model structures (M_k), with $\boldsymbol{\theta}$ as the quantity of interest, the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ is:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{\theta} \mid M_k, \boldsymbol{y}) p(M_k \mid \boldsymbol{y}).$$

Thus, the distribution of $\boldsymbol{\theta}$ is an average of the posterior distributions for each of the models considered $p(\boldsymbol{\theta} \mid M_k, \boldsymbol{y})$, weighted by their posterior model probability $p(M_k \mid \boldsymbol{y})$. Where data exist to test each structural assumption, formal model averaging can be used to weight the inferences derived from different model structures to aid the decision-makers in their interpretation of the results.

An slightly different approach to Bayesian model averaging, proposed by Jackson et al. (2010) for cost-effectiveness models, formally takes into account structural uncertainty by constructing a probability distribution over model structures. The required distribution over the choice of model structures can be obtained by assessing the relative plausibility of the scenarios against the data using some measure of the adequacy of each model. These measures are then used to construct a model-averaged posterior distribution which allows for sampling uncertainty about model selection. Model weights can be estimated by a bootstrap procedure as the probability that each model is selected by the predictive criterion (i.e. highest expected predictive utility for a replicate data set among the models being compared), rather than in standard Bayesian model averaging, where each model is weighted by its posterior probability of being true.

Alternative structural assumptions can produce very different conclusions and it is therefore essential, for decision-making purposes, to incorporate structural uncertainty in the decision modelling. By explicitly characterizing the sources of structural uncertainty in the model as measurable parameters, it is possible to quantify the increase in decision uncertainty. Essentially, structural uncertainty could be regarded as the uncertainty related to elements of the decision model that are weakly informed by evidence. In these setting, judgement is required, either with respect to which scenarios are most plausible, which probabilities should be assigned in model averaging, or what values the missing parameters are likely to take. The latter approach is appealing as it enables the formal elicitation of parameter values and facilitates an analysis that is able to inform research decisions about these uncertainties.

Individual versus *Aggregated* Level Data

From a statistical point of view, one important distinction in the way in which economic evaluations are performed, depends on the nature of the underlying data. Increasingly often, *Individual Level Data* (ILD) are collected alongside information on relevant clinical outcomes as part of RCTs and provide an important source of data on economic evaluations. However, the use of a single patient-level data set can have some limitations with respect to evaluating healthcare as delivered in the real world, i.e. outside controlled environments (Drummond et al., 2005b). These include the partial nature of the comparisons undertaken, short-term follow-up, use of intermediate rather than ultimate measures of health outcomes, and unrepresentative patients, clinicians, and locations.

Consequently, it is advised that economic evaluations in healthcare take information from as many sources as possible to address some of these problems (Sculpher et al., 2005). Much of the health economic literature focuses on decision analytic models, which are mostly based on *Aggregated Level Data* (ALD). The decision model represents an important analytic framework to generate estimates of cost-effectiveness based on a synthesis of available data across multiple sources and for the comparison of multiple options that might not have been included as part of an RCT (e.g. from a literature review). Nevertheless, this synthesis of information is not always straightforward and different methodologies for deciding whether or not to include information or account for gaps in the literature are available (Drummond et al., 2005a; Briggs et al., 2006).

Individual Level Data

Typically, individual-level benefits are expressed in terms of generic health related quality of life (HRQoL) outcomes, the most common of which are derived using preference-based instruments (e.g. EQ-5D; www.euroqol.org). Integrating HRQoL weights and survival it is possible to estimate an individual's *Quality-Adjusted Life Years* or QALYs (Loomes and McKenzie, 1989). When one has access to ILD, QALYs can be computed for each study participant. The same principle applies to the derivation of costs at the individual level, which can be estimated combining resource use data (e.g. hospital days, intensive care unit days) with the unit costs or prices of those resources, obtained from self-reported methods or raw data extracted from healthcare services (Thorn et al., 2013; Franklin et al., 2017). Within the context of RCTs, these data are typically collected through a combination of case report forms, patient diaries, and locally administered questionnaires (Ridyard and Hughes, 2010).

When ILD are used, the parameters derived from the statistical model typically consist in the relevant population summaries $(\mu_t^{(e)}, \mu_t^{(c)})$, which can be used to derive a range of model input parameters and better characterise their distribution. However, these types of data are often characterised by some complexities (e.g. correlation, non normality, spikes and missingness) which, if not accounted for in the statistical model using appropriate methods, could lead to biased inferences and mislead the cost-effectiveness assessment. By virtue of its modular nature, Bayesian modelling is very flexible, which means that a basic structure can be relatively easily extended to account for the increasing complexity required to formally and jointly allow for these complexities.

A General Bayesian Modelling Framework for ILD

Individual-level benefit and cost data (e,c) are typically subject to some level of correlation, and thus it is important to formally account for this in the statistical modelling. One useful strategy (O'Hagan and Stevens, 2001; Nixon and Thompson, 2005; Baio, 2012) is to factorise the joint sampling distribution $p(e,c)$ in terms of a conditional and a marginal distribution

$$p(e,c) = p(e)p(c \mid e) = p(c)p(c \mid e). \tag{11}$$

Note that while it is possible to use interchangeably either factorisation, without loss of generality, the framework in the following is described by expressing the joint distribution in Equation 11 through a marginal distribution for the benefits $p(e)$ and a conditional distribution of the costs given the benefits $p(c \mid e)$.

For example, for each individual $i = 1,...,n_t$ in each treatment or intervention arm $t = 0,...,T$, the distribution of the benefits is defined as $p\left(e_{it} \mid \boldsymbol{\theta}_t^{(e)}\right)$, indexed by a set of parameters $\boldsymbol{\theta}_t^{(e)}$. These typically consist in a *location* $\phi_{it}^{(e)}$ and a set of *ancillary* parameters ψ_{et}, which can include some measure of marginal variance, $\sigma_t^{2(e)}$. The location parameter can be modelled using a generalised linear structure, e.g.

$$g^{(e)}\left(\phi_{it}^{(e)}\right) = \alpha_{0t}\ [+\ldots], \tag{12}$$

where α_{0t} is the intercept and the notation $[+\ldots]$ indicates that other terms (e.g. quantifying the effect of relevant covariates) may or may not be included in the model. For example, the baseline utilities are likely to be highly correlated with the QALYs and should be included in the regression model to obtain adjusted mean estimates (Manca et al., 2005; Hunter et al., 2015). In the absence of covariates or assuming that a centered version $x^*_{it} = (x_{it} - \bar{x}_t)$ is used, the parameters $\mu_t^{(e)} = g^{(e)-1}(\alpha_{0t})$ in Equation 12 represent the population average benefits in each group.

As for the costs, the conditional model $p\left(c_{it} \mid e_{it}, \boldsymbol{\theta}_t^{(c)}\right)$ is specified so to explicitly depend on the benefit variable, as well as on a set of quantities $\boldsymbol{\theta}_t^{(c)}$, again comprising a location and ancillary parameters. Note that in this case $\psi_t^{(c)}$ includes a conditional variance $\tau_t^{2(c)}$ which, within a linear or generalised linear model structure, can be expressed as a function of the marginal variance $\sigma_t^{2(c)}$ (Nixon and Thompson, 2005; Baio, 2012). The cost location can be modelled as a function of the benefits as

$$g^{(c)}\left(\phi_{it}^{(c)}\right) = \beta_{0t} + \beta_{1t}(e_{it} - \mu_t^{(e)})\ [+\ldots]. \tag{13}$$

Here, $\left(e_{it} - \mu_t^{(e)}\right)$ is the centered version of the benefits variable, while β_{1t} captures the extent of linear dependency of c_{it} on e_{it}. As for the benefits, other covariates may or may not be included in Equation 13, e.g. the baseline costs (Van Asselt et al., 2009). Assuming these covariates are either also centered or absent, $\mu_t^{(c)} = g^{(c)-1}(\beta_{0t})$ are the population average costs in each group.

Figure 1.3 shows a graphical representation of the general modelling framework described in Equation 11. The benefit and cost distributions are represented in terms of combined "modules" — the blue and the red boxes — in which the random quantities are linked through logical relationships. This ensures the full characterisation of the uncertainty for each variable in the model. Notably, this is general enough to be extended to any suitable distributional assumption, as well as to handle covariates in either or both the modules.

Arguably, the easiest way of jointly modelling two variables is to assume Bivariate normality, which in our context can be factorised into marginal and conditional Normal distributions for e_{it} and $c_{it} \mid e_{it}$, using an identity link function for the location parameters. However, benefit (e.g. as measured in terms of QALYs) and, especially, cost data can be characterised by a large degree of skewness, which makes the assumption of normality unlikely to be adequate. In a frequentist framework, a popular approach among practitioners to account for skewness in the final estimates is non-parametric bootstrapping (Barber and Thompson, 2000). This method typically generates the distribution of average costs and effects across repeated samples by drawing from the original data with replacement. Although this procedure may accommodate the skewed nature of the data, reliance on using simple averages typically gives similar results to assuming normal distributions and can lead to incorrect inferences (O'Hagan and Stevens, 2001). To deal with skewness, particularly within a Bayesian approach, the use of more appropriate parametric modelling has been proposed in the literature (Nixon and Thompson, 2005; Thompson and Nixon, 2005). For example, one can

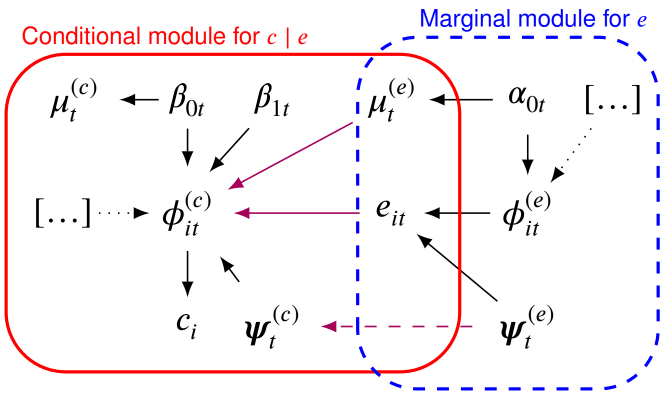


Figure 1.3: Joint distribution $p(e,c)$, expressed in terms of a marginal distribution for the effectiveness and a conditional distribution for the costs, respectively indicated with a solid red line and blue dashed line. The parameters indexing the corresponding distributions or "modules" are indicated with different Greek letters, while i and t denote the individual and treatment index, respectively. The solid black and magenta arrows show the dependence relationships between the parameters within and between the two models, respectively. The dashed magenta arrow indicates that the ancillary parameters of the cost model may be expressed as a function of the corresponding effectiveness parameters. The dots enclosed in the square brackets indicate the potential inclusion of other covariates at the mean level for both modules.

specify a Beta marginal for the benefits and a Gamma conditional for the costs:

$$e_{it} \mid \boldsymbol{\theta}_t^{(e)} \sim \text{Beta}\left(\phi_{it}^{(e)}\tau_{it}^{(e)}, \left(1-\phi_{it}^{(e)}\right)\tau_{it}^{(e)}\right) \qquad \text{and} \qquad c_{it} \mid e_{it}, \boldsymbol{\theta}_t^{(c)} \sim \text{Gamma}\left(\phi_{it}^{(c)}\tau_{it}^{(c)}, \tau_{it}^{(c)}\right).$$

The Beta distribution can be parameterised in terms of location $\phi_{it}^{(e)}$ and scale $\tau_{it}^{(e)} = \left(\frac{\phi_{it}^{(e)}\left(1-\phi_{it}^{(e)}\right)}{\sigma_t^{2(e)}} - 1\right)$, while the Gamma distribution can be parameterised in terms of location $\phi_{it}^{(c)}$ and rate $\tau_{it}^{(c)} = \frac{\phi_{it}^{(c)}}{\sigma_t^{2(c)}}$. The generalised linear model for the location parameters is then specified using a logit and logarithmic link functions for $\phi_{it}^{(e)}$ and $\phi_{it}^{(c)}$, respectively. The marginal means for the benefits and costs in each group can then be obtained using the respective inverse link functions

$$\mu_t^{(e)} = \frac{\exp(\alpha_{0t})}{1+\exp(\alpha_{0t})} \qquad \text{and} \qquad \mu_t^{(c)} = \exp(\beta_{0t}).$$

Within a Bayesian approach, it is straightforward to define a prior distribution on the parameters $\boldsymbol{\theta}_t^{(e)} = (\alpha_{0t}, \sigma_t^{(e)})$ and $\boldsymbol{\theta}_t^{(c)} = (\beta_{0t}, \beta_{1t}, \sigma_t^{(c)})$ and then induce a prior on the mean and, a fortiori, on the other model parameters. These models can also be further extended to handle additional features in the outcomes: from handling data with a hierarchical structure (Nixon and Thompson, 2005) to dealing with administrative censoring on the cost scale (Willan et al., 2005).

An interesting feature of the Bayesian approach is that it allows the inclusion of relevant prior information on the natural scale parameters, e.g. the population average costs and benefits (Baio, 2014). This is particularly relevant in studies where the sample size is limited – for example, in the case of pilot trials.

Hurdle Models to Handle Spikes

Another potential issue when modelling ILD is that e_{it} and c_{it} may exhibit spikes at one or both of the boundaries of the range for the underlying distribution. For example, some patients in a trial may not accrue any cost at all (i.e. $c_{it} = 0$) thus invalidating the assumptions for the Gamma distribution, which is defined on the range $(0,+\infty)$. Similarly, individuals who are associated with perfect health (i.e. $e_{it}=1$) may be observed, which makes it difficult to use a Beta distribution, defined on the open interval $(0,1)$. When the proportion of these values is substantial, they may induce high skewness in the data and the application of simple methods may lead to biased inferences (Mihaylova et al., 2011; Basu and Manca, 2012).

A solution suggested to handle the spikes is the application of *hurdle models* (Ntzoufras, 2009; Mihaylova et al., 2011; Basu and Manca, 2012; Baio, 2014; Gabrio et al., 2018). These are mixture models defined by two components: the first one is a mass distribution at the spike, while the second is a parametric model applied to the natural range of the relevant variable. Usually, a logistic regression is used to estimate the probability of incurring a "structural" value (e.g. 0 for the costs, or 1 for the QALYs); this is then used to weight the mean of the "non-structural" values estimated in the second component.

The modelling framework in Figure 1.3 can be expanded to a hurdle version in a relatively easy way for either or both outcomes. For example, assume that some unit QALYs are observed in e_{it}; an indicator variable $d^{(e)}_{it}$ can be defined taking value 1 if the $i-$th individual is associated with a structural value of one ($e_{it}=1$) and 0 otherwise ($e_{it}<1$). This is then modelled as

$$\begin{aligned} d^{(e)}_{it} &:= \mathbb{I}(e_{it}=1) \sim \text{Bernoulli}\left(\pi^{(e)}_{it}\right) \\ \text{logit}\left(\pi^{(e)}_{it}\right) &= \gamma_0\,[+\ldots], \end{aligned} \tag{14}$$

where $\pi^{(e)}_{it}$ is the individual probability of unit QALYs, which is estimated on the logit scale as a function of a baseline parameter γ_0. As for the benefit and cost models, other relevant covariates can be additively included in Equation 14 (e.g. the baseline utilities). Within this framework, the quantity

$$\bar{\pi}^{(e)}_t = \frac{\exp(\gamma_{0t})}{1+\exp(\gamma_{0t})} \tag{15}$$

represents the estimated marginal probability of unit QALYs.

Depending on the value of $d^{(e)}_{it}$, the observed data on e_{it} can be partitioned into two subsets. In the first subset, formed by the n^1 subjects for whom $d^{(e)}_{it}=1$, is identified with a variable $e^1_{it}=1$. Conversely, the second subset consists of the $n^{<1}_t=(n_t-n^1_t)$ subjects for whom $d^{(e)}_{it}=0$ and these individuals are identified with a variable $e^{<1}_{it}$. Because this is less than 1, it can be modelled directly using a Beta distribution characterised by an overall mean $\mu^{<1(e)}_t$ [2]. The overall population average benefit measure in both treatment groups is then computed as the linear combination

$$\mu^{(e)}_t = \left(1-\bar{\pi}^{(e)}_t\right)\mu^{<1(e)}_t + \bar{\pi}^{(e)}_t, \tag{16}$$

where $\bar{\pi}^{(e)}_t$ and $\left(1-\bar{\pi}^{(e)}_t\right)$ in effect represent the weights used to mix the two components.

Using a similar approach, it is possible to specify a hurdle model for the cost variables and define an indicator variable $d^{(c)}_{it}$ to partition c_{it} into the subsets of the individuals associated with a null (c^0_{it}) and positive ($c^{>0}_{it}$) cost. The overall population average costs can then be obtained by computing the cost measures that are analogous to those in Equation 15 and 16. Specifically, $\mu^{(c)}_t$ is derived as a weighted average using the estimated marginal probability of zero costs $\bar{\pi}^{(c)}_t$ and the mean parameter $\mu^{>0(c)}_t$, derived from fitting a Gamma distribution to $c^{>0}_{it}$.

Missing Data

Finally, ILD are almost invariably affected by the problem of missingness. Numerous methods are available for handling missing values in the wider statistical literature, each relying on specific assumptions whose

[2] Since the Beta distribution is defined on the open interval $(0,1)$, when negative values are present in $e^{<1}_{it}$, some tranformation is required in order to fit the model to the data (Basu and Manca, 2012)

validity must be assessed on a case-by-case basis. Whilst some guidelines exist for performing economic evaluations in the presence of missing outcome values (Ramsey et al., 2015), they tend not to be consistently followed in published studies, which have historically performed the analysis only on individuals with fully-observed data (Wood et al., 2004; Groenwold et al., 2012; Noble et al., 2012; Gabrio et al., 2017; Leurent et al., 2018). However, despite being easy to implement, these analyses are inefficient, may yield biased inferences, and lead to incorrect cost-effectiveness conclusions (Harkanen et al., 2013; Faria et al., 2014).

Multiple Imputation (MI; Rubin, 1987) is a more flexible method, which increasingly represents the *de facto* standard in clinical studies (Manca and Palmer, 2005; Burton et al., 2007; Diaz-Ordaz et al., 2014). In a nutshell, MI proceeds by replacing each missing data point with a value simulated from a suitable model. M complete (i.e. without missing data) replicates of the original dataset are thus created, each of which is then analysed separately using standard methods. The individual estimates are pooled using meta-analytic tools such as *Rubin's rules* (Rubin, 1987), to reflect the inherent uncertainty in imputing the missing values. For historical reasons, as well as on the basis of theoretical considerations, the number of replicated datesets M is usually in the range 5-10 (Rubin, 1987; Schafer, 1997, 1998).

Even though it has been shown that MI performs well in most standard situations, when the complexity of the analysis increases, a full Bayesian approach is likely to be a preferable option as it jointly imputes missing values and estimates the model parameters. In this way, the analyst is not required to explicitly specify which components should enter the imputation model to ensure the correct correspondence with the analysis model to avoid biased results (Erler et al., 2016). An example of this danger is when the imputation model includes less variables that those that will be used as predictors or covariates in the final analysis, or when it excludes the outcome(s) of interest. These situations should be avoided, as they generally result in both inconsistent estimators of the analysis model parameters and invalidity of the Rubin's variance estimator (Carpenter and Kenward, 2013). By contrast, especially in settings where the variables are characterised by complex dependence relationships, a full Bayesian approach ensures the coherence between the analysis and imputation steps through the joint imputation of the missing values and estimation of the parameters of interest.

In addition, in many applications, MI is based upon assuming a *Missing At Random* (MAR) mechanism, i.e. the observed data can explain fully the reason for why some observations are missing. However, this may not be reasonable in practice (e.g. for self-reported questionnaire data) and it is important to explore whether the resulting inferences are robust to a range of plausible *Missing Not At Random* (MNAR) mechanisms, which cannot be explained fully by the observed data.

Neither MAR nor MNAR assumptions can be tested using the available data alone and thus it is crucial to perform sensitivity analysis to explore how variations in assumptions about the missing values impact the results (Carpenter and Kenward, 2013; Molenberghs et al., 2015; Mason et al., 2018). The Bayesian approach naturally allows for the principled incorporation of external evidence through the use of prior distributions, e.g. by eliciting expert opinions (Mason et al., 2017), which is often crucial for conducting sensitivity analysis to a plausible range of missingness assumptions, particularly under MNAR.

Example: The MenSS trial

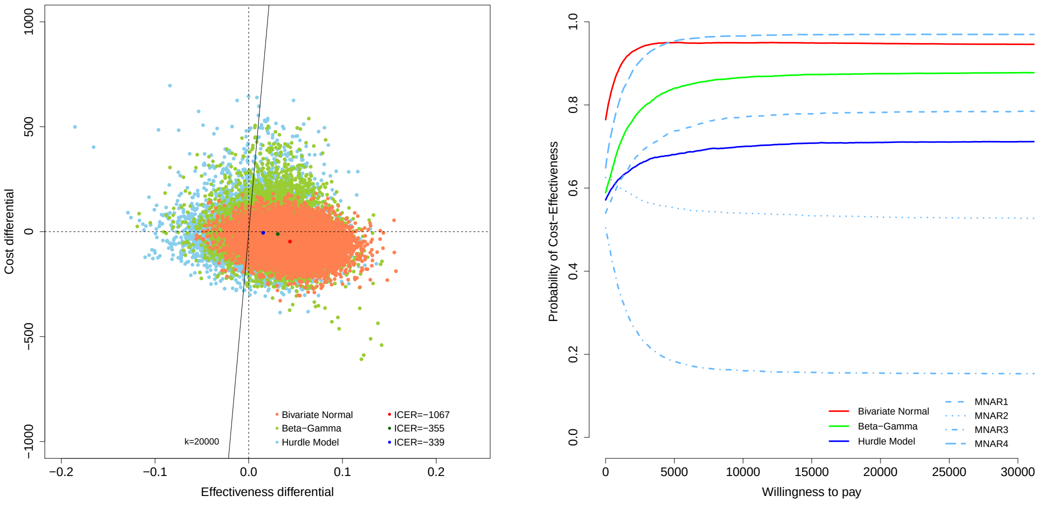
The results from the analysis of a case study, taken from Gabrio et al. (2018), are reported to show the importance of adopting a comprehensive modelling approach to ILD and the strategic advantages of building these complex models within a Bayesian framework. In this pilot RCT, the MenSS trial (Bailey et al., 2016), 159 individuals were randomised to receive either usual clinical care only ($n_0 = 75$) or a combination of of usual care and a new digital intervention ($n_1 = 84$) to reduce the incidence of sexually transmitted infections in young men. The outcomes of the economic evaluation are individual level QALYs and costs, which are computed based on fully-observed EQ-5D and resource use data for only $27(36\%)$ and $19(23\%)$ individuals in the control and intervention group, respectively.

Using the framework described in Equation 11, three models that account for a different number of data complexities are contrasted. These are: Bivariate Normal for the two outcomes; Beta marginal for the

QALYs and Gamma conditional for the costs; Beta marginal with a hurdle approach for spikes at 1 in the QALYs and Gamma conditional for the costs.

The models are fitted to the full data (observed and missing), imputing the missing values either under MAR (for all models) or alternative MNAR scenarios (only for the hurdle model). Specifically, for the Hurdle model, a sensitivity analysis is conducted to explore the robustness of the results to some departures from MAR. Four "extreme" MNAR scenarios are defined with respect to the number of missing individuals that could be associated with a unit QALYs in either or both treatment groups. These are: all the missing values in both groups (MNAR1); none of the missing values in both groups (MNAR2); all the missing values in the control and none in the intervention (MNAR3); all the missing values in the intervention and none in the control (MNAR4).

Figure 1.4 shows the CEPs and CEACs associated with the implementation of the three models to the MenSS data. The results for each model under MAR are indicated with different coloured dots and solid lines (red–Bivariate Normal, green–Beta-Gamma and blue–Hurdle model). In the CEAC plot, the results under the four MNAR scenarios explored are indicated with different types of dashed lines.



(a) Cost-Effectiveness Plane

(b) Cost-Effectiveness Acceptability Curve

Figure 1.4: CEPs (panel a) and CEACs (panel b) associated with the Hurdle (blue dots and line), Bivariate Normal (red dots and line) and Beta-Gamma (green dots and line) models. In the CEPs, the ICERs based on the results from the three model specifications under MAR are indicated with corresponding darker coloured dots. For the CEACs, in addition to the results under MAR (solid lines), the probability values for four MNAR scenarios are represented with different types of dashed lines.

In the CEP (panel a), for all three models more than 70% of the samples fall in the sustainability area at $k = £20,000$, even though the cloud of dots is more spread out for the Beta-Gamma and, especially, for the Hurdle model compared with the Bivariate Normal. All models are associated with negative ICERs, which suggests that, under MAR, the intervention can be considered as cost-effective by producing a QALYs gain at virtually no extra costs.

In the CEAC (panel b), the results under MAR for the Bivariate Normal and Beta-Gamma models indicate the cost-effectiveness of the new intervention with a probability above 0.8 for most values of k. Conversely, for the Hurdle model, the curve is shifted downward by a considerable amount with respect to the other models, and suggests a more uncertain conclusion. However, none of these results is robust to the MNAR scenarios explored: the probability of cost-effectiveness is very sensitive to the assumed number of structural ones in both treatment groups.

The results from this trial show considerable variations in the cost-effectiveness assessment depending on the number of complexities that are handled in the statistical model. Since this study is representative of the "typical" dataset used in trial-based economic evaluations, then it is highly likely that the same features (and potentially the same contradictions in the results, upon varying the complexity of the modelling assumptions) apply to other cases. The Bayesian approach allows to construct a flexible framework to jointly handle the complexities of ILD in economic evaluations, which may avoid biased results and misleading cost-effectiveness conclusions.

Bayesian Model Averaging

Conigliani and Tancredi (2009) show how Bayesian model averaging can be used to assess structural uncertainty in economic evaluations with respect to the choice of the distribution for the costs c_{it}. This approach requires the consideration of a set of plausible cost models (e.g. alternative parametric distributions) $\mathcal{M}_t = \{M_{1t}, \ldots, M_{Kt}\}$ for each treatment t being evaluated. The quantities of interest are the population mean costs $\mu_t^{(c)}$, which are unknown parameters of all models in $\mathcal{M}_t$. Bayesian model averaging can be used to obtain the posterior distribution of $\mu_t^{(c)}$ as a mixture of its posterior marginal distributions under each of the models in $\mathcal{M}_t$:

$$p\left(\mu_t^{(c)} \mid c_{it}\right) = \sum_{k=1}^{K} p\left(\mu_t^{(c)} \mid c_{it}, M_{kt}\right) p\left(M_{kt} \mid c_{it}\right), \tag{17}$$

where the mixing probabilities are given by the posterior model probabilities $p(M_{kt} \mid c_{it})$ (Hoeting et al., 1999). Rather than studying how the conclusions change across different cost models, Bayesian model averaging takes into account the inferences obtained with all the models in $\mathcal{M}_t$ that have non-zero posterior probability. The main difficulty is the specification of the set of models which should include a wide range of plausible choices, i.e. for the costs the distributions should be positively skewed and offer a range of different tail behaviours for the data (e.g. LogNormal, Gamma, etc.). It is also convenient to re-parameterise all the distributions in $\mathcal{M}_t$ in terms of means μ_{ct} and standard deviations σ_{ct} (or some other parameters) to make the prior specification easier. This implies that the same prior distribution can be introduced under the various models in $\mathcal{M}_t$ and that the unknown parameters have a clear meaning. Under these assumptions, the posterior marginal distribution in Equation 17 can be written as:

$$p\left(\mu_t^{(c)} \mid c_{it}\right) = \sum_{k=1}^{K} \left[\int p\left(\mu_t^{(c)}, \sigma_t^{(c)} \mid c_{it}, M_{kt}\right) d\sigma_t^{(c)}\right] p(M_{kt} \mid c_{it}). \tag{18}$$

For each treatment group t, let $p(c_{it} \mid \mu_t^{(c)}, \sigma_t^{(c)}, M_{kt})$ and $p(\mu_t^{(c)}, \sigma_t^{(c)})$ be the distribution of the cost data and the prior distribution of the parameters under model M_{kt} in $\mathcal{M}_t$, respectively. Moreover, let $p(M_{kt})$ be the prior model probability of M_{kt} such that $\sum_{k=1}^{K} p(M_{kt}) = 1$ for each t. Then, the corresponding posterior distribution of the parameters under M_{kt} and the posterior model probability of M_{kt}, which need to be substituted in Equation 18, are derived via Bayes' theorem with posterior inferences about $\mu_t^{(c)}$ typically obtained using MCMC methods (O'Hagan and Foster, 2004).

Because different models may produce rather different conclusions in terms of cost-effectiveness, one should expect that the results of Bayesian model averaging is sensitive to which models are included in $\mathcal{M}_t$. Therefore, it is important to assess the sensitivity of the inferences to the choice of the prior distribution for the unknown parameters under the various models in $\mathcal{M}_t$. If all the models considered share the same parameterisation and prior distribution, one can simply vary the hyperprior values for μ_{ct} and σ_{ct} to assess how the posterior model probabilities and the posterior summaries of μ_{ct} from Bayesian model averaging change. Typically, the choice of the prior distribution for the unknown parameters of the models in $\mathcal{M}_t$ has a substantial impact on the inferences and care should be used in eliciting these priors from experts of the problem under consideration (Conigliani and Tancredi, 2009). Finally, Bayesian model averaging could be used to handle structural uncertainty not only about the distribution of cost data, but also about other model assumptions, such as the distribution of effects, the type of relationship between costs and effects, or the prior distributions.

Aggregated Level Data

Economic evaluations based on ALD typically use information about relevant parameters that, however, are not directly the population average benefits and costs. Rather, they are core parameters that may describe disease progression, rate of clinical events, prevalence and incidence of a disease. These parameters are then combined using mathematical models to simulate the expected costs and benefits under different treatments regimens and scenarios (Briggs et al., 2006). When there are multiple sources of aggregate data to inform the estimation of the same model parameter (e.g. relative effectiveness) for a given model, researchers tend to synthesise these using meta-analytic techniques.

This information is usually collected through systematic literature reviews, before being quantitatively synthesised into a single estimate. For example, a systematic review of the literature might inform the baseline prevalence of a given disease, as well as the relative effectiveness of a new intervention. Furthermore, suitable models can be constructed to allow studies of different designs (e.g. RCT and observational studies) to be pooled in order to estimate the quantities of interest. These different pieces of information can then be combined in a decision-analytic model to estimate the incremental benefits Δ_e and costs Δ_c needed to perform the economic analysis.

A popular approach to synthesise information from several studies or evidence sources is based on *hierarchical* or *multilevel* models. Typically, these models assume the existence of J clusters (e.g. studies), each reporting data on n_j units (e.g. individuals) on which an outcome of interest y_{ij} is observed. The underlying idea is that it is possible to learn about clusters made by only a few observations by obtaining some indirect evidence from other (possibly larger) clusters. This feature is particularly relevant in the case of "indirect comparisons", where no head-to-head comparison between two interventions is available, but inference can be made using studies testing each of them against a common comparator (e.g. placebo).

Bayesian modelling is particularly effective to represent multi-level data structures by exploiting conditional exchangeability assumptions in the data (Gelman and Hill, 2007). In general terms, a hierarchical structure is represented by assuming the existence of a cluster-specific parameter θ_j and modelling the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ as draws from a probability distribution characterised by a vector of hyperparameters $\boldsymbol{\psi}$. Hierarchical modelling accounts for both possible levels of correlations (within and between clusters). The process of relating the different clusters in the hierarchical structure is sometimes referred to as "borrowing strength": it is still possible to learn about clusters made by only a few observations by obtaining some indirect evidence from the other (possibly larger) subgroups.

Typically, multilevel models are implemented by first modelling the observed data using some probability distribution $p(y_{ij} \mid \theta_j)$ conditionally on the parameters θ_j. Then (some function of) the parameters is associated with a probability distribution encoding the assumption of exchangeability. For example, consider

$$g(\theta_j) \sim \text{Normal}(\mu_\theta, \sigma^2_\theta),$$

where $g(\cdot)$ can be the identity function for continuous data or the logit function for binary data, and $\boldsymbol{\psi} = (\mu_\theta, \sigma^2_\theta)$ is the vector of hyperparameters identifying the common effect for all the clusters (on which some suitable prior distributions must be specified).

Hierarchical models are commonly used in the process of evidence synthesis, particularly when individual data are not available (Welton et al., 2012; Bujkiewicz et al., 2016). To illustrate the idea underlying Bayesian hierarchical modelling a simplified example taken from Welton et al. (2012) is considered. The model uses the event data r_{jt} (e.g. number of adverse events of interventions) and numbers of individuals n_{jt} in two treatment groups $t = 0, 1$ from J different studies:

$$\begin{aligned} r_{j0} &\sim \text{Binomial}(p_{j0}, n_{j0}) & r_{j1} &\sim \text{Binomial}(p_{j1}, n_{j1}) \\ \text{logit}(p_{j0}) &= \alpha_j & \text{logit}(p_{j1}) &= \alpha_j + \beta_j \\ \alpha_j &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) & \beta_j &\sim \text{Normal}(\mu_\beta, \sigma^2_\beta), \end{aligned} \tag{19}$$

where p_{jt} are the probabilities of an event in the two groups in each of the $j = 1, \ldots, J$ studies, α_j and β_j

are the estimated log-odds of an event in group 0 and between the two groups, $\boldsymbol{\mu} = (\mu_\alpha, \mu_\beta)$ are the overall pooled estimates of α_j and β_j, while $\boldsymbol{\sigma}^2 = (\sigma^2_\alpha, \sigma^2_\beta)$ are the between-study variances. Typically, when the effect size of interest is μ_β, the intercept terms of the logistic regressions in Equation 19 are assigned vague prior distributions by fixing the corresponding hyperparameter values, e.g. $\alpha_j \sim \text{Normal}(0, 10^5)$. Vague prior distributions are then specified for the hyperparameters of the effect sizes, such as $\mu_\beta \sim \text{Normal}(0, 10^5)$ and $\sigma^2_\beta \sim \text{Uniform}(0, 10)$. Alternative priors, especially for the study-level variance parameters σ^2_β, are usually considered to assess the sensitivity of the results to different prior choices (e.g. Half-Normal or Half-Cauchy distributions).

Non-Bayesian methods can also be used to handle multi-level data and are routinely implemented by practitioners in the context of meta-analysis. Three popular approaches are iterative generalised least squares (IGLS) and restricted maximum likelihood (REML) for continuous outcomes and quasi-likelihood (QL) methods for dichotomous outcomes. IGLS is a sequential procedure which relies on an iterative generalised least squares estimation for fitting Normal multilevel models (convergence is assumed to occur when two successive sets of estimates differ by no more than a given tolerance). As with many maximum likelihood procedures, IGLS produces biased estimates in small samples, where the algorithm tends to underestimate the variance of the random-effects $\boldsymbol{\theta}_j$. Bias-adjusted estimates can be obtained by adding correction terms at each iteration of the procedure, which then takes the name of restricted IGLS and coincides with REML in Normal models. Estimated asymptotic standard errors of the estimates are then derived from the final values at convergence of the covariance matrices of the parameters (Goldstein, 1995). For binary outcomes, general multilevel models are typically implemented through QL methods by linearising the model via Taylor series expansion. Estimated asymptotic standard errors for QL estimates are typically derived from a version of the observed Fisher information based on the quasi-likelihood function underlying the estimation process (Breslow and Clayton, 1993)

Compared with these meta-analytic methods, the Bayesian procedure in evidence synthesis guarantees a better characterisation of the underlying variability in the structured parameters, which leads to better precision, e.g. estimations that tend to be unbiased and well calibrated, especially for non-Gaussian data (Browne and Draper, 2006). This is essentially due to the fact that the full uncertainty about the higher level parameters is reflected in the precision of the estimation, while in general non-Bayesian methods such as IGLS or REML produce artificially narrow confidence intervals for the parameters of interest. A further advantage is the possibility of estimating functions of parameters in a relatively straightforward way. For example, by using a MCMC approach it is sufficient to monitor some parameters θ and then define the parameter of actual interest ϕ using a suitable deterministic relationship $\phi = f(\theta)$. Uncertainty on ϕ is automatically accounted for.

Another popular approach for evidence synthesis is based on *multistate* or *Markov* models (Ades, 2003; De Angelis et al., 2014). These are typically used to describe the natural history of a disease through patients' movements or *transitions* over time and a finite (and usually discrete) set of states that are assumed to be representative of the management of the disease. Usually, time is modelled through a set of discrete cycles (e.g. years). Markov models are very often used to model the economic impact of noncommunicable diseases (e.g. cancer or cardiovascular disease).

A typical issue in Markov models is that it may not always be possible to find direct evidence to estimate the relevant transition probabilities between each state of the model. For example, there may be no study evaluating the chance that patients move from one particular state to another state. In these cases, the transition probabilities can be estimated by linking them to some other relevant parameters in the model. In a full Bayesian approach, these parameters can be estimated using the available evidence, perhaps through a synthesis of the available literature. Thus, the resulting transition probabilities are computed as functions of random quantities, which induces a full posterior distribution accounting for the uncertainty in the economic model. This automatically produces a framework for PSA that is particularly straightforward to implement (Spiegelhalter and Best, 2003; Welton and Ades, 2005).

Although decision-analytic models are often entirely informed by ALD, there is no reason why some parameters could not be directly estimated by using (perhaps small) experimental datasets. By incorporating ALD and ILD into a single modelling process, the performance of decision-analytic models can be improved by balancing the strengths and limitations associated with each type of data. For example, ALD are usually population based but do not contain personal level information, while ILD contain rich personal level information, but the sample size is usually small and problems of selection biases and missing data are common. The combination of different types of data types, however, often comes at the price of considering data that may be affected by different types of bias, and thus suitable statistical methods need to be used (Spiegelhalter et al., 2004; Welton et al., 2012; Philippo et al., 2016).

A Bayesian modelling framework is particularly suited to deal with this situation because it allows to build up a series of local submodels, each of which can be based on different sources of data, and to link them together into a coherent global analysis (Spiegelhalter, 1998; Richardson and Best, 2003; Molitor et al., 2009). However, a potential problem may arise when combining ILD and ALD in practice. In some cases, evidence for at least one type of data may be difficult to find or, even if available, it may not be relevant to the research question. Indeed, care is needed to ensure that the data and sources of information being combined within the modelling framework are compatible and can be assumed to be representative of the same underlying population.

Bayesian Methods in Health Economics Practice

HTA has been slow to adopt Bayesian methods; this could be due to a reluctance to use prior opinions, unfamiliarity, mathematical complexity, lack of software, or conservatism of the healthcare establishment and, in particular, the regulatory authorities. However, the use of Bayesian approach has been increasingly advocated as an efficient tool to integrate statistical evidence synthesis and parameter estimation with probabilistic decision analysis in an unified framework for HTA (Ades et al., 2006; Spiegelhalter et al., 2004; Baio, 2012). This enables a transparent "evidence-based" decision modelling, reflecting the uncertainty and the structural relationships in all the available data.

With respect to trial-based analyses, the flexibility and modularity of the Bayesian modelling structure are well-suited to jointly account for the typical complexities that affect ILD. In addition, prior distributions can be used as convenient means to incorporate external information into the model when the evidence from the data is limited or absent (e.g. for missing values). In the context of evidence synthesis, the Bayesian approach is particularly appealing in that it allows for all the uncertainty and correlation induced by the often heterogeneous nature of the evidence (either ALD only or both ALD and ILD) to be synthesised in a way that can be easily integrated within a decision modelling framework.

The availability and spread of Bayesian software among practitioners since the late 1990s, such as `OpenBUGS` (Lunn et al., 2012) or `JAGS` (Plummer, 2010), has greatly improved the applicability and reduced the computational costs of these models. Thus, analysts are provided with a powerful framework, which has been termed *comprehensive decision modelling* (Cooper et al., 2004), for simultaneously estimating posterior distributions for parameters based on specified prior knowledge and data evidence, and for translating this into the ultimate measures used in the decision analysis to inform cost-effectiveness conclusions.

Further Reading

Baio, G. (2012). *Bayesian Methods in Health Economics*. Chapman and Hall/CRC, University College London London, UK.

Baio, G., Berardi, A., and Heath, A. (2017). *Bayesian Cost Effectiveness Analysis with the R package BCEA*. Springer, New York.

Berger, J. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science and Business Media.

Bojke, L., Claxton, K., Palmer, S., and Sculpher, M. (2006). *Defining and characterising structural uncertainty in decision analytic models*. Centre for Health Economics, University of York, York.

Briggs, A., Sculpher, M., and Claxton, K. (2006). *Decision modelling for health economic evaluation*. OUP, Oxford, UK.

Drummond, M., Drummond, M., and McGuire, A. (2001). *Economic evaluation in health care: merging theory with practice*. OUP, Oxford.

Drummond, M., Schulpher, M., Claxton, K., Stoddart, G., and Torrance, G. (2005). *Methods for the economic evaluation of health care programmes. 3rd ed.* Oxford university press, Oxford.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Jones, A., Rice, N., d'Uva, T., and Balia, S. (2012). *Applied health economics (second edition)*. Routledge.

Lindley, D. (1985). *Making Decisions*. Wiley.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley and Sons, West Sussex, UK.

Raiffa, H. (1968). *Decision analysis: introductory lectures on choices under uncertainty*. AddisonWesley, Reading.

Smiith, J. (1988). *Decision Analysis: A Bayesian approach*. Chapman and Hall, London, UK.

Spiegelhalter, D., Abrams, K., and Myles, J. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley and Sons, Chichester, UK.

Welton, N., Sutton, A., Cooper, N., and Adams, K. (2012). *Evidence Synthesis for Decision Making in Healthcare*. Wiley, UK.

References

Ades, A. (2003). A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine*, 22:2995–3016.

Ades, A., Sculpher, M., Sutton, A., Abrams, K., Cooper, N., Welton, N., and Lu, G. (2006). Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics*, 24:1–19.

Bailey, J., Webster, R., Hunter, R., Griffin, M., N., F., Rait, G., Estcourt, C., Michie, S., Anderson, J., Stephenson, J., Gerressu, M., Sinag Ang, C., and Murray, E. (2016). The men's safer sex project: intervention development and feasibility randomised controlled trial of an interactive digital intervention to increase condom use in men. *Health Technology Assessment*, 20.

Baio, G. (2012). *Bayesian Methods in Health Economics*. Chapman and Hall/CRC, University College London London, UK.

Baio, G. (2014). Bayesian models for cost-effectiveness analysis in the presence of structural zero costs. *Statistics in Medicine*, 33:1900–1913.

Baio, G., Berardi, A., and Heath, A. (2017). *Bayesian Cost Effectiveness Analysis with the R package BCEA*. Springer, New York.

Baio, G. and Dawid, A. (2015). Probabilistic sensitivity analysis in health economics. *Statistical Methods in Medical Research*, 24:615–634.

Barber, J. and Thompson, S. (2000). Analysis of cost data in randomised trials: an application of the non-parametric bootstrap. *Statistcis in Medicine*, 19:3219–3236.

Barton, G., Briggs, A., and Fenwick, E. (2008). Optimal cost-effectiveness decisions: the role of the cost-effectiveness acceptability curve (ceac), the cost-effectiveness acceptability frontier (ceaf), and the expected value of perfection information (evpi). *Value in Health*, 11(5):886–897.

Basu, A. (2009). Individualization at the heart of comparative effectiveness research: the time for i-cer has come. *Medical Decision Making*, 29:9–11.

Basu, A. and Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, 1:56–69.

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, 27:112–127.

Bernardo, J. and Smith, A. (1999). *Bayesian Theory*. Wiley, New York.

Bilcke, J., Beutels, P., Brisson, M., and Jit, M. (2011). Accounting for methodological, structural, and parameter uncertainty in decision- analytic models: A practical guide. *Medical Decision Making*, 31:675–692.

Black, W. (1990). A graphic representation of cost-effectiveness. *Medical Decision Making*, 10:212–214.

Bojke, L., Claxton, K., Palmer, S., and Sculpher, M. (2006). *Defining and characterising structural uncertainty in decision analytic models*. Centre for Health Economics, University of York, York.

Bojke, L., Claxton, K., Sculpher, M., and Palmer, S. (2009). Characterizing structural uncertainty in decision analytic models:a review and application of methods. *Value in Health*, 12:739–749.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Briggs, A., Sculpher, M., and Claxton, K. (2006). *Decision modelling for health economic evaluation*. OUP, Oxford, UK.

Browne, W. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1:473–514.

Bujkiewicz, S., Thompson, J., Riley, L., and Abrams, K. (2016). Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in Medicine*, 35:1063–1089.

Burton, A., Billingham, L., and Bryan, S. (2007). Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. *Clinical Trials*, 4:154–161.

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. John Wiley and Sons, Chichester, UK.

Claxton, K. (1999). The irrelevance of inference: a decision making approach to stochastic evaluation of health care technologies. *Journal of Health Economics*, 18:342–364.

Claxton, K. (2001). Bayesian value of information analysis. *International Journal of Technology Assessment in Health Care*, 17:38–55.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Hakehurst, R., Buxton, M., Brazier, J., and O'Hagan, T. (2005). Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Heath Economics*, 27:339–347.

Conigliani, C. and Tancredi, A. (2009). A bayesian model averaging approach for cost-effectiveness analyses. *Health Economics*, 18:807–821.

Cooper, N., Sutton, A., Abrams, K., Turner, D., and Wailoo, A. (2004). Health economics. *Medical Decision Making*, 23:203–226.

De Angelis, D., Presanis, A., Conti, S., and Ades, A. (2014). Estimation of hiv burden through bayesian evidence synthesis. *Statistical Science*, pages 9–17.

Dias, S., Sutton, A., Welton, N., and Ades, A. (2013). Evidence synthesis for decision making 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis. *Medical Decision Making*, 33:671–678.

Diaz-Ordaz, K., Kenward, M., and Grieve, R. (2014). Handling missing values in cost effectiveness analyses that use data from cluster randomized trials. *Journal of the Royal Statistical Society: Series A*, 177:457–474.

Drummond, M., Manca, A., and Sculpher (2005a). Increasing the generalizability of economic evaluations: Recommendations for the design, analysis, and reporting of studies. *International Journal of Technology Assessment in Health Care*, 21:165–171.

Drummond, M., Schulpher, M., Claxton, K., Stoddart, G., and Torrance, G. (2005b). *Methods for the economic evaluation of health care programmes. 3rd ed.* Oxford university press, Oxford.

Erler, N., Rizopoulos, D., Van Rosmalen, J., Jaddoe, V., Francob, O., and Lesaffre, E. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full bayesian approach. *Statistics in Medicine*, 35:2955–2974.

Espinoza, M., Manca, A., Claxton, K., and Sculpher, M. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: Conceptual framework and application. *Medical Decision Making*, 34:951–964.

Faria, R., Gomes, M., Epstein, D., and White, I. (2014). A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *PharmacoEconomics*, 32:1157–1170.

Fenwick, E., Claxton, K., and Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics*, 10(8):779–787.

Franklin, M., Davis, S., Horspool, M., Sun Kua, W., and Julious, S. (2017). Economic evaluations alongside efficient study designs using large observational datasets: the pleasant trial case study. *PharmacoEconomics*, 35:561–573.

Gabrio, A., Mason, A., and Baio, G. (2017). Handling missing data in within-trial cost-effectiveness analysis: A review with future recommendations. *PharmacoEconomics-Open*, 1:79–97.

Gabrio, A., Mason, J., and Baio, G. (2018). A full bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. `https://arxiv.org/abs/1801.09541`.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Goldstein, H. (1995). *Multilevel Statistical Models - Second Edition*. Edward Arnold, London.

Groenwold, R., Rogier, A., Donders, T., Roes, K., Harrell, F., and Moons, K. (2012). Dealing with missing outcome data in randomized trials and observational studies. *American Journal of Epidemiology*, 175:210–217.

Harkanen, T., Maljanen, T., Lindfors, O., Virtala, E., and Knekt, P. (2013). Confounding and missing data in cost-effectiveness analysis: comparing different methods. *Health Economics Review*, 3.

Hay, J., Jackson, J., Luce, B., Avorn, J., and Ashraf, T. (1999). Panel 2: Methodological issues in conducting pharmacoeconomic evaluations – modeling studies. *Value in Health*, 2:78–81.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417.

Howard, R. (1966). Information value theory. *IEEE Trans. Syst. Sci. Cybern.*, 2:22–26.

Hunter, R., Baio, G., Butt, T., Morris, S., Round, J., and Freemantle, N. (2015). An educational review of the statistical issues in analysing utility data for cost-utility analysis. *PharmacoEconomics*, 33:355–366.

International Network of Agencies for Health Technology Assessment (2018). `http://www.inahta.org/HTA./`. Accessed July 2018.

Jackson, C., Sharples, L., and Thompson, S. (2010). Structural and parameter uncertainty in bayesian cost-effectiveness models. *Appl. Statist*, 59:233–253.

Jackson, C., Thompson, S., and Sharples, L. (2009). Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society: Series A*, 172:383–404.

Koerkamp, B., Hunink, M., Stijnen, T., Hammitt, J., Kuntz, K., and Weinstein, M. (2007). Limitations of acceptability curves for presenting uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 27:101–111.

Leurent, B., Gomes, M., and Carpenter, J. (2018). Missing data in trial-based cost-effectiveness analysis: An incomplete journey. *Health Economics*.

Loomes, G. and McKenzie, L. (1989). The use of qalys in health care decision making. *Soc. Sci. Med.*, 28:299–308.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

Manca, A., Hawkins, N., and Sculpher, M. (2005). Estimating mean qalys in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics*, 14:487–496.

Manca, A. and Palmer, S. (2005). Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials. *Appl Health Econ Health Policy*, 4:65–75.

Mason, A., Gomes, M., Grieve, R., and Carpenter, J. (2018). A bayesian framework for health economic evaluation in studies with missing data. *Health Economics*, 27:1670–1683.

Mason, A., Gomes, M., Grieve, R., Ulug, P., Powell, J., and Carpenter, J. (2017). Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the improve trial. *Clinical Trials*, 14:357–367.

Mihaylova, B., Briggs, A., O'Hagan, A., and Thompson, S. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20:897–916.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (2015). *Handbook of Missing Data Methodology*. Chapman and Hall, Boca Raton, FL.

Molitor, N., Best, N., Jackson, C., and Richardson, S. (2009). Using bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by products. *J. R. Statist. Soc: Series A*, 172:615–637.

NICE (2013). *Guide to the Methods of Technological Appraisal*. NICE, London, UK.

Nixon, R. and Thompson, S. (2005). Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics*, 14:1217–1229.

Noble, S., Hollingworth, W., and Tilling, K. (2012). Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Economics*, 21:187–200.

Ntzoufras, I. (2009). *Bayesian Modelling Using WinBUGS*. John Wiley and Sons, New York, US.

O'Hagan, A. and Foster, J. (2004). *Bayesian Inference, Kendall's Advanced Theory of Statistics, Second Edition*. Arnold, London.

O'Hagan, A. and Stevens, J. (2001). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics*, 10:303–315.

Philippo, D., Ades, A., Dias, S., Palmer, S., Abrams, K., and Welton, N. (2016). *NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submissions to NICE*. London: Natl. Inst.Health Care Excell.

Plummer, M. (2010). JAGS: Just Another Gibbs Sampler. `http://www-fis.iarc.fr/~martyn/software/jags/`.

Ramsey, S., Willke, R., Glick, H., Reed, S., Augustovski, F., Johnsson, B., Briggs, A., and Sullivan, S. (2015). Cost-effectiveness analysis alongside clinical trials ii-an ispor good research practices task force report. *Value in Health*, 18:161–172.

Richardson, S. and Best, N. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, 14:129–147.

Ridyard, C. and Hughes, D. (2010). Methods for the collection of resource use data within clinical trials: A systematic review of studies funded by the uk health technology assessment program. *Value in Health*, 13:867–872.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York,USA.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, New York, US.

Schafer, J. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioural Research*, 33:545–571.

Sculpher, M., Claxton, K., Drummond, M., and McCabe, C. (2005). Whither trial-based economic evaluation for health decision making? *Health Economics*, 15:677–687.

Snowling, S. and Kramer, J. (2001). Evaluating modelling uncertainty for model selection. *Ecological Modelling*, 138:17–30.

Spiegelhalter, D. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics*, 47:115–133.

Spiegelhalter, D., Abrams, K., and Myles, J. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley and Sons, Chichester, UK.

Spiegelhalter, D. and Best, N. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, 22:3687–3709.

Stinnett, A. and Mullahy, J. (1998). A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18:68–80.

Thompson, S. and Nixon, R. (2005). How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making*, 4:416–423.

Thorn, J., Coast, J., Cohen, D., Hollingworth, W., Knapp, M., and Noble, S. (2013). Resource-use measurement based on patient recall: issues and challenges for economic evaluation. *Appl Health Econ Health Policy*, 11:155–161.

Van Asselt, A., van Mastrigt, G., Dirksen, C., Arntz, A., Severens, J., and Kessels, A. (2009). How to deal with cost differences at baseline. *PharmacoEconomics*, 27:519–528.

Van Gestel, A., Grutters, J., Schouten, J., Webers, C., Beckers, H., Joore, M., and Severens, J. (2012). The role of the expected value of individualized care in cost-effectiveness analyses and decision making. *Value in Health*, 15:13–21.

Van Hout, B., Al, M., Gordon, G., Rutten, F., and Kuntz, K. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics*, 3:309–319.

Weinstein, M., O'brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., and Luce, B. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ispor task force on good research practices – modeling studies. *Value in Health*, 6:9–17.

Welton, D. and Ades, A. (2005). Estimation of markov chain transition probabilities and rates from fully and partially observed data: Uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, 25:633–645.

Welton, N., Sutton, A., Cooper, N., and Adams, K. (2012). *Evidence Synthesis for Decision Making in Healthcare*. Wiley, UK.

Welton, N. and Thom, H. (2015). Value of information: We' ve got speed, what more do we need? *Medical Decision Making*, 35:564–566.

Willan, A., Briggs, A., and Hock, J. (2005). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics*, 13:461–475.

Wood, A., White, I., and Thompson, S. (2004). Are missing outcome data adequately handled?a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376.