Full Bayesian Methods to Handle Missing Data in Health Economic Evaluation

Andrea Gabrio

University College London (UCL)

Department of Statistical Science

A thesis presented for the degree of

Doctor of Philosophy in Statistics

Declaration of Authorship

I, Andrea Gabrio, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: _____

Date: _____

Abstract

Trial-based economic evaluations are performed on individual-level data, which almost invariably contain missing values. Missingness represents a threat for the analysis because any statistical method makes assumptions about the unobserved values that cannot be verified from the data at hand; when these assumptions are not realistic, they could lead to biased inferences and mislead the cost-effectiveness assessment.

We start by investigating the current missing data handling in economic evaluations and provide recommendations about how information about missingness and related methods should be reported in the analysis. We illustrate the pitfalls and issues that affect the methods used in routine analyses, which typically do not account for the intrinsic complexities of the data and rarely include sensitivity analysis to the missingness assumptions. We propose to overcome these problems using a full Bayesian approach. We use two case studies to demonstrate the benefits of our approach, which allows for a flexible specification of the model to jointly handle the complexities of the data and the uncertainty around the missing values.

Finally, we present a longitudinal bivariate model to handle nonignorable missingness. The model extends the standard approach by accounting for all observed data, for which a flexible parametric model is specified. Missing data are handled through a combination of identifying restrictions and sensitivity parameters. First, a benchmark scenario is specified and then plausible nonignorable departures are assessed using alternative prior distributions on the sensitivity parameters. The model is applied to and motivated by one of the two case studies considered.

Impact Statement

My PhD research has focussed on the handling of missing data in health economic evaluation, mainly from a Bayesian statistical perspective. Through two real case examples, I have identified the limitations of the standard approach used by practitioners and proposed an alternative statistical framework to perform economic evaluations that can improve the current practice. This has resulted in one first author publication and three original manuscripts that have already been submitted for publication to different academic journals.

My research has important implications within the health economics community due to the fact that considerable proportions of missing data often occur in trial-based analyses, but their impact on the results is rarely assessed. The results from this thesis show that failure to conduct sensitivity analysis to plausible missingness assumptions can have substantial implications in terms of decision-making and could lead to different cost-effectiveness assessments. This is a problem of great interest to many clinicians, health economists and, crucially, decision-makers (e.g. NICE in the UK), who typically use the results from these analyses to inform resource allocation decisions and the funding of new healthcare technologies. A statistical software package is under development to facilitate the implementation of the methods presented in this thesis in routine analyses among practitioners and make them available to a wider audience.

During my PhD I have been invited to present my work at the Centre for Statistical Methodology Early Career Researcher Showcase (London School of Hygiene and Tropical Medicine, London, 2018). I have given oral presentations at the PRIMENT Statistics, Health Economics and Methodology Seminar (UCL, London, 2018) as well as poster presentations at the Third European Health Economics Association (EuHEA) PhD student-supervisor and early career researcher conference (Universitat Internacional de Catalunya, Barcelona, 2016) and at the Health Economics Symposium (UCL, London, 2018). I have also been awarded the Costas Goutis Prize (2017) and a research grant from the Foundation BLANCEFLOR Boncompagni - Ludovisi née Bildt (2015-2018).

Contents

List of Tables

List of Figures

Glossary

AC Available Cases

CC Complete Cases

CCA Complete Case Analysis

CEA Cost-Effectiveness Analysis

CEAC Cost-Effectiveness Acceptability Curve

CEP Cost-Effectiveness Plane

CUA Cost Utility Analysis

DIC Deviance Information Criterion

EIB Expected Incremental Benefit

EQ-5D EuroQol-5D

IB Incremental Benefit

ICER Incremental Cost-Effectiveness Ratio

LVCF Last Value Carried Forward

MAR Missing At Random

MCAR Missing Completely At Random

MCMC Markov Chain Monte Carlo

MenSS Men's Safer Sex

MI Multiple Imputation

MNAR Missing Not At Random

NHS National Health Service

NICE National Institute for Health and Care Excellence

PBS Positive Behaviour Support

QALY Quality Adjusted Life Years

RCT Randomised Controlled Trial

SI Single Imputation

Research Question and Outline of the Thesis

Individual-level data in health economic evaluations are generally characterised by some missing outcome values. If these unobserved values are not appropriately handled, they may affect the inferences and possibly mislead the cost-effectiveness assessment. The modelling task is particularly challenging as the problem of handling missingness is often embedded within a more complex framework, where outcome data typically present a series of complexities that need to be simultaneously addressed to avoid biased results.

Trial-based routine analyses do not typically account for all these complexities, are conducted on cross-sectional quantities based only on the complete cases and rarely assess the robustness of the results to different assumptions about the missing values. The failure to appropriately account for the uncertainty generated by missingness may have important consequences on the results and, more importantly, on the approval or reimbursement of new health care technologies.

Bayesian methods are well-suited for addressing decision-making problems. By taking a probabilistic approach, based on decision rules and available information, they can explicitly account for relevant sources of uncertainty in the decision process and obtain an "optimal" decision output. In addition, the flexibility of the Bayesian approach can handle the complexities of the data in a relatively easy way and naturally allows the incorporation and assessment of transparent assumptions about the missing data.

The aim of this research is to develop a full Bayesian approach to handle missingness in health economic evaluations and compare its performance with respect to the standard approach used by practitioners. We address the research question based on three key **objectives**: **1)** to review the missingness methods used in trial-based economic evaluations and evaluate the quality of routine analyses with respect to the handling of missing data; **2)** to identify potential limitations of the standard approach in terms of unrealistic missing data assumptions and provide a Bayesian approach that can improve the current practice and avoid biased results; **3)** to develop a full Bayesian approach to handle missingness in a principled way by combining a model for the data and explicit assumptions about the missing values.

The rest of the thesis is structured as follows. Chapter 1 summarises the theoretical background related to the main topics of this thesis. First, we introduce the health economics evaluation framework and purpose; then we provide a brief summary of some key concepts of Bayesian analysis and its advantages in dealing with decision-making problems; finally, we present the topic of missing data analysis and we review some of the most popular methods and approaches to handle missingness. Chapter 2 shows the results from a literature review on missingness methods in trial-based analyses and provides recommendations to improve the current practice. In addition, guidelines for assessing the quality of missing data analyses are provided in the form of a structural framework, which is described and applied to the articles studied in the review (**objective** 1). Chapter 3 presents the two case studies (MenSS and PBS trials) that will be analysed in this thesis and describes the standard approach for performing economic evaluations in

routine analyses. Chapter 4 illustrates a pitfall related to the different missing data assumptions associated with alternative implementations of the mean baseline adjustment methods used in routine analyses. Data from both case studies are used to demonstrate the potential bias associated with this method (**objective** 2). Chapter 5 presents a general Bayesian analytic framework that improves the standard approach and leads to more realistic imputations by jointly tackling the typical complexities that affect the data. We demonstrate the benefits and flexibility of our approach on the data from the two case studies (**objective** 2). Chapter 6 proposes a parametric Bayesian longitudinal approach that extends the modelling framework of economic evaluations to handle missingness more efficiently, while incorporating a sensitivity analysis to alternative missingness assumptions. We motivate and apply our approach using the data from the PBS trial (**objective** 3). Finally, Chapter 7 summarises the main conclusions from this thesis and suggests directions for future research.

Chapter 1

Background

We first introduce health economic evaluations, the type of data analysed and the decision-making problem involved in the cost-effectiveness assessment. Next, we briefly review some key concepts of Bayesian inference and describe the advantages of using a Bayesian approach to account for multiple forms of uncertainty in economic evaluations. Finally, we present the topic of missing data from a general statistical perspective, review some of the most popular methods used to handle missingness and focus on a principled approach for conducting sensitivity analysis to plausible missingness assumptions.

1.1 Health Economic Evaluation

Economic evaluations are applied in the field of healthcare with the principle aim of improving the economic efficiency of resource allocation, i.e. help maximise benefits from available (and constrained) resources. This has become an increasingly important problem in many countries over the last decades due to a continuous increase in the costs associated with healthcare services that now affect a great portion of the total economy expenditures (OECD, 2015). Naturally, this is the result of an increase of the average life expectancy and the development of new and more expensive technologies. A major consequence of this process is the need to balance the total healthcare spending, i.e. to define what is the optimum expenditure level and specify how to reach it. In countries where there is a predominant public funding of healthcare, such as the UK, this is typically achieved by defining a governmental health scheme that can provide the highest clinical benefit level for the patients and society, given the availability of limited resources.

In the UK, the *National Institute for Health and Care Excellence* (NICE) plays an important role with regard to the approval of new healthcare technologies and uses economic evaluations to inform decisions about a wide range of interventions: medical devices, diagnostic technologies, surgical procedures and pharmaceuticals (NICE, 2013). NICE provides evidence-based guidelines on how a particular disease or condition should be treated and assesses whether new treatments provide value for money as they become available in England and Wales (Meltzer, 2001).

The increasing use of economic evaluation for decision-making has placed requirements on the type of evidence and analytic methods to be used in order to define an appropriate framework for the analysis (Sculpher et al., 2005). NICE typically relies on decision models to evaluate the cost-effectiveness of new treatment regimens. These models are based on information collected from the literature, among which a key role is played by *Randomised Controlled Trials* (RCTs). They provide individual patient data where the randomisation of patients to treatment acts to reduce bias, and therefore can be used to perform head-to-head comparisons in controlled envi-

ronments. As a result, RCTs are commonly used as a vehicle for economic evaluations. Many funders, such as the UK *National Institute for Health Research Health Technology Assessment Programme*, routinely request that assessments of cost effectiveness are incorporated in the design of randomised trials to inform policy makers about the feasibility of extending the treatment to the overall target population.

Economic evaluations can be formally defined as the comparison of alternative options in terms of both their *costs* and *benefits* (Drummond et al., 2005). The joint consideration of both outcome measures represents a key element in determining whether a new treatment option should be given priority in terms of resource allocation with respect to an alternative. The most popular types of economic evaluation in healthcare are *Cost-Effectiveness Analysis* (CEA) and *Cost-Utility Analysis* (CUA), which share a similar rationale but typically differ by the types of measures used to describe the benefits.

In CEA, the benefits of an intervention are measured in terms of a pre-defined unit of health outcome such as lives saved or life years gained and the task of an analyst performing the evaluation is to estimate the cost per unit of health outcome achieved, i.e. the cost per life saved. CEA does not permit a direct comparison of costs and benefits across interventions yielding different outcomes (for instance, cases prevented vs. life years gained) but is restricted to the comparisons of interventions that use the same disease-specific outcome measures. However, healthcare providers, such as the *National Health Service* (NHS) in the UK, require the comparison of interventions across different disease areas to inform resource allocation and prioritisation decisions.

To avoid the problem of non-comparability, benefits in CUA are expressed in terms of utility or quality of life. Among these non-monetary measures, one of the most popular is *Quality Adjusted Life Years* (QALYs), an index comprising both length and quality of life. Although this has been debated (Mooney, 1989; Neumann and Greenberg, 2009), it is generally assumed that the QALY is a comprehensive measure of health that captures enough aspects of health to be considered an appropriate instrument for measuring outcomes in the field of curative healthcare.

More specifically, QALYs measure the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life (one QALY is generally associated with one year of life in perfect health). QALYs are calculated by estimating the years of life remaining for a patient following a particular treatment or intervention and weighting each year with a quality-of-life or utility score (see Section 1.2 for more details), which is often measured in terms of the person's ability to carry out the activities of daily life, and freedom from pain and mental disturbance.

The result of a CUA is usually expressed in terms of total net cost per unit of utility or quality (e.g. cost per QALY gained). As in CEA, the value of measures is still implicitly defined as part of the QALY component but has the general advantage of being used across many different disease areas. This makes CUA very attractive because effectiveness, measured in terms of life expectancy, is a straightforward concept for most clinicians and utilities are a quantitative measure about the strength of patients' preferences for certain health states. However, issues arise when calculating utility measures due to a variety of available statistical methodologies as well as different types of patients' data that could be used (Hunter et al., 2015). Although CUA is more resource and time consuming than CEA, it is the recommended analytic framework for economic evaluations in many jurisdictions such as the UK.

Due to an established practice, especially for analyses on individual-level data, practitioners often use the term CEA to also indicate a cost-utility analysis. Throughout this thesis we will focus on trial-based analyses and use the terms economic evaluations, cost-effectiveness analysis and cost-utility analysis interchangeably, as is commonly done.

1.2 Individual-Level Data

A multivariate outcome constitutes the main focus of economic evaluations: the costs and benefits associated with the treatments under examination. We now present a brief description of how these individual-level variables are typically computed and the features that characterise the data.

In a trial setting, e.g. RCTs, health benefit and resource use data are typically collected on each individual in the study ($i = 1, \ldots, n$) at baseline ($j = 0$) and at successive follow-up times ($j = 1, \ldots, J$) for each treatment group t. Data on some baseline demographic variables (e.g. age, gender, ethnicity, etc.) are also typically collected. Table 1.1 displays an example of a typical trial-based individual level dataset for economic evaluation.

		Demographics			**utility data**				**cost data**			
Individual	t	Gender	Age	...	u_0	u_1	...	u_J	c_0	c_1	...	c_J
1	1	M	23	...	0.32	0.66	...	0.44	£103	£241	...	£80
2	1	M	21	...	0.12	0.16	...	0.38	£1204	£1808	...	£877
3	2	F	19	...	0.49	0.55	...	0.88	£16	£12	...	£22
4	1	F	20	...	0.23	0.37	...	0.52	£99	£150	...	£85
...	...	...	...	...	...	...	...	...	...	...	...	...

Table 1.1: Example of a typical trial-based dataset used in economic evaluations.

Different types of effectiveness or benefit measures can be considered, even though decision-making bodies typically favour outcomes that are comparable across as many different disease areas as possible. Generic preference-based measures of *health related quality of life* are typically used for economic evaluations, and are obtained from short health questionnaires that measure patients' health and well-being across a number of domains. The questionnaire most favoured in the UK is the *EuroQol 5D* (EQ-5D, `http://euroqol.org`) and the variant currently most commonly used is the 3-level version. The EQ-5D is constructed on the basis of five different domains (mobility, self-care, usual activities, pain and anxiety/depression) for which patients are asked whether they have no, some or extreme problems (three levels), for a total of $3^5 = 243$ potential distinct health states. For example, the health state (11223) is associated with an individual who reports no problems (level 1) in the mobility and self-care domains, some problems (level 2) in usual activities and pain, and extreme problems (level 3) in anxiety/depression.

Each health state is then associated with a country-specific utility score representing the preferences of a sample of the general population for that specific health state. Utility scores are calculated using preference-based algorithms which typically anchor the scores so that a value of 1 corresponds to perfect health and a value of 0 is equivalent to the state of death; sometimes, depending on the specific method used, negative scores are possible, representing states that are theoretically worse than death. For example, in the UK, the utility scores for the EQ-5D 3-level version range from a value of 1 for perfect health to -0.594, which corresponds to the worst possible health state. NICE's recommended algorithm to calculate the utilities is the *time-trade off* (Dolan and Gutex, 1995), in which the utility scores associated with health states considered better or worse than death are valued using different approaches. Figure 1.1 shows how the algorithm derives the utility scores associated with an hypothetical impaired health state h, which is either valued better (panel a) or worse (panel b) than death.

For an impaired state h, which is considered better than death (panel a), the respondent faces a choice between two hypothetical lives: one involving x years of healthy life, followed by death (alternative 1); the other involving t years in state h (where $x \leq t$), followed by death (alternative 2). If the respondent prefers alternative 2 to alternative 1, x is increased to make alternative 1 more attractive; if the respondent prefers alternative 1 to alternative 2, x is reduced to make alternative 1 less attractive. This iterative procedure continues until the respondent is unable to

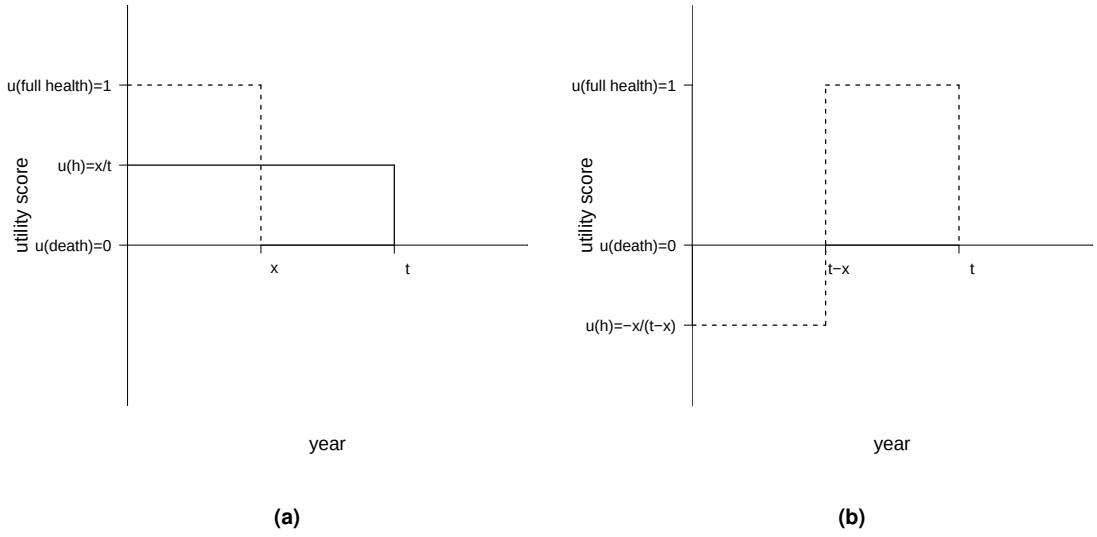


Figure 1.1: Schematic representation of the time-trade off algorithm to derive the utility scores associated with some impaired health state, which is either valued better (panel a) or worse (panel b) than death. Legend: h impaired health state, $u(h)$ utility score of state h, x time in full health, t time in state h.

choose between the two lives. The utility value of h, i.e. $u(h)$, is calculated according to how much healthy time the respondent is willing to forgo at this point of indifference, and is given by x/t. If the state h is considered by the respondent to be worse than death (panel b), the respondent is presented with a different choice: between a life involving $t - x$ years in h, followed by x healthy years and then death (alternative 1); and immediate death (alternative 2). The value of x is varied until the respondent's point of indifference is identified, where $u(h) = -x/(t - x)$. Dolan and Gutex (1995) validated the time-trade off algorithm on a representative sample of the general UK population to derive a preference-based tariff which directly provides the utility scores associated with each of the health states of the EQ-5D questionnaire. For example, using this tariff, the utility score associated with the state (11223) is equal to 0.255.

The decision of which types of resource use are included in an economic analysis is usually tailored to the requirements of the decision-makers, i.e. the target audience of the evaluation. For a number of countries, healthcare decision-making bodies only consider healthcare costs when assessing the cost-effectiveness of a new intervention. However, in principle, other costs could be included, e.g. societal costs. In addition, the types of cost and resource data collected can be different (patient questionnaires, clinic records, administrative records), depending on the perspective of the decision-maker.

Aggregate measures of effectiveness (e.g. QALYs) and total costs for each individual and treatment option are typically computed by combining the utilities u_{ijt} and costs c_{ijt} collected at each time point into cross-sectional quantities (e_{it}, c_{it})[1] as

$$e_{it} = \sum_{j=1}^{J} \left(u_{ijt} + u_{ij-1t}\right) \frac{x_j}{2} \quad \text{and} \quad c_{it} = \sum_{j=1}^{J} c_{ijt}, \tag{1.1}$$

where $x_j = \frac{\text{Time}_j - \text{Time}_{j-1}}{\text{Unit of time}}$ is the fraction of the time unit (e.g. 12 months) between consecutive measurements, e.g. $x_2 = \frac{(6 \text{ months} - 3 \text{ months})}{12 \text{ months}} = 0.25$. For the utilities, this approach is often referred to as the *area under the curve* (Drummond et al., 2005).

Essentially, QALYs are calculated as a series of preference-weighted health states, where the

[1] We slightly abuse the notation and denote the longitudinal and aggregate cost variables with c_{ijt} and c_{it}, respectively.

weights (i.e. the utility scores at each time point) reflect the desirability of living in those states. Once the weights are derived, they are multiplied by the time spent in the associated state and these products are then summed to obtain the QALYs. Thus, the QALY values associated with an individual, and their range, depend both on the utility scores at each time point j and the time horizon J considered. For example, given a time duration of 1 year, the QALY value associated with an individual who has always lived in perfect health (i.e. utility scores of 1 at each time point j) is equal to 1. However, if a time duration of 2 years is considered, the QALY value associated with the same individual living in perfect health throughout this period is equal to 2. Similarly, the QALY value associated with an individual living in the worst possible health state ($u_j = -0.594$ at each j) is equal to -0.594 when $J = 1$ year and -1.188 when $J = 2$ years. When the QALYs and total cost outcomes are evaluated over a time period longer than 1 year, their values are typically discounted using some yearly discount rate r to account for time preferences in the receipt of costs and benefits, where the value for the discount rate currently recommended by NICE is equal to 3.5% (NICE, 2013). Since for both the case studies analysed in this thesis the time horizon is equal to 1 year, the range of the QALYs always coincides with that of the utility scores, i.e. between -0.594 and 1, and no discounting is required for both QALYs and total costs.

In general, cost data can vary widely between individuals, are defined on the range $[0, \infty)$ and tend to be positively skewed. The skewness of cost data is typically due to the fact that for many evaluations a smaller number of patients will accrue substantially higher costs compared to other patients. This may be due to, for example, long inpatient stays or expensive interventions. It is also common to observe individuals who are associated with a null cost and that induces a spike at 0 in the cost distribution.

In the UK, the utility scores for the EQ-5D 3-level version are defined on the range $[-0.594, 1]$, where negative values are associated with health states that are considered worse than death (Dolan and Gutex, 1995). Among the general population, utility data tend to be negatively skewed, with most of the values lying at the higher end of the measurement scale and some observations displaying extremely low utility levels. Similarly to the costs, some individuals are typically associated with a perfect health state that induces a spike at 1 in the utility distribution. Right-skewed distributions of utilities are occasionally observed among certain groups of patients (e.g. terminally ill patients or individuals with chronic conditions) where most of the individuals in the sample report poor health states.

A typical feature of individual-level utility and cost data is that they can be either positively or negatively correlated. The first case may arise, for example, when effective treatments are innovative and are associated with higher unit costs. The second case, instead, may result when more effective treatments reduce total care pathway costs e.g. by reducing hospitalisations, side effects, etc. In addition, when the data are collected from different centres or clusters, utilities and costs can be differently correlated at the individual and cluster level.

This typically occurs in cluster randomised trials, where the unit of randomization is the "cluster" – for example, the hospital or primary care physician, not the individual. A cluster design may be chosen because the intervention operates at a group rather than at an individual level (e.g. changing incentives for providers) or if there is a high risk of "contamination" among the individuals within clusters (e.g. evaluating different advertising strategies to encourage smoking cessation). In cluster RCTs, individuals within a cluster are likely to be somewhat similar in their characteristics and the care they receive, and therefore, individual utilities or costs within the same cluster tend to be more homogeneous than those in different clusters (Gomes et al., 2012a). For example, cost data based on process measures such as length of stay, typically have a relatively high proportion of the variation at the cluster rather than at the individual level (Campbell et al., 2005). Thus, in cluster RCTs, the size or direction of the correlation may differ according to whether the focus is at the individual or the cluster level (Gomes et al., 2011). For example, within

clusters, individuals with lower health status may incur higher costs (i.e. at the individual level, there is a strong negative correlation), while clusters that have higher mean costs per patient may have on average higher utilities.

Finally, some observations for either or both outcomes, are almost invariably missing. Reasons for missingness may differ according to the context considered or the individuals' characteristics (e.g. age, sex, etc.) and outcome values. For example, when data for utilities and costs are derived from similar types of sources throughout the trial, e.g. using self-reported questionnaires, then missingness at a given time typically occurs in both outcomes, e.g. when an individual drops out from the study. However, in trial-based analyses, outcome data may also be derived using different types of sources, e.g. EQ-5D questionnaires for the utilities and a combination of observational datasets and NHS average unit prices for the costs. In this case, reasons and patterns of missing data may be different between utilities and costs and missingness in one outcome may imply or be related to missingness in the other outcome. In addition, when utility and/or cost data for the $i-$th individual in the study are not observed at all time points, then (e_{it}, c_{it}) cannot be directly computed as in Equation 1.1 for those subjects and are recorded as missing. When the data are partially-observed or "incomplete", there are important implications for their analysis that cannot be ignored (we address the implications of missing data analysis in Section 1.5).

1.3 Bayesian Analysis in Economic Evaluation

From the statistical point of view, trial-based CEAs have historically been performed using frequentist methods: these include power calculations at the design stage and calculation of p-values and confidence intervals (Briggs and Gray, 1998; Laska et al., 1999; Willan, 2001; Glick, 2011). However, the increasing sophistication of economic evaluations is highlighting the limitations of this approach. For example, unlike standard statistical analyses, economic evaluations do not just focus on estimation (e.g. the computation of point or interval estimation, or hypothesis testing), but are used as a tool to aid decision making (Claxton, 1999). Thus, rather than relying merely on statistical and clinical significance, economic evaluations need 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 effectiveness of a new intervention affects the decision about whether it is paid for by the public provider). To this aim, much of the recent research has been oriented towards building the economic evaluation on sound statistical decision-theoretic foundations (Spiegelhalter et al., 2004; Briggs and Gray, 1999), and increasingly often under a Bayesian statistical approach (O'Hagan and Stevens, 2001; Baio, 2012). In particular, NICE advocates the use of this decision-theoretic framework as a standardised approach in health economic evaluations to ensure the comparability of results and the consistency of decision-making (NICE, 2013). Therefore, although alternative approaches to decision-making exist in other application areas, the thesis discusses and focuses on the decision-theoretic approach recommended by NICE, to which all economic evaluations in the UK are expected to adhere.

There are several reasons which make the use of the Bayesian approach in economic evaluations particularly appealing. First, Bayesian methods are naturally embedded in the wider scheme of decision theory; by taking a probabilistic approach, based on decision rules and available information, they can explicitly account for relevant sources of uncertainty in the decision process and obtain an "optimal" course of actions. Second, Bayesian modelling is characterised by extreme flexibility, which allows to account for the typical complexities of the data (e.g. correlation, skewness, spikes and missing data) in a relatively easy way. Third, the Bayesian approach naturally allows the incorporation of evidence from different sources (e.g. expert opinion or multiple studies) into the analysis, which may improve the estimation of the quantities of interest compared with us-

ing the evidence from a single source (e.g. a single trial). Finally, under the Bayesian approach, it is straightforward to assess and quantify the impact of uncertainty in all inputs of the decision process; this is extremely relevant in economic evaluations as it is a required component in the approval or reimbursement of a new intervention for many decision-making bodies, such as NICE in the UK (Claxton et al., 2005).

We briefly introduce the main aspects of Bayesian inference and focus on how the economic evaluation process is performed within a Bayesian framework. The concepts illustrated in this chapter constitute only a limited insight into the more general and complex Bayesian statistical theory. For a more comprehensive and exhaustive presentation of these topics we refer the reader to Gelman et al. (2013), Jackman (2009) and Lee (2012).

1.3.1 Bayesian Inference and Computation

In contrast with the frequentist approach, which assumes a unique, correct or "true" value of the probability attached to any uncertain event, the Bayesian approach interprets the probability as a "subjective" degree of belief, which depends on the individual whose uncertainty is being expressed and the information available to him/her. Under this perspective, each individual is entitled to his/her own subjective probability, which can be updated according to the evidence that becomes sequentially available.

From a modelling perspective, the Bayesian interpretation of probability allows to make probabilistic statements directly on the quantities of interest, i.e. some unobservable feature of the process under study, typically represented by a set of parameters. More specifically, a Bayesian model specifies a full probability distribution to describe uncertainty in terms of the data, which are subject to sampling variability, and unobserved quantities (e.g. parameters or future observations), which are not typically known to the experimenter. As a consequence, probability is used in the Bayesian framework to assess any form of limited knowledge. The experimenter needs to identify a suitable probability distribution to describe the overall uncertainty about the data $\boldsymbol{y}$ and the unknown parameters $\boldsymbol{\theta}$, which we indicate with $p(\boldsymbol{y}, \boldsymbol{\theta})$.

By the basic rules of probability we can re-express this joint distribution as the product of the marginal distribution of the data $p(\boldsymbol{y})$ and the conditional distribution of the parameters given the data $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ or vice versa, i.e. $p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{y})p(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})$. From this, we can derive the fundamental theorem of Bayesian inference known as Bayes theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{1.2}$$

The basic idea underlying the theorem is that we can update our level of uncertainty about the parameters before observing the data, expressed through a *prior distribution* $p(\boldsymbol{\theta})$, with the evidence from the data, expressed through the *likelihood* $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, into a *posterior distribution* $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. This allows us to make inference in terms of direct probabilistic statements.

When little information is contained in the prior $p(\boldsymbol{\theta})$, which is then typically referred to as a *diffuse* or *vague* prior, the resulting posterior will be mostly informed by the likelihood. Thus, inference will be numerically similar to that achieved in a frequentist setting, where only information from the data is considered. However, because of the different assumptions underlying the Bayesian and frequentist statistical frameworks, the interpretation of the results between the two approaches remains different.

Alternatively, we can use *informative* priors, i.e. distributions that represent some knowledge about the model parameters before observing the data and that, together with the likelihood, drives posterior inferences. The most serious issue with using informative priors is related to the way information is elicited, i.e. brought into the model. More generally, the *elicitation process*

implies that the people providing the information to be included in the model (e.g. clinical experts) possess some kind of knowledge or beliefs "in their heads" and it is the analyst task to devise the right kind of questions to "extract" this information from them (O'Hagan et al., 2006).

It is crucial to find an appropriate way to express the external information collected in order to adequately inform the priors on the parameters of interest in the setting analysed. There is a wide literature about the process of eliciting experts probabilities and alternative approaches are available (Stevens and O'Hagan, 2002; Grigore et al., 2016; Mason et al., 2017). Elicited probabilities may suffer from biases and non-coherence in practice, but the goal of the elicitation is to represent the expert knowledge and beliefs as accurately as possible. We will not further cover this subject as it falls outside the focus of this work and we refer to O'Hagan et al. (2006) for a more comprehensive examination on the topic.

A Bayesian analysis can assess the level of uncertainty for any unobserved quantity, be it a parameter or some future observation, given the information from the observed data and model assumptions. For example, it is possible to evaluate probabilistically unobserved data $\tilde{\boldsymbol{y}}$, assumed to be of the same nature as those already observed, by means of the (posterior) *predictive distribution*

$$p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}. \tag{1.3}$$

The expression indicates that, once the uncertainty regarding the unknown parameters has been integrated out, $p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$ does not depend on $\boldsymbol{\theta}$ anymore and indicates what we know about the distribution of the (future) $\boldsymbol{y}$. When only prior information is available, we replace $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ by $p(\boldsymbol{\theta})$ in Equation 1.3. This yields the *prior predictive distribution* $p(\tilde{\boldsymbol{y}}) = \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, which corresponds to the denominator of Equation 1.2.

While the posterior distribution might be known in closed form, it is often not analytically tractable. Iterative approximation methods are typically used to approximate the posterior distribution and produce the required inference. Among these, one of the most popular techniques is *Monte Carlo Integration* (Kroese et al., 2013), which uses a series of random values to numerically compute a definite integral. More details on how Monte Carlo integration can be used to derive posterior summaries is available in Appendix A.1.

When the posterior distribution is not available in closed form, a more general class of algorithms which can be used to derive posterior inferences is provided by *Markov Chain Monte Carlo* (MCMC) methods (Brooks et al., 2011). MCMC methods are a class of iterative algorithms for sampling from generic probability distributions, which are based on the construction of a Markov chain that converges to the desired target distribution, i.e. the joint posterior distribution of the parameters we are interested in. Provided that some regularity conditions are satisfied (Jackman, 2009; Brooks et al., 2011), after a sufficiently large number of iterations the chain will forget its initial state and will converge to the unique stationary target distribution, which does not depend on time or the initial position.

There are different types of MCMC algorithms. One of the most popular is *Gibbs sampling* (Geman and Geman, 1984), which samples sequentially from the full conditional distribution of each parameter or block of parameters. For a detailed presentation of different types of MCMC methods we refer to Brooks et al. (2011). Over time, many software that are specifically designed for the analysis of Bayesian analysis using MCMC techniques have been developed. Perhaps, the software that have most proliferated among applied statisticians are those based on the model description language known as `BUGS` (*Bayesian inference Using Gibbs Sampling*; Gilks et al., 1994; Lunn et al., 2012). Some examples are: `WinBUGS`, and its open source variant *OpenBUGS*; and `JAGS` (*Just Another Gibbs Sampler*; Plummer, 2010). Recently, another probabilistic programming language, called `STAN` (Carpenter et al., 2017), has been developed for conducting Bayesian inference using a specific class of MCMC methods known as *Hamiltonian Monte Carlo* (Brooks et al.,

2011).

Once the MCMC sampling has successfully been performed, for each parameter of interest we can typically access a vector of S simulations $(\theta^{(1)}, \ldots, \theta^{(S)})$ from the posterior distribution. This can then be summarised by computing a variety of quantities, such as the posterior mean or median, with uncertainty typically represented via "credible" intervals obtained from the percentiles of the posterior. For example, we can compute the posterior mean $\mathrm{E}[\theta \mid \boldsymbol{y}]$ by averaging across the simulated values $\theta^{(s)}$ or derive an approximate 95% *credible interval* (CI) using the empirical 2.5% and 97.5% quantiles of the posterior distribution of θ. A slightly different type of interval is the 95% *highest posterior density* (HPD) interval. This corresponds to the CI that contains the values of θ that are a posteriori more plausible, i.e. $p(\theta \mid \boldsymbol{y})$ is higher for all θs inside the interval than for values outside the interval.

Generally speaking, there is no measure that can definitely assess whether the MCMC has converged because the chain will reach the target distribution (forgetting its initial state) only asymptotically. However, some diagnostic measures such as the *potential scale reduction factor* and the *effective sample size* (Gelman et al., 2013) may provide some useful insights to detect failures in convergence or high autocorrelation in the MCMC sampler. We do not go further into details about MCMC methods as they are not the main focus of this work, and we refer to Brooks et al. (2011) for a comprehensive text about MCMC methods and inference.

1.3.2 Model Checking

A crucial part of any statistical analysis is assessing the fit of the model. In Bayesian analysis this includes checking for any sensitivity to the choice of the prior in addition to more "standard" checks regarding, for example, residuals and predictions (Gelman et al., 2013). In particular, when comparing models, it can be informative to evaluate their predictive accuracy and perform model selection based on their fit to the data.

A typical approach to assess the predictive accuracy of a Bayesian model uses *posterior predictive checks* (Gelman et al., 2013). The basic technique is to draw simulated values from the posterior predictive distribution (Equation 1.3), and compare these to the observed data, either graphically or using discrepancy measures. If the model fit is reasonable, then the predictions generated by the model should look similar to the observed data. Different types of graphical assessments can be used, including a display of all the data, data summaries or residuals (Gelman et al., 2013). A potential issue with using posterior predictive checks is that, since the data have influenced the estimation of the parameters, they are in fact used twice, i.e. for model fitting and checking. This is not optimal as, in principle, the fit of the models should be evaluated on some external data that were not used for estimation. However, posterior predictive checks can provide useful insights about potential failings of the model, provided that usage is limited to study model adequacy, not for model comparison and inference (Meng, 1994).

Alternative measures of model checking, known as *information criteria*, have been proposed. These measures evaluate the predictive ability of competing models in terms of the accuracy of the models' predictions based on the observed data and some bias correction terms. For a comprehensive review of different types of criteria we refer to Gelman et al. (2014) and Vehtari et al. (2017). For the analyses in this thesis, we specifically focus on the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), which is the predictive measure of choice in many Bayesian applications, in part because of its incorporation in the popular `BUGS` software. A detailed description of how DIC can be computed and the potential pitfalls in its implementation is available in Appendix A.2.

1.4 Decision Modelling in Economic Evaluation

Health economic evaluations are a typical problem of decision-making under uncertainty. Their main objective is to provide decision-makers with a useful tool that permits the comparison of competing technologies in terms of the benefits they provide and the resource use required to reach these benefits. This leads to the need of some methods to compare alternative options, reflect uncertainty in the conclusions and choose which option should be applied to the whole population, given the available evidence (e.g. coming from a RCT).

Figure 1.2 graphically represents the general process of doing a Bayesian analysis (with a view of using the results of the model to perform an economic evaluation). Four steps form the process:

Statistical Model. At the beginning of the process a statistical model is used to estimate some relevant parameters, such as the population mean effectiveness and costs. The type of statistical analysis varies with the nature of the underlying data (e.g. individual level versus aggregated level data).

Economic Model. The estimates from the statistical model are then combined in the economic model, with the objective of obtaining some relevant population parameters indicating the average benefits and costs for a given intervention in comparison to another one. Depending on the type of available data and statistical model used, this step may simply correspond to using the parameters as derived from the statistical model or it may involve complex combinations thereof (Baio, 2017).

Decision Analysis. Once appropriate outcome measures are obtained from the economic model, these are used as the basis for the decision analysis, which aims at identifying the optimal intervention that should be applied to the target population by computing suitable measures of “cost-effectiveness".

Uncertainty Analysis. The final aspect is represented by the evaluation of how the uncertainty (e.g. in parameters or model structure) impacts the final decision outcome and obtain the best course of action given current evidence. However, when the available evidence is limited, it is important to assess the impact of current uncertainty on decision-making.

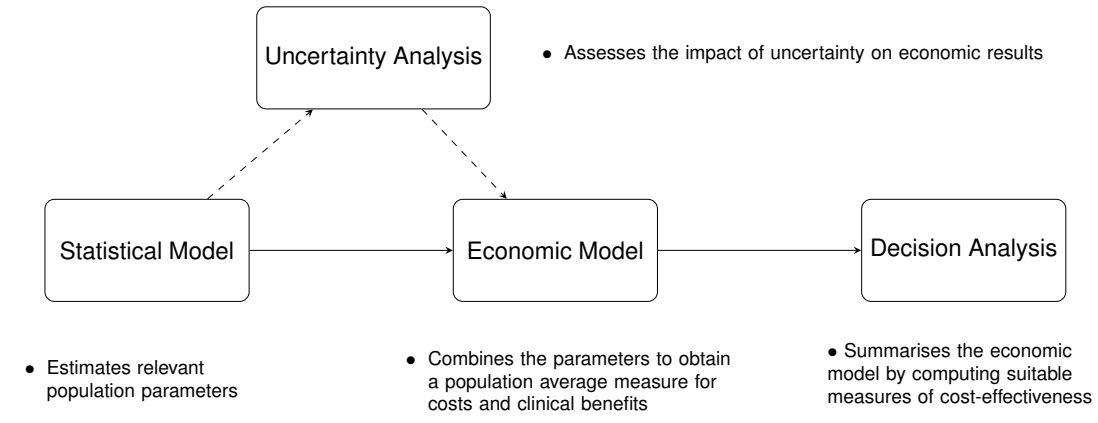


Figure 1.2: Economic evaluation process. The process is formed by four main steps, represented as rectangles, connected to each other by arrows (indicating the dependency relationships). The flow of the process is indicated by solid arrows, while dashed arrows link the steps subjected to uncertainty assessments. Source: Baio et al. (2017).

The first two steps (Statistical and Economic Model) are related to the construction of appropriate models to obtain inferences and their forms mainly depend on the available data and modeller

approach. Conversely, the last two steps (Decision and Uncertainty Analysis) can be represented in a more standardised way. Specifically, their purpose is to obtain summary CEA measures of interest for decision makers and assess the current level of uncertainty in the decision problem. We now focus on these steps and on the tools typically used to perform their tasks.

1.4.1 Comparing Health Interventions

As discussed earlier, the purpose of economic evaluations is to provide information to decision-makers about the costs and benefits for each treatment under evaluation. Decision analysis has been defined as a systematic approach to decision-making under uncertainty (Raiffa, 1968). In the context of economic evaluation, a decision analytic model uses mathematical relationships to define a series of possible consequences, each associated with specific costs and benefits, for the set of alternative treatments being evaluated. These consequences are then used to inform the best course of action and choose the "optimal" treatment given current evidence, where optimality is defined in decision-theoretic terms (O'Hagan and Stevens, 2001; Briggs et al., 2006; Baio, 2012).

A simple and popular criterion to select an "optimal" strategy in the decision problem of economic evaluation is the *maximisation of the expected utility*. The main idea is to choose, among the set of available interventions, the option maximising the probability of the outcome preferred by the decision-maker, i.e. in terms of benefits and costs. When certain conditions are satisfied (Raiffa, 1968; Smiith, 1988), the decision-making can be performed by simply computing an average. In particular, we need to compute for each intervention the expectation of a *utility function* with respect to both "population" (parameters) and "individual" (sampling) uncertainty/variability. For simplicity, we consider the example where there are only two treatments being compared and generically term the clinical benefits as "effectiveness". In the following, we indicate the economic outcome variables as (c_{it}, e_{it}), where $t = 1, 2$ denotes the treatment option (e.g. new intervention vs control), and drop the individual index i to ease notation. Although the definition and type of the variables (c_t and e_t) may vary depending on the specific conditions considered (e.g. quality of life or survival data), throughout the thesis we assume that they are always expressed in terms of health care costs and QALYs, which are derived from the cost and utility data collected at different time points in the trial using the formulae illustrated in Section 1.2.

As for the type of utility function, health economic evaluations are generally based on the *Net monetary Benefit* (Stinnett and Mullahy, 1998)

$$\text{Net Benefit} = ke_t - c_t, \tag{1.4}$$

where k is a willingness-to-pay parameter used to put the outcomes on the same scale. Note that the definition provided in Equation 1.4 is based on the standard representation of the net benefit which is typically used in the health economic literature (Claxton, 1999) and that alternative formulations exist (Willan et al., 2004).

The net benefit represents the budget that the decision-maker is willing to invest to increase the benefits by one unit. The main advantage of using the net benefit framework is its fixed form, once (e_t, c_t) are defined, which provides easy guidance to the evaluation of the interventions. Moreover, Equation 1.4 is linear in (e_t, c_t), which facilitates interpretation and calculation tasks. However, the use of the net benefit presupposes that the decision-maker is risk neutral, which is by no means always appropriate in health policy problems (Koerkamp et al., 2007). For simplicity, we do not focus on this aspect and we assume that the net benefit framework can adequately describe the utility function of the decision-maker, in line with the vast majority of the health economics literature (Briggs, 1999; O'Hagan and Stevens, 2001; Spiegelhalter et al., 2004; Grieve et al.,

2005; Gomes et al., 2012b; Diaz-Ordaz et al., 2014b; Ng et al., 2016).

Since the net benefit is assumed to be linear in both outcomes, the focus of the analysis can be moved to the estimation of the expected effectiveness and cost quantities, which can then be used in Equation 1.4 to easily derive the expected net benefit associated with each treatment t. In particular, a statistical model $p(e_t, c_t \mid \boldsymbol{\theta})$ is typically applied to the cost and effectiveness variables to derive some relevant parameters $\boldsymbol{\theta}$. The model is usually fitted to all treatments under the assumption that some of the parameters $\boldsymbol{\theta}$ are shared between treatment groups. In principle, however, it is possible to separately fit a model to each treatment to derive treatment specific estimates for all model parameters, i.e. $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The interest lies in the population mean parameters

$$\mu_{et} = \mathsf{E}\left[e_t \mid \boldsymbol{\theta}\right] \quad \text{and} \quad \mu_{ct} = \mathsf{E}\left[c_t \mid \boldsymbol{\theta}\right]. \tag{1.5}$$

The quantities μ_{et} and μ_{ct} are then used in assessing the relative cost-effectiveness of the interventions. More specifically, resource allocation decisions are typically based on suitable economic summaries that compare average differences in c_t and e_t between options. For example, we can re-express the parameters in Equation 1.5 in terms of the population average increments

$$\begin{aligned} \Delta_e &= \mathsf{E}[e \mid \theta_2] - \mathsf{E}[e \mid \theta_1] = \mu_{e2} - \mu_{e1} \\ \Delta_c &= \mathsf{E}[c \mid \theta_2] - \mathsf{E}[c \mid \theta_1] = \mu_{c2} - \mu_{c1}, \end{aligned} \tag{1.6}$$

where Δ_e and Δ_c are respectively the average increment in the effectiveness and costs from selecting $t = 2$ compared to $t = 1$. Once the measures in Equation 1.6 are obtained, using the net benefit as the utility function, it is possible to obtain the *Incremental Benefit* (IB) of treatment 2 over treatment 1

$$\mathsf{IB} = k\Delta_e - \Delta_c. \tag{1.7}$$

Notice that, under the Bayesian framework, the quantities (Δ_e, Δ_c) in Equation 1.7 are random variables because, while sampling variability is averaged out, these are defined as functions of the parameters of the model $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The second layer of uncertainty (i.e. the population, parameters domain) can be further averaged out. Thus, according to the net benefit framework, decision-making can be effected by considering the so-called *Expected Incremental Benefit* (EIB)

$$\mathsf{EIB} = \mathsf{E}[k\Delta_e - \Delta_c] = k\mathsf{E}[\Delta_e] - \mathsf{E}[\Delta_c]. \tag{1.8}$$

where, the increment in mean effectiveness and costs $(\mathsf{E}[\Delta_e], \mathsf{E}[\Delta_c])$ are actually pure numbers

$$\begin{aligned} \mathsf{E}[\Delta_e] &= \overline{e}_2 - \overline{e}_1 = \mathsf{E}[\mu_{e2}] - \mathsf{E}[\mu_{e1}] \\ \mathsf{E}[\Delta_c] &= \overline{c}_2 - \overline{c}_1 = \mathsf{E}[\mu_{c2}] - \mathsf{E}[\mu_{c1}]. \end{aligned} \tag{1.9}$$

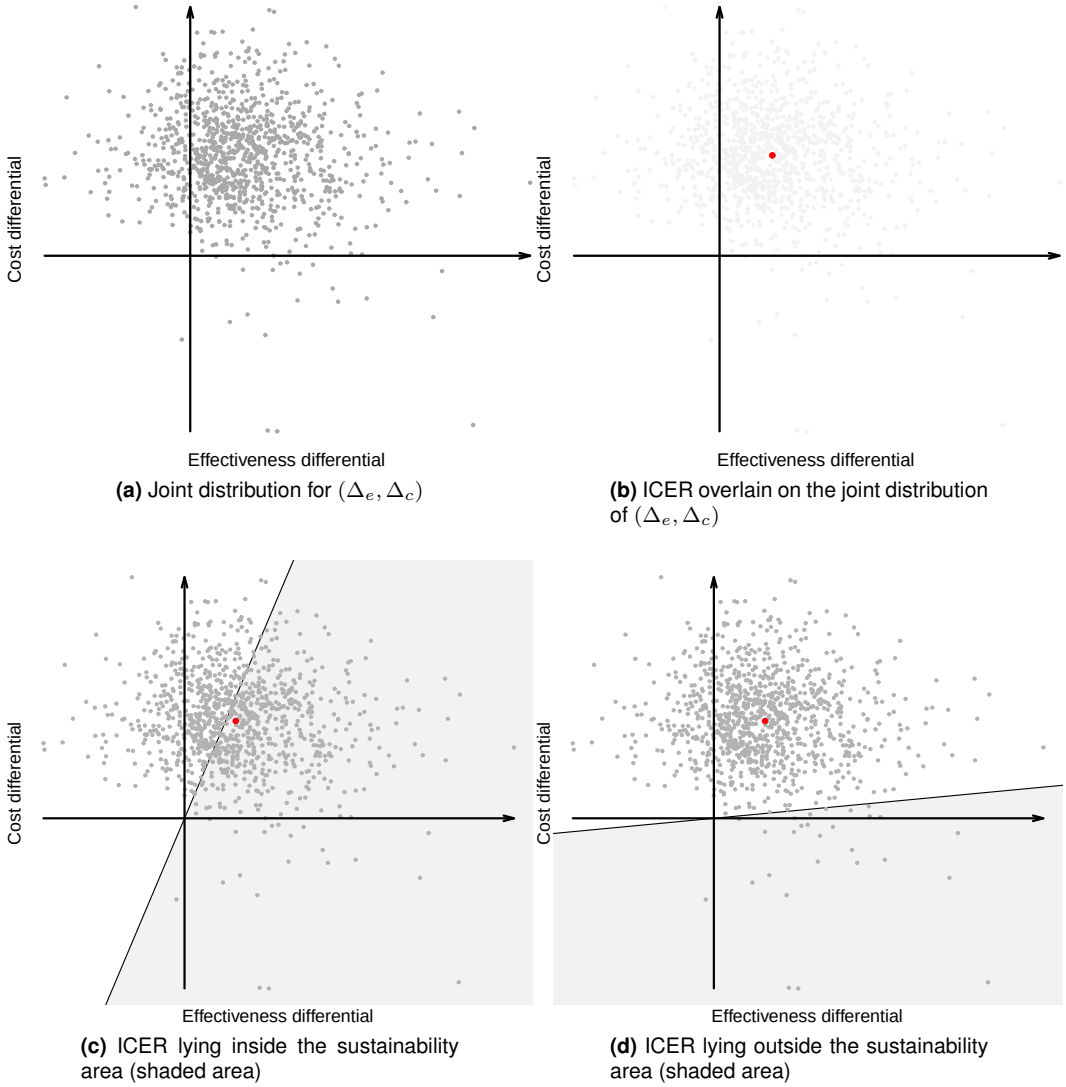
In particular, $\overline{c}_2$ and $\overline{e}_2$ are the population averages of cost and benefit measures for the reference treatment, to be assessed against those of the comparator $(\overline{c}_1, \overline{e}_1)$. The quantities in Equation 1.8 and Equation 1.9, in turn, can be used to compute the *Incremental Cost-Effectiveness Ratio* (ICER), defined as

$$\mathsf{ICER} = \frac{\mathsf{E}\left[\Delta_c\right]}{\mathsf{E}\left[\Delta_e\right]}. \tag{1.10}$$

The ICER represents the cost per incremental unit of effectiveness (e.g. cost per QALY gained) and provides a ratio summary of the additional cost and benefit that result from one option compared to another. From Equation 1.8 and Equation 1.10, we see that when $\mathsf{EIB} > 0$ and so $k >$ ICER, then $t = 2$ is the optimal treatment (associated with the highest expected utility). Thus, decision-making can be equivalently effected by comparing the ICER to the willingness-to-pay threshold.

1.4.2 Cost-Effectiveness Assessment

The two layers of uncertainty underlying the decision-making process, as well as the relationships between benefits and costs, can be best appreciated through the inspection of the *Cost-Effectiveness Plane* (CEP; see Black, 1990; Briggs and Gray, 1999; Baio, 2012), shown in Figure 1.3 which is taken from Baio (2012).



(a) Joint distribution for (Δ_e, Δ_c)

(b) ICER overlain on the joint distribution of (Δ_e, Δ_c)

(c) ICER lying inside the sustainability area (shaded area)

(d) ICER lying outside the sustainability area (shaded area)

Figure 1.3: 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 sustainability areas under two alternative choices for k. Source Baio (2012).

The CEP depicts the joint distribution of the random variables (Δ_e, Δ_c) on the x and y axis, respectively, and is divided into four quadrants. Depending on where the point $(\mathrm{E}(\Delta_e), \mathrm{E}(\Delta_c))$, which is related to the ICER through Equation 1.10 (i.e. the ICER corresponds to the slope of the line joining the origin with the point), the CEP can provide a general idea about the cost-effectiveness of the new intervention with respect to the comparator. For example, when the point associated with the ICER falls in the north-east quadrant, the intervention generates more health gains but is also more expensive. Other quadrants are relevant when the intervention generates poorer health outcomes (north-west or south-west quadrants) or lower costs (south-west or south-east quadrants). CEPs are also useful to show the uncertainty around model parameters and cost-effectiveness outcomes, often represented as a cloud of points on the plane corresponding to different iterations of an economic model, Figure 1.3 (a). Specifically, in the graphs, the samples

are obtained from the joint posterior distribution of the expected mean effectiveness and costs using a full Bayesian approach (through MCMC methods, see Section 1.3.1). Taking the expectations over the marginal distributions for Δ_e and Δ_c, we then marginalise out the uncertainty and obtain the point $(\mathsf{E}(\Delta_e), \mathsf{E}(\Delta_c))$ in the plane which is closely related to the ICER, Figure 1.3 (b).

Figure 1.3 (c) shows the "sustainability area" (Baio, 2012), i.e. the part of the CEP that lies below the line $\mathsf{E}[\Delta_c] = k\mathsf{E}[\Delta_e]$, for a given value of the willingness-to-pay k. Interventions for which the point associated with the ICER is in the sustainability area are more cost-effective than the comparator. Changing the value for the threshold can modify the decision as to whether $t = 2$ is the most cost-effective intervention. The target values for k, indicated by NICE for economic evaluations, usually lie between $£20,000 - 40,000$ per QALY gain (NICE, 2008). Recently, Claxton et al. (2015) have also suggested to reduce this generic threshold benchmark value. However, no agreement has been reached yet as to whether this new threshold should be used and routine analyses still use $£20,000 - 40,000$ as the reference range.

Figure 1.3 (d) shows the sustainability area for a different choice of the parameter k. In this case, because the ICER and, for that matter, most of the entire distribution of (Δ_e, Δ_c) lie outside the sustainability area, the new intervention $t = 2$ is not cost-effective.

Theoretically, the identification of the expected utility with respect to both individual variations and uncertainty in the value of the parameters is all that is needed to reach the best decision given current evidence. However, when implementing an intervention, decision-makers are typically faced with the problem of whether to make the decision based on the current evidence or to defer the decision in order first to gather more evidence. Thus, the decision-process is also subject to the uncertainty related to the trade-off between the collection of new data, with the aim at resolving at least partially the current uncertainty about the model parameters, and the sampling costs associated with the collection of the additional evidence. For these reasons, economic evaluations are typically subjected to some form of uncertainty analysis to quantify and qualify the uncertainty underlying the decision problem.

A particular class of analysis that is suited to assess this type of uncertainty is *probabilistic sensitivity analysis* (Claxton et al., 2005; Baio, 2012; Baio and Dawid, 2015), which considers all input parameters as random quantities and assigns them a probability distribution that describes the associated state of knowledge. Very briefly, the idea behind probabilistic sensitivity analysis is to compare the actual decision process to the ideal one, characterised by the (currently unknown) quantities. This is done with a view to assessing whether the information provided by the current evidence is sufficient to take a decision on the optimal treatment or it would be more effective to defer the decision until after additional evidence is collected.

A common tool used to summarise the results of probabilistic sensitivity analysis is the *Cost-Effectiveness Acceptability Curve* (CEAC; see Van Hout et al., 1994). Using the net benefit as the utility function, the CEAC can be expressed as

$$\text{CEAC} = \Pr(k\Delta_e - \Delta_c > 0) \tag{1.11}$$

The CEAC depends on the willingness-to-pay parameter k. When Net benefit > 0, i.e. the optimal decision is treatment $t = 2$, the CEAC is the probability that learning the value of $\boldsymbol{\theta}$ (i.e. resolving the uncertainty on the parameters) would not change that decision. The main advantage of Equation 1.11 is to allow a simple summary of the probability of cost-effectiveness upon varying the willingness-to-pay parameter k. In particular, within a Bayesian approach, the CEAC can be straightforwardly interpreted as the probability that the intervention $t = 2$ is cheaper, compared to $t = 1$, as k is reduced, and that it is more effective as k increases.

Some limitations about the CEACs have been pointed out since they do not allow explicit for any possible change in the payoffs. In particular, CEACs are only concerned with currently avail-

able information, but do not consider explicitly the possibility of gathering additional evidence, therefore providing only a partial evaluation of the overall decision process (Koerkamp et al., 2007). Despite their limitations, however, CEACs represent the standard tools used by practitioners in trial-based analyses to report the economic results, particulalry for small and pilot trials, where the objective is to provide a simple summarisation of the probability of cost-effectiveness of the new interventions to inform the decision about the feasibility of conducting larger trials to collect more evidence. Since the economic analyses in this thesis will only focus on realtively small trials (Section 3.1), we will follow the current practice in trial-based analyses which uses the CEACs as the main tool to summarise the uncertainty about the cost-effectiveness of a given intervention (NICE, 2013; Ramsey et al., 2015).

Finally, an important and almost inevitable source of uncertainty in trial-based economic evaluations is represented by missingness, as it induces a lower level of information in the data. Missing data may seriously complicate the analysis due to a potential bias introduction that can affect all treatment regimens and outcome measures. Handling missingness is a particularly well-suited task for Bayesian analysis where any unknown quantity (e.g. missing data and parameters) is attached to a probability distribution that can be used to incorporate in the model some external information about that quantity which is not contained in the data.

It is therefore essential, within the decision model, to investigate and describe how uncertainty in all model inputs or parameters (including the missing values) translates in terms of uncertainty over the model outputs. Indeed, variations in posterior inferences will be an indication of the impact that this uncertainty has on the inferences and on the decision-making process.

1.5 Missing Data Analysis

There is a large literature on handling missing data. For the purpose of this thesis, we only introduce some key concepts for missingness analysis and present some of the most popular methods to deal with missing outcome data in trial-based economic evaluations. For a more comprehensive discussion on the topic we refer to the handbooks Rubin (1987); Schafer (1997); Molenberghs and Kenward (2007); Daniels and Hogan (2008); Little et al. (2010); Carpenter and Kenward (2013); Molenberghs et al. (2015).

The CEA setting is particularly challenging and worthwhile in terms of missing data handling. First, due to its intrinsic characteristics, CEA deals with a bivariate outcome whose relationships should be accounted for in the analysis. Thus, specifying and modelling the impact missing data may have on both outcomes, their dependence structure, and how these are translated in terms of changes in cost-effectiveness assessments is a major issue. Second, unlike standard statistical analyses, CEA does not merely focus on estimation purposes but needs to quantify the impact of missing data uncertainty both on the inferences and decision-making.

When *nonrespondents* are systematically different from respondents in terms of characteristics or outcomes, biased estimates may result from the (often implicit) formulation of unrealistic missingness assumptions (Molenberghs and Kenward, 2007). This is particularly likely when the reasons for missingness are ignored or not properly recorded. As a result, the statistical and economic analysis will be impaired because assumptions about unobserved data will be forced by the lack of available information about the missing values. Inferences may be strongly affected by the method chosen to deal with missingness, with the effect that decisions about the cost-effectiveness of a new treatment may be totally or partially misrepresented (Manca and Palmer, 2005; Marshall et al., 2009).

1.5.1 Full Data Models

We denote with $\boldsymbol{y}_{ij} = (u_{ij}, c_{ij})$ the typical bivariate outcome of trial-based CEAs, formed by the utility and cost pair collected at time j for subject i. Although, in principle, different groups of subjects enrolled in a trial could be followed-up at different times, i.e. $\boldsymbol{y}_{ij_{(i)}}$, however, this is not typically the case for trial-based economic evaluations, where a consistent follow-up is usually scheduled to ensure that cost and utility data are collected at the same set of occasions for all subjects in the analysis (Ramsey et al., 2015). Thus, in the following, we will assume that both economic outcomes are collected at the same time points $j = (1, \dots, J)$ for all subjects $i = (1, \dots, n)$.

When missingnes occurs at one or more occasions in the study, data are necessarily unbalanced over time since there are at least some individuals who have a different number of repeated measurements at a common set of occasions. This poses a serious problem for the analysis because nonresponses may occur at any follow-up time and lead to distinct missingness patterns.

Dropout or *attrition* refers to the specific situation where subjects are observed uninterruptedly from the beginning of the study until a given point in time, after which they are never observed again. More generally, a missing data pattern is said to be *monotone* if missingness is all due to dropout, i.e. given that $\boldsymbol{y}_{ij}$ is unobserved, then $\boldsymbol{y}_{ij+1}$ is also unobserved. When this condition does not hold, the missing data pattern is said to be *nonmonotone* or *intermittent*.

We denote with $\boldsymbol{r}_{ij} = (r^u_{ij}, r^c_{ij})$ a pair of indicator variables that take value 1 if the corresponding outcome for subject i at time j is observed and 0 otherwise. Next, we denote with $\boldsymbol{r}_i = (\boldsymbol{r}_{i1}, \dots, \boldsymbol{r}_{iJ})$ the missingness pattern to which subject i belongs, where each pattern is associated with different values for $\boldsymbol{r}_{ij}$. For example, the pattern $\boldsymbol{r} = ((1,1), \dots, (1,1)) = \mathbf{1}$ is associated with the set of ones at each occasion and therefore corresponds to the completers pattern.

Given the set of the responses $\boldsymbol{y}$ and missingness indicators $\boldsymbol{r}$, the *full data model* describes the joint distribution $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$, indexed by a finite-dimensional parameter $\boldsymbol{\omega}$. In almost all practical settings, interest typically lies in drawing inference about some parameters $\boldsymbol{\theta}$ that are a subset of $\boldsymbol{\omega}$ and that index the *full data response model* $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. The relationship between the two full data models is

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{r} \in \mathscr{R}} p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}), \tag{1.12}$$

where $\mathscr{R}$ denotes the sample space of $\boldsymbol{r}$. Equation 1.12 shows that inference about $\boldsymbol{\theta}$ will depend crucially on choices made in specifying the full data model. These are distinguished between the choices related to the model for the data and the choices about the missingness assumptions. However, while the observed data can be used to validate the modelling choices about the observed responses, they offer no information to validate choices about the missing data assumptions. Thus, either explicit or implicit assumptions about the missing values are needed to identify the full data model and will generally influence the final inferences about $\boldsymbol{\theta}$.

Before briefly reviewing some of the most commonly used techniques to handle missing data in CEA, we introduce the key concept of the *missing data mechanism*, which provides a convenient setting for the understanding and addressing of missingness.

1.5.2 Missing data mechanism

When analysing partially observed data it is essential to investigate the possible reasons behind the missingness, which translates into an assumed missing data mechanism that is linked to the data generating process. An analysis that is able to express and explicitly model this relationship can address missingness in a “principled way”. We specifically refer to principled methods for

missing data as those based on a well-defined statistical model for the complete data, and explicit assumptions about the missing value mechanism.

More formally, we use the term *missing data mechanism* to refer to the conditional distribution of the missing data indicators given the full data response $p(\boldsymbol{r} \mid \boldsymbol{y}, \psi)$, where the parameter ψ is a subset of $\boldsymbol{\omega}$. Any full data model can be factored as the product of a full data response model and the associated missing data mechanism

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{r} \mid \boldsymbol{y}, \psi). \tag{1.13}$$

The characterisation of the missing data mechanism through the factorisation in Equation 1.13, which is commonly referred to as a *selection model factorisation*, is only one possible representation. The implied missing data mechanism can, in principle, be derived from any specification of the full data model, but it may not always take a closed form.

The most accepted classification of missing data mechanisms is given by Rubin (1987) and is based on three classes, distinguished according to how the probability of missingness in the missing data mechanism is modelled.

Definition 3.1 *Missing Completely At Random (MCAR).*

Missing responses are missing completely at random (MCAR) if, for all ψ

$$p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \psi) = p(\boldsymbol{r} \mid \psi), \tag{1.14}$$

where $\boldsymbol{y}_{obs}$ and $\boldsymbol{y}_{mis}$ indicate the subsets of the response data that are observed and missing, respectively. Under MCAR there is no systematic difference between partially and fully observed individuals in terms of the outcome $\boldsymbol{y}$. In other words, Equation 1.14 assumes that the observed cases are a representative sample of the full sample. MCAR is a strong assumption that ensures the validity of any fully-observed data method to any case of partially-observed data. An example of MCAR is when a sample of patient questionnaires is lost and therefore no data is recorded.

Definition 3.2 *Missing At Random (MAR).*

Missing responses are missing at random (MAR) if, for all $\boldsymbol{y}_{obs}$ and ψ

$$p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \psi) = p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \psi). \tag{1.15}$$

Under MAR, the partially observed cases are systematically different from the fully observed cases; crucially, however, the difference is fully captured by $\boldsymbol{y}_{obs}$. Equation 1.15 is less restrictive than Equation 1.14 and is therefore considered a more plausible assumption in many situations. For this reason, MAR is normally considered as the reference assumption for the analysis of partially observed data. An example of MAR is when an individual is removed from a study as soon as the value of a specific observed variable falls outside a certain range. Missingness is therefore under the control of the investigator and is related to some observed components.

The MAR assumption is one component of the *ignorability* condition, which allows to draw valid inference about $\boldsymbol{\theta}$ in the full data response model based on the likelihood for $\boldsymbol{y}_{obs}$. More specifically, missingness is said to be *ignorable* if the following three conditions hold (Little and Rubin, 2002): (1) the missing data mechanism is MAR; (2) the parameter $\boldsymbol{\omega}$ of the full data model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ can be decomposed as $(\boldsymbol{\theta}, \psi)$, with $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and $p(\boldsymbol{r} \mid \boldsymbol{y}, \psi)$; (3) the parameters of the full data response model and missing data mechanism are a priori independent, that is $p(\boldsymbol{\omega}) = p(\boldsymbol{\theta}) p(\psi)$ (only required for a Bayesian analysis). When any of the conditions for ignorability is not satisfied, missingness is said to be *nonignorable* or *informative*. Often, this is due to the failure of the first condition, known as Missing Not At Random (MNAR).

Definition 3.3 *Missing Not At Random (MNAR).*

Missing responses are missing not at random (MNAR) if, for some $\boldsymbol{y}_{mis} \neq \boldsymbol{y}'_{mis}$ and $\boldsymbol{\psi}$

$$p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{\psi}) \neq p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}'_{mis}, \boldsymbol{\psi}), \tag{1.16}$$

that is, if $\boldsymbol{r}$ depends on some part of the unobserved data, say $\boldsymbol{y}'_{mis}$, even after conditioning on $\boldsymbol{y}_{obs}$. Since we cannot definitely distinguish between MAR and MNAR from the data at hand, the full data model $p(\boldsymbol{y}, \boldsymbol{r})$ always requires untestable assumptions about the missing data mechanism in order to be identified. Thus, model assumptions about $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\psi})$ become crucial in that any assumption on the missingness process cannot be tested from the observed data. The impossibility of testing Equation 1.16 based on the observed data leads to an extremely important role played by sensitivity analysis in assessing the robustness of the results across a range of plausible assumptions about the missingness mechanism. Some examples of these types of models, typically referred to as "nonignorable models", will be discussed in Section 1.6.

Since the scope of this section is to provide a broad overview for Rubin's classification, we assumed the simplest case where there are no covariate variables $\boldsymbol{x}$. If these covariates are fully observed, they can be incorporated in either or both the response and missingness model in a relatively easy way. However, when there are missing values in $\boldsymbol{x}$, the task of modelling missingness becomes more complex as there are potentially as many missingness mechanisms as the number of partially observed variables considered.

1.5.3 Missing data methods

This section briefly summarises some of the most commonly used methods to handle missing data in CEA. A review of the methods used in trial-based economic evaluations is presented and discussed in Chapter 2. Here, we focus our attention on three popular approaches: Complete Case Analysis, Single Imputation and Multiple Imputation. These are quite well-known methods used in the CEA literature to deal with missingness.

Complete Case Analysis

One of the popular methods for handling missing data in trial-based CEAs is *Complete Case Analysis* (CCA), as shown in recent reviews (Eekhout et al., 2012; Noble et al., 2012; Diaz-Ordaz et al., 2014a; Hughes et al., 2016; Gabrio et al., 2017; Leurent et al., 2018a). This approach includes in the analysis only individuals in the completers pattern (i.e. $\boldsymbol{r} = \mathbf{1}$) and is straightforward to implement. However, estimates under CCA are inefficient because some of the observations are discarded and the predictive information contained in the partially observed cases is completely ignored. In addition, non-negligible rates of missingness on a few variables of interest may cause large portions of the sample to be discarded and induce a greater loss of efficiency. When missingness occurs only in $\boldsymbol{y}$, the condition of validity of CCA corresponds to the MAR condition. However, the condition for validity of CCA does not fit neatly into Rubin's classes (White and Carlin, 2010) in the important cases when: some covariates in the analysis are partially-observed; or the outcome is longitudinal. For example, when the focus of the analysis is in some aspect of the conditional distribution of the response $\boldsymbol{y}$ given some covariate $\boldsymbol{x}$ (e.g. regression coefficients), then CCA will lead unbiased estimates as long as the distribution of $\boldsymbol{y} \mid \boldsymbol{x}$ in the completers is the same as in the target population. This condition does not match exactly the definition of the Rubin's classes since it holds when the missingness mechanism is MCAR, but also under certain MAR and MNAR mechanisms (White and Carlin, 2010). In general, the condition for the validity of CCA is that missingness is (conditionally) independent of response $\boldsymbol{y}$. Appendix A.3 provides a

simple example to show the validity of the estimates derived from CCA under different scenarios when the interest of the analysis is in some aspect of the conditional distribution of the response given some partially-observed covariates.

Single Imputation

Single Imputation (SI) methods replace the missing values with a single imputed value. The most attractive feature of these methods is that, once the imputation is performed, standard methods for complete data can be used. However, a serious drawback is that they always fail to account for the uncertainty associated with the missing data. Two of the most popular SI methods are marginal or conditional *mean imputation* (Buck, 1960) and *Last Value Carried Forward* (LVCF; Kenward and Molenberghs, 2010; Shao and Zhong, 2010).

Likelihood-Based Methods

Under MAR, the marginal distribution of the observed data $\boldsymbol{y}_{obs}$, also called the observed data likelihood and denoted with $L(\boldsymbol{\theta}, \boldsymbol{y}_{obs})$, provides the correct likelihood for the parameters $\boldsymbol{\theta}$ when the model fitted to the complete data $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is realistic, where $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$. The logarithm of this function, called the observed-data log-likelihood

$$\log L(\boldsymbol{\theta}, \boldsymbol{y}_{obs}) \tag{1.17}$$

has a crucial role in estimation. In particular, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ which maximises $\log L(\boldsymbol{\theta}, \boldsymbol{y})$ tends to be approximately unbiased and highly efficient in large samples (Cox and Hinkley, 1974). Confidence intervals and regions are often computed by appealing to the fact that, in regular problems with large samples, $\hat{\boldsymbol{\theta}}$ is approximately normally distributed about the true parameter $\boldsymbol{\theta}$ with approximate covariance matrix given by

$$V(\hat{\boldsymbol{\theta}}) \approx [-\log L^{'}(\hat{\boldsymbol{\theta}})]^{-1}, \tag{1.18}$$

where $\log L^{'}(\hat{\boldsymbol{\theta}})$ is the matrix of second partial derivatives of Equation 1.17 with respect to the elements of $\boldsymbol{\theta}$. The matrix $-\log L^{'}(\hat{\boldsymbol{\theta}})$ in Equation 1.18, which is often called the "observed information", describes how quickly the log likelihood function drops as we move away from the maximum likelihood estimate.

When the expressions for the maximum likelihood estimates cannot be written down in closed, iterative computation methods are typically required for estimation. A general method for maximum likelihood in missing data problems is the so called expectation-maximisation algorithm (Dempster et al., 1977), which proceeds by "filling in the missing data" with a best guess at what it might be under the current estimate of the unknown parameters, then re-estimate the parameters from the observed and filled-in data. The process is reiterated until the changes in parameter estimates between iterations are less than a pre-specified threshold at which point the algorithm is stopped. In theory, likelihood methods are more attractive than CCA or SI methods as they provide estimates that are generally valid under MAR (Little and Rubin, 2002; Schafer and Graham, 2002). However, they still rest on a few crucial assumptions. First, they assume that the sample is large enough for the ML estimates to be approximately unbiased and normally distributed. Second, the likelihood function comes from an assumed parametric model for the complete data (both observed and missing).

Multiple Imputation

Multiple Imputation (MI) overcomes the problems of single imputation by replacing missing data with a set of H plausible values, thereby taking into account the uncertainty about the imputations (Rubin, 1987, 1988, 1996). A common way to represent how MI works is based on three steps. The first is the *Imputation* step, where H completed datasets (observed and imputed data) are created to express the uncertainty about the imputed values. Then, in the *Analysis* step, the H completed datasets are analysed with some statistical method to produce H different sets of parameter estimates. Finally, in the *Combination* step, the results from the H different analyses are combined into a single estimate.

The main idea behind MI is that each imputed data set is not considered as the true completion of the actual data, as in SI, but instead they are used jointly to produce valid inferences. Rubin's rules (Rubin, 1987) are typically used to obtain an estimate of the mean (θ_{MI}) and variance (V_{MI}) of the parameters of interest across the H imputed datasets. These are typically computed as

$$\theta_{MI} = \frac{1}{H}\sum_{h=1}^{H}\theta_h \quad \text{and} \quad V_{MI} = \bar{V} + \left(1 + \frac{1}{H}\right)B, \tag{1.19}$$

where the mean estimate θ_{MI} is obtained as a simple average of the H separate parameter estimates θ_h, while the variance estimate V_{MI} is derived as a specific combination of two components. The first component, $\bar{V} = \sum_{h=1}^{K}\frac{V_h}{H}$, is the average of the within-imputation variance estimates for all the H imputed data sets. $\bar{V}$ does not account for the implied correlation in the imputed data. The second component, $B = \sum_{h=1}^{H}\frac{(\theta_h - \theta_{MI})^2}{(H-1)}$, is the overall between-imputation variance estimate, reflecting the uncertainty due to missing data captured by the imputation process. The more we expect the imputed data sets to differ from each other, the greater the between-imputation variance estimate should be.

Originally, a small number of imputations (roughly 10) was recommended to achieve sufficiently realistic estimates for the quantities in Equation 1.19 (Rubin, 1987), but nowadays this number can be greatly increased at a limited computational cost. The increased availability of functions that allow the implementation of MI in standard software packages (e.g. `STATA` or `R`; see Van Buuren and Groothuis-Oudshoorn, 2011), with limited customisation from the user, in combination with the lack of expertise and understanding about MI may have lead to some unsupervised used of these methods, as noted by Molenberghs et al. (2015). Recently, guidelines and recommendations have started to appear in the CEA literature to move practitioners towards a more critical approach to the implementation of MI and other imputation methods (Faria et al., 2014; Diaz-Ordaz et al., 2014a; Leurent et al., 2018a), but these require further development in order to be incorporated into the current practice.

On the one hand, MI generally allows the inclusion of a larger number of variables/predictors in the imputation model than used in the analysis model, which potentially makes the assumption of MAR more plausible and thus the overall analysis less likely to be biased. On the other, the performance of MI depends on the correct specification of the imputation model (i.e. complexity in the analysis model is reflected in the imputation model) and care is required in its construction. When there is a mismatch between the imputation model and the analysis model, often referred to as "non-congeniality", the final results may be biased (Van Buuren and Groothuis-Oudshoorn, 2011; Carpenter and Kenward, 2013). This situation occurs when modelling choices made in performing imputation are not compatible with those that will be made in the ultimate analysis of interest. Examples include the use of inappropriate distributions for the imputation of the data and omission of covariates or random effect terms in the imputation models which, instead, are included in the analysis model. In all these cases the relationships between the variables in the analysis are misrepresent and may lead to biases in the estimates of association, e.g. Rubin's

variance rules in Equation 1.19 (Molenberghs et al., 2015)

MI is commonly available in many statistical software packages using different approaches. Perhaps, the most popular version of MI is the one known as *chained equations*, which imputes missingness using appropriate univariate regression specifications for each variable included in the imputation model. Multiple imputation using chained equations can simultaneously handle issues about an appropriate scaling and modelling of the data distributions together with the incorporation of non-linear terms in the univariate models. An alternative version of MI is known as *joint modelling* and is based on the construction of a joint model for the complete data where missing values are iteratively sampled from proper conditional distributions. In CEA, an example of multiple imputation using joint modelling has been recently proposed to handle multilevel structures in the data through a flexible parametric approach (Gomes et al., 2013; Diaz-Ordaz et al., 2016).

1.6 Nonignorable Models

In the presence of missingness, identification of a full data response model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ requires making unverifiable assumptions about the full data model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$, with $\boldsymbol{\theta}$ a subset of $\boldsymbol{\omega}$. When the assumptions for ignorability are believed to be unrealistic, a more general class of models that allow missing data indicators to depend on the missing responses themselves can be used. These models are called *nonignorable models* and allow to parameterise the conditional dependence between $\boldsymbol{r}$ and $\boldsymbol{y}_{mis}$ given $\boldsymbol{y}_{obs}$. This association structure can only be identified through some untestable assumptions about $\boldsymbol{y}_{mis}$.

The Bayesian approach is well-suited to handle missingness, particularly under a nonignorable modelling framework. More specifically, assumptions about the missing values can be easily incorporated in a Bayesian model using informative prior distributions, which can be elicited using sources external to the data. Inferences are derived based on the information from the available data and the missingness assumptions specified, with uncertainty that is fully propagated throughout the model. Then, the sensitivity of the results to alternative assumptions (i.e. using different priors) and their impact on the final results can be easily assessed in terms of variations in the posterior quantities of interest.

There are different types of nonignorable models, typically distinguished based on the factorisation used to identify the full data model. Three popular techniques are the *Selection Models* (Heckman, 1979; Diggle and Kenward, 1994; Mason et al., 2012b), *Shared Parameters Models* (Wu and Carroll, 1994; Hogan and Laird, 1997) and *Pattern Mixture Models* (Little, 1993, 1994; Daniels and Hogan, 2008). The general intuition and a summary of the potential advantages and disadvantages of each of these nonignorable approaches are now briefly presented.

Selection Models

The selection model approach factors the full-data distribution as

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\psi}), \tag{1.20}$$

so that the full-data response model $p(\boldsymbol{y}, \boldsymbol{\theta})$ and the missing data mechanism $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\psi})$ must be specified by the analyst. Equation 1.20 can be attractive because it directly specifies the full-data response distribution and because its factorisation appeals to the missing data taxonomy described in Section 1.5.2, which allows an easy characterization of the missing data mechanism. However, an important downside of selection models is that identification of the missing

data distribution is accomplished primarily through parametric assumptions about the full-data response model $p(\boldsymbol{y}, \boldsymbol{\theta})$ and the explicit form of the missing data mechanism (e.g. linear in $\boldsymbol{y}$). This could lead to the situation where selection models allow the parameters indexing the association between $\boldsymbol{r}$ and the partially-observed $\boldsymbol{y}$ (i.e. ψ) to be identified from the observed data. Thus, this feature of the models places considerable importance on assumptions that cannot be verified, and has the potential to make sensitivity analysis problematic.

Shared Parameter Models

Shared parameter models specify the full data distribution using an explicitly multilevel formulation, where some random effects $\boldsymbol{b}$ are modelled jointly with $\boldsymbol{y}$ and $\boldsymbol{r}$. The general form of the full-data model using a shared parameter approach is

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = \int p(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{b}, \boldsymbol{\omega}) d\boldsymbol{b}, \tag{1.21}$$

Alternative specifications are formulated by making different assumptions about the joint distribution under the integral sign. Notice that in Equation 1.21 the full-data parameter $\boldsymbol{\omega}$ also includes parameters indexing the distribution of the random effects. The main advantage of shared parameter models is their simplified specification for the response and missingness components and the possibility of handling multilevel response data in a relatively easy way. However, a critical disadvantage is that the underlying missing data mechanism can be difficult to understand and may not even have a closed form (i.e. it requires the integration over $\boldsymbol{b}$).

Pattern Mixture Models

Pattern mixture models factor the full data model as

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta}) p(\boldsymbol{r} \mid \boldsymbol{\lambda}), \tag{1.22}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ respectively index the two factors in Equation 1.22. To notice that, the decomposition of $\boldsymbol{\omega} = (\boldsymbol{\theta}, \psi)$ used in Section 1.5.2 refers to a selection model factorisation of the full-data model (Equation 1.20), while the partition $\boldsymbol{\omega} = (\boldsymbol{\eta}, \boldsymbol{\lambda})$ is based on the pattern mixture model factorisation and the two are therefore not equivalent.

The full data response model can be derived as a mixture

$$p(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{r} \in \mathscr{R}} p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta}) p(\boldsymbol{r} \mid \boldsymbol{\lambda}). \tag{1.23}$$

Pattern mixture models essentially specify a different distribution for each missing data pattern and retrieve the full data response model by marginalising over $\boldsymbol{r}$. The main advantage of these models is to make explicit the parameters that cannot be identified by the observed data. There are two potential downsides of pattern mixture models. First, the full data response model is not directly available and must be retrieved through Equation 1.23. Second, the implementation of these models may become cumbersome: as the dimension of $\boldsymbol{y}$ increases, so will the dimension of the unidentified parameters.

Although all three types of nonignorable models can be used to explicitly handle missingness under MNAR, the choice of which approach to use is typically guided by the specific objective of the analysis and context analysed. In particular, the focus of this thesis is on the factorisation of pattern mixture models and how it can be used to assess the sensitivity of the inferences to alternative nonignorable assumptions. Specifying a different distribution for each missing data pattern may seem cumbersome, but it has the advantage of making explicit the parameters that

cannot be identified by the observed data. These parameters, often called "sensitivity parameters" (see Section 1.6.3) make a suitable basis for formulating sensitivity analysis and for identifying the conditional distribution of missing responses using informative prior distributions; in general, mixture models lend themselves well to parameterisations having these properties.

When missingness is monotone it is possible to summarise the patterns by dropout time and directly model the dropout process (Daniels and Hogan, 2008; Gaskins et al., 2016). In addition, for monotone patterns, a precise definition of the MAR condition exists, requiring that that the marginal probability of dropping out at a specific time j cannot depend on measurements at or beyond j (Molenberghs et al., 1997). However, when missingness is non-monotone, the plausibility and appropriateness of the MAR/ignorability assumption has been debated in the literature. Robins and Gill (1997) and Vansteelandt et al. (2007) argued that a specific MAR condition for non-monotone missingness patterns implies the existence of MAR missingness mechanisms in different portions of the data. Nevertheless, due to the precise way in which this must occur in order for the MAR condition to jointly hold for all patterns, the "natural missing data processes" for non-monotone missingness patterns, especially within a longitudinal framework, will typically be MNAR. The general definition of the MAR condition and the ignorability assumption provided in Section 1.5.2 offer a natural starting point for an analysis of partially-observed data even if, especially for non-monotone patterns, they are unlikely to hold precisely. This is why in longitudinal applications, precise discussions over the MAR mechanism are usually of secondary interest with respect to including appropriate (relatively) fully observed variables which can be used to improve the estimate of the distribution of the missing data given the observed data (Molenberghs et al., 2015).

When missingness is non-monotone and the data are sparse, alternative strategies for identifying the distribution of the missing data can be used. Linero and Daniels (2018) provide a review of these approaches, which are now briefly presented.

Permutation missingness. A first approach, proposed by Robins and Gill (1997), is the class of permutation missingness models. Let $\overline{\boldsymbol{y}}_j = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j)$ and $\tilde{\boldsymbol{y}}_j = (\boldsymbol{y}_{j+1}, \ldots, \boldsymbol{y}_J)$ denote the history of the response up to time j and the future of the response strictly after time j, respectively. Thus, the full response data vector can be expressed as $\boldsymbol{y} = (\overline{\boldsymbol{y}}_j, \tilde{\boldsymbol{y}}_j)$, and similarly for the missingness indicator vector $\boldsymbol{r} = (\overline{\boldsymbol{r}}_j, \tilde{\boldsymbol{r}}_j)$. Let also $\overline{\boldsymbol{o}}_j$ and $\tilde{\boldsymbol{o}}_j$ denote the observed data (including the $\boldsymbol{r}_j$) up to time j and the observed data strictly after time j. The permutation missingness model assumes:

$$p(\boldsymbol{r}_j \mid \boldsymbol{y}, \tilde{\boldsymbol{r}}_j) = p(\boldsymbol{r}_j \mid \overline{\boldsymbol{y}}_{j-1}, \tilde{\boldsymbol{o}}_j),$$

possibly after applying an a priori known permutation to $\boldsymbol{y}$. In words, these models assume that missingness at time j can depend on the "past" and the "observed future", but not on the present, where the notion of time is determined by the given permutation.

Sequential explainability. Vansteelandt et al. (2007) proposed the sequential explainability restriction, which assumes

$$p(\boldsymbol{y}_j \mid \overline{\boldsymbol{o}}_j, \boldsymbol{r}_j = \boldsymbol{0}) = p(\boldsymbol{y}_j \mid \overline{\boldsymbol{o}}_j, \boldsymbol{r}_j = \boldsymbol{1}),$$

that is, it assumes that the observed responses prior to time j are sufficient to predict whether or not a subject will be measured at time j, while the outcome at time j is not predictive.

Nearest identified pattern. Linero (2017) introduced the nearest identified pattern restriction,

which assumes

$$p(\boldsymbol{y}_j \mid \boldsymbol{r}_j, \boldsymbol{y}) = p(\boldsymbol{y}_j \mid \boldsymbol{r}_j^\star, \boldsymbol{y}),$$

where, $\boldsymbol{r}_j^\star$ corresponds to $\boldsymbol{r}_j$ but with the $j-$th component fixed at 1. The model assumes that, conditional on all other observed quantities being equal, missingness at time j is not predictive of the response at time j.

Itemwise conditional independence. Sadinle and Reiter (2017) proposed the itemwise conditional independence assumption

$$p(\boldsymbol{y}_j \mid \boldsymbol{r}_j = \boldsymbol{0}, \boldsymbol{r}_{-j}, \boldsymbol{y}_{-j}) = p(\boldsymbol{y}_j \mid \boldsymbol{r}_j = \boldsymbol{1}, \boldsymbol{r}_{-j}, \boldsymbol{y}_{-j}),$$

where, $\boldsymbol{r}_{-j}$ and $\boldsymbol{y}_{-j}$ denote the full vectors $\boldsymbol{r}$ and $\boldsymbol{y}$ with the $j-$th component removed. The model assumes that, conditional on all other quantities (both observed and unobserved) being equal, missingness at time j is not predictive of the response at time j.

Pairwise missing at random. Tchetgen Tchetgen et al. (2016) introduced the pairwise missing at random restriction

$$p(\boldsymbol{y}_j \mid \boldsymbol{r}_j = \boldsymbol{0}, \boldsymbol{y}) = p(\boldsymbol{y}_j \mid \boldsymbol{r} = \boldsymbol{1}, \boldsymbol{y}),$$

which assumes that the distribution of the missing values of a subject at each time j can be approximated using the observed response at time j of an equivalent subject who was observed at all measurement times (i.e. a completer). Linero and Daniels (2018) showed that pairwise missing at random is an extension of the complete case missing value restriction (defined for monotone missingness) to non-monotone missingness.

All these classes of restrictions are phrased in terms of conditional independence relationships which, however, are not themselves particularly plausible when $\boldsymbol{y}_j$ is thought to directly influence $\boldsymbol{r}_j$. Indeed, these assumptions are not used because the conditional independences they suggest are believed to hold, but rather they are only used as benchmark assumptions from which alternative non-ignorable departures can be explored in sensitivity analysis. Because of this, an alternative and more intuitive modelling strategy which can handle both monotone and non-monotone missingness patterns, based on the so-called extrapolation factorisation, is now introduced.

1.6.1 Extrapolation Factorisation

The full data model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ can be expressed as

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}}) p(\boldsymbol{y}_{mis}^{\boldsymbol{r}} \mid \boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r}, \boldsymbol{\omega}_{\mathsf{E}}), \tag{1.24}$$

where $\boldsymbol{\omega}_{\mathsf{E}}$ and $\boldsymbol{\omega}_{\mathsf{O}}$ are (possibly overlapping) subsets of $\boldsymbol{\omega}$, while $\boldsymbol{y}_{obs}^{\boldsymbol{r}}$ and $\boldsymbol{y}_{mis}^{\boldsymbol{r}}$ indicate the observed and missing responses within pattern $\boldsymbol{r}$, respectively. Equation 1.24 (Linero and Daniels, 2015) is often called the *extrapolation factorisation* and factors the joint distribution in two components. The observed data distribution $p(\boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$ is completely identified by the data, while the extrapolation distribution $p(\boldsymbol{y}_{mis}^{\boldsymbol{r}} \mid \boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r}, \boldsymbol{\omega}_{\mathsf{E}})$ remains unidentified by the data in the absence of unverifiable assumptions about the full data (Daniels and Hogan, 2008).

To specify the observed data distribution $p(\boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$ a possible strategy is to use a *working model* $p^\star$ (Linero and Daniels, 2015) for the joint distribution of the response and missingness

$$p(\boldsymbol{y}_{obs}^{\boldsymbol{r}}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}}) = \int p^\star(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) d\boldsymbol{y}_{mis}. \tag{1.25}$$

Since $p^\star$ is used only to obtain a model for $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$ and not as a basis for inference, the extrapolation distribution is left unidentified. Thus, any inference depending on the observed data distribution may be obtained using Equation 1.25 as the true model, with the advantage that it is often easier to specify a model for the full data compared with a model for the observed data. In principle, any factorisation can be used to specify $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$. For example, we can use the factorisation of pattern mixture models, which lend themselves particularly well to parameterisations based on the extrapolation factorisation.

This modelling approach is particularly appealing for handling missingness because the parameters of the extrapolation distribution cannot be identified from data and make therefore a suitable basis for conducting sensitivity analysis. Typically, the values of these parameters are arbitrarily set or informative priors (under a Bayesian framework) are used to assess the robustness of the inferences to plausible variations.

1.6.2 Sensitivity Analysis

Exploring the sensitivity of the results with respect to different missingness assumptions is a crucial task in any analysis that deals with partially-observed data.

Sensitivity analysis represents an extremely valuable tool to deal with the uncertainty induced by missingness. Formally, sensitivity analysis is a technique used to determine how different input values in a model will impact the output, under a given set of assumptions. When applied to missing data, sensitivity analysis corresponds to exploring plausible missing data assumptions and assessing how consistent results are across different scenarios. The degree to which conclusions (inferences) are stable across these scenarios provides an indication of the confidence that can be placed in the results.

Bayesian methods are particularly well-suited for conducting sensitivity analysis by specifying alternative prior distributions to encode different assumptions about missingness. In addition, the full propagation of the uncertainty in a Bayesian model allows to easily assess and quantify the impact of these assumptions on the posterior results.

In principle, sensitivity analysis can be implemented in any type of nonignorable models, but the nature and interpretation of the assessment changes according to the specific features of each approach. Here, we focus on the extrapolation factorisation, where the separation between the parameters that are identified and those that cannot be identified by the observed data offers a convenient framework for conducting sensitivity analysis.

1.6.3 Identifying Restrictions and Sensitivity Parameters

Recall that the full data distribution can be factored into an extrapolation distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r}, \boldsymbol{\omega}_{\mathsf{E}})$ and an observed data distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$, where $\boldsymbol{\omega}_{\mathsf{E}}$ and $\boldsymbol{\omega}_{\mathsf{O}}$ denote the parameters indexing the two distributions. In general, the extrapolation distribution can be identified through *identifying restrictions* and informative priors on the parameters of the distribution of the missing responses, typically called *sensitivity parameters*.

Identifying restrictions correspond to assumptions about $p(\boldsymbol{y}, \boldsymbol{r})$, which link the observed data distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}_{\mathsf{O}})$ to the extrapolation distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r}, \boldsymbol{\omega}_{\mathsf{E}})$. According to the way the two factors are linked, different types of restrictions can be used. For example, the *complete case missing value restriction* identifies the full data distribution by equating the extrapolation distribution to the distributions of the completers $p(\boldsymbol{y} \mid \boldsymbol{r} = \boldsymbol{1}, \boldsymbol{\omega}_{\mathsf{C}})$, indexed by the parameters $\boldsymbol{\omega}_{\mathsf{C}} \subset \boldsymbol{\omega}_{\mathsf{O}}$. We refer to Daniels and Hogan (2008) for a comprehensive review of the most popular classes of identifying restrictions.

It is common practice to specify a single identifying restriction as a benchmark assumption and consider interpretable deviations from that benchmark to assess how inferences are driven by our assumptions (Wang and Daniels, 2011). When identification of the joint $p(\boldsymbol{y}, \boldsymbol{r})$ is not required, e.g. if the interest lies only on some parameters rather than the entire distribution, then *partial identifying restrictions* (Linero and Daniels, 2018) can be used to identify the distribution only up to the specific parameters of interest.

Identifying restrictions are often combined with sensitivity parameters to conduct sensitivity analysis to MNAR assumptions. These parameters are generally not identifiable from the observed data but, when their values are fixed, the full data model is identified. To illustrate this nonignorable strategy we consider the example of the mixture model $p(\boldsymbol{y} \mid \boldsymbol{r}) \sim \text{Normal}(\boldsymbol{\mu}^{\boldsymbol{r}}, \boldsymbol{\Sigma}^{\boldsymbol{r}})$, where $\boldsymbol{\mu}^{\boldsymbol{r}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{r}}$ are the mean vector and the variance matrix of $\boldsymbol{y}$ in each pattern $\boldsymbol{r}$, respectively. For the purpose of this example, we assume that the variance matrix is common across the patterns $\boldsymbol{\Sigma}^{\boldsymbol{r}} = \boldsymbol{\Sigma}$ and that it can be identified from the data.

We assume that only dropout missingness occurs and that the only missing responses are those at the last follow-up, i.e. $\boldsymbol{y}_{mis} = \boldsymbol{y}_J^{\boldsymbol{r}^{\neq \mathbf{1}}}$. Within this setting, $\boldsymbol{\omega}_{\mathsf{O}} = (\boldsymbol{\mu}_1^{\boldsymbol{r}}, \ldots, \boldsymbol{\mu}_{J-1}^{\boldsymbol{r}}, \boldsymbol{\mu}_J^{\boldsymbol{r}=\mathbf{1}}, \boldsymbol{\Sigma})$ and $\boldsymbol{\omega}_{\mathsf{E}} = \boldsymbol{\mu}_J^{\boldsymbol{r}^{\neq \mathbf{1}}}$ are the sets of parameters indexing the observed data and extrapolation distributions, respectively. Using the complete case missing value restriction, we identify $\boldsymbol{\omega}_{\mathsf{E}}$ using the subset of $\boldsymbol{\omega}_{\mathsf{O}}$ related to the corresponding parameters that index the completers $\boldsymbol{\omega}_{\mathsf{C}} = \boldsymbol{\mu}_J^{\boldsymbol{r}=\mathbf{1}}$. This amounts to assuming a very informative (point mass) prior:

$$p(\boldsymbol{\omega}_{\mathsf{E}} \mid \boldsymbol{\omega}_{\mathsf{O}}) = I\{\boldsymbol{\omega}_{\mathsf{E}} = \boldsymbol{\omega}_{\mathsf{C}}\}, \tag{1.26}$$

where $\boldsymbol{\omega}_{\mathsf{C}} \subset \boldsymbol{\omega}_{\mathsf{O}}$. Identifying restrictions typically impose very restrictive assumptions about the parameters of the extrapolation distribution and fail to characterise the missingness uncertainty. However, Equation 1.26 provides a useful framework for incorporating some sensitivity parameters.

More generally, we can reparameterise the model in terms of $\boldsymbol{\xi}(\boldsymbol{\omega}) = (\boldsymbol{\xi}_{\mathsf{S}}, \boldsymbol{\xi}_{\mathsf{M}})$, where $\boldsymbol{\xi}_{\mathsf{M}}$ are the parameters fully identified from the data while $\boldsymbol{\xi}_{\mathsf{S}}$ are the sensitivity parameters. Daniels and Hogan (2008) provide a formal definition of sensitivity parameters, which must satisfy specific conditions. Consider a full data model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ and its extrapolation factorisation as in Equation 1.24. Let $\boldsymbol{\xi}(\boldsymbol{\omega}) = (\boldsymbol{\xi}_{\mathsf{S}}, \boldsymbol{\xi}_{\mathsf{M}})$ denote a reparameterisation of the full data parameter $\boldsymbol{\omega}$ such that: 1) $\boldsymbol{\xi}_{\mathsf{S}}$ is a nonconstant function of $\boldsymbol{\omega}_{\mathsf{E}}$; 2) the observed data distribution is a constant function of $\boldsymbol{\xi}_{\mathsf{S}}$; 3) at a fixed value of $\boldsymbol{\xi}_{\mathsf{S}}$, the observed data distribution is a nonconstant function of $\boldsymbol{\xi}_{\mathsf{M}}$. Then, when all these three conditions are satisfied, $\boldsymbol{\xi}_{\mathsf{S}}$ is a sensitivity parameter.

The use of a pattern mixture approach allows a straightforward incorporation of $\boldsymbol{\xi}_{\mathsf{S}}$ through the reparameterisation of the full data parameter $\boldsymbol{\omega} = (\boldsymbol{\omega}_{\mathsf{O}}, \boldsymbol{\omega}_{\mathsf{E}})$ as $\boldsymbol{\xi}(\boldsymbol{\omega}) = (\boldsymbol{\xi}_{\mathsf{S}}, \boldsymbol{\xi}_{\mathsf{M}})$. For example, in the mixture model considered before, the parameters indexing the observed data distribution correspond to those fully identified from the data ($\boldsymbol{\omega}_{\mathsf{O}} = \boldsymbol{\xi}_{\mathsf{M}}$) while the parameters indexing the extrapolation distribution can be identified using the sensitivity parameters ($\boldsymbol{\omega}_{\mathsf{E}} = \boldsymbol{\xi}_{\mathsf{S}}$). Then, within the framework of Equation 1.26, $\boldsymbol{\xi}_{\mathsf{S}}$ can be expressed as a function of $\boldsymbol{\omega}_{\mathsf{C}} \in \boldsymbol{\omega}_{\mathsf{O}}$ and some redundant parameters $\boldsymbol{\delta}$

$$\boldsymbol{\omega}_{\mathsf{E}} = \boldsymbol{\xi}_{\mathsf{S}} = \boldsymbol{\omega}_{\mathsf{C}} + \boldsymbol{\delta}, \tag{1.27}$$

whose value must be specified according to the specific missingness assumptions made in the given context. Thus, Equation 1.27 provides a convenient framework to incorporate $\boldsymbol{\xi}_{\mathsf{S}}$ into the model, where the robustness of the results to departures from a benchmark assumption, e.g. $\boldsymbol{\xi}_{\mathsf{S}} = \boldsymbol{\omega}_{\mathsf{C}}$, can be easily assessed through the specification of alternative values of $\boldsymbol{\delta}$. Essentially, $\boldsymbol{\xi}_{\mathsf{S}}$ are parameters that index the extrapolation distribution which cannot be identified by the observed data and provide a framework for assessing sensitivity of model-based inferences to assumptions

about the missing data mechanism.

In a frequentist setting, a possible approach to incorporate ξ_S into the model is the so-called "delta-shift" method, which corresponds to adding/subtracting a constant value to the imputed data (for example obtained through MI) under a benchmark missingness assumption, typically MAR (Leurent et al., 2018b). This approach approximately corresponds to a degenerate Bayesian analysis which assumes a point mass prior on ξ_S, with uncertainty about the sensitivity parameters that can be captured by repeating the analysis for a fixed set of values for the sensitivity parameters. By contrast, in a Bayesian framework, uncertainty about ξ_S is typically captured using proper probability distributions, whose impact on the inferences is available in terms of the variations observed in the posterior quantities of interest. This would allow to specify a joint prior on the sensitivity parameters to account for the possible dependence among ξ_S while exploring some plausible missingness departures, which should be defined according to the available information (e.g. expert opinion).

1.6.4 Specifying Priors on the Sensitivity Parameters

Prior distributions on the sensitivity parameters should be constructed in order to reflect uncertainty about the missing data assumptions within the context analysed. A common approach is to anchor full data models at a benchmark assumption (e.g. MAR), which typically coincides with a fixed point in the range of plausible values of the sensitivity parameters ξ_S. Inference under MNAR is then assessed in terms of departures from this benchmark scenario.

Given the reparameterisation $\xi(\omega) = (\xi_S, \xi_M)$, the priors on ξ_S are typically specified in terms of the parameters identified by the observed data ξ_M and some redundant parameters δ that explicitly capture the departures from the benchmark. More specifically, given a one-to-one function $h(\xi_M, \delta)$, the sensitivity parameters can be re-expressed as (Daniels and Hogan, 2008)

$$\xi_S = h(\xi_M, \delta), \tag{1.28}$$

where δ captures the information about the missing data mechanism. The quantity in Equation 1.28 is typically anchored to a benchmark by setting δ equal to some value δ_0 (e.g. $\delta_0 = \mathbf{0}$). This corresponds to assuming a point mass prior at δ_0 which encodes a particular missingness assumption with certainty.

Alternatively, priors that convey uncertainty about the missing data mechanism $\delta \sim p(\delta^\star)$ can be specified, where $\delta^\star$ are the hyperparameters of the non-degenerate priors on δ. Typically, values for $\delta^\star$ are informed by a credible source of external information, such as expert opinion if available, or calibrated on the observed data to obtain a range of plausible departures from the benchmark. The Bayesian framework is ideal for this task as it allows to quantify uncertainty about missing data assumptions through these priors.

Key points of this chapter:

- Trial-based economic evaluations are an important source of evidence for decision-makers to inform the decision about the funding or reimbursement of new healthcare technologies. Analyses typically compare alternative options in terms of some measures of effectiveness and costs that are collected for each participant at different time points in the study. These measures are then summarised into cross-sectional quantities (e.g. QALYs) on which the analysis is performed.
- Bayesian methods are well-suited to addressing decision-making problems, such as that in economic evaluations. By taking a probabilistic approach, based on decision rules and available information, they can explicitly account for both sampling and parameter uncertainty in the decision process and obtain an "optimal" decision output.
- Trial-based outcome data are almost invariably affected by missingness, which considerably complicates the task of the analyst as assumptions about the unobserved values cannot be checked from the data. Most of the missing data methods in the CEA literature typically rely on ignorability of the missing data mechanism as the default assumption when drawing inference. However, this cannot be verified from the observed data and plausible nonignorable departures should be explored.
- Nonignorable models are typically specified using some factorisation of the full data model. Pattern mixture models can be used in combination with the extrapolation factorisation as a particularly appealing way to model nonignorable missingness. It allows to separately fit a model to the observed data and identify the distribution of the missing data through suitably-defined identifying restrictions and sensitivity parameters. The Bayesian approach naturally allows the implementation of sensitivity analysis through the incorporation of external information into the model though informative priors on the sensitivity parameters.

Chapter 2

Literature Review

In this section we present the results of a focused [2] literature review that assesses the quality of the information reported and type of methods used to handle missing outcome data in trial-based economic evaluations. Recent reviews on missing data methods in within-trial CEAs (Noble et al., 2012; Diaz-Ordaz et al., 2014a; Hughes et al., 2016; Leurent et al., 2018a) concluded that CCA has historically represented the standard approach, but transparent information about missingness has rarely been provided to justify its use.

The purpose of this review is to critically appraise the current literature in within-trial CEAs with respect to the quality of the information reported and the methods used to deal with missingness for both effectiveness and costs. The review complements previous work (Noble et al., 2012), covering 2003-2009 (88 articles) with a new review, covering 2009-2015 (81 articles) and focuses on two perspectives.

First, we provide guidelines on how the information about missingness and related methods should be presented to improve the reporting and handling of missing data. We propose to address this issue by means of a *quality evaluation scheme*, providing a structured approach that can be used to guide the collection of information, formulation of the assumptions, choice of methods, and considerations of possible limitations for the given missingness problem. Second, we review the description of the missing data, the statistical methods used to deal with them and the quality of the judgement underpinning the choice of these methods. A synthesised version of this chapter in the form of a research article is published in *PharmacoEconomics-Open*.

2.1 Quality Evaluation Scheme

In order to judge whether missing data in CEAs have been adequately handled, we assembled guidelines from previous review articles (Wood et al., 2004; Noble et al., 2012; Faria et al., 2014; Diaz-Ordaz et al., 2014a) on how information relating to the missing data should be reported (Table 2.1). In particular, we defined three broad components of the analysis that are related to the description of the missingness problem (Description), details of the methods used to address it (Methods) and a discussion on the uncertainty in the conclusions resulting from the missingness (Limitations). For each component, information that is considered to be vital for transparency is listed under “key considerations”, while other details that could usefully be provided as supplementary material are suggested under “optimal considerations”.

[2] A *focused* or *trageted* review is similar to a systematic literature review in that it uses explicit methods to identify, select, critically appraise some key relevant research question. However, this approach is less comprehensive than a systematic review since the focus is on key (rather than exhaustive) research questions, and is typically performed by a single reviewer (Moher et al., 2015; Huelin et al., 2015)

Description	*Method*	*Limitations*
Key considerations	**Key considerations**	**Key considerations**
1. Report the number of individuals with missing data for each variable in the reported analysis by treatment group.	1. Identify a plausible missingness assumption for the specific patterns and setting analysed.	1. Acknowledge and quantify the impact of the missing data on the results.
2. Describe the missing data patterns for all variables included in the economic analysis (is missingness on one variable associated with missingness on another variable?, is there a longitudinal aspect to the data?)	2. State the method and software used in the base-case analysis.	2. State possible weaknesses and issues with respect to the method and assumptions.
3. Discuss plausible reasons why values are missing (e.g. death).	3. For more general methods provide details about their implementation *	
	4. Perform a plausible robustness analysis; provide and discuss the results.	
Optimal considerations	**Optimal considerations**	
1. Provide supplementary material about the preliminary analysis on missingness (e.g. descriptive plots and tables)	1. Provide supplementary material about the method implementation in the base-case and robustness analysis (e.g. software implementation code)	

*For example, in Multiple Imputation, state the imputation model specification and variables included, the number of imputations, post imputation checks.

Table 2.1: List of the information content for each of the three components that we would like to observe in the studies in order to achieve a full reporting of the missing data analysis. The contents are divided into two subgroups: key and optimal considerations. The former are mandatory for transparency in the presence of missing data while the latter are additional considerations that can be provided.

Using the list of key considerations in Table 2.1, we determine whether null (all key considerations absent), partial (one or more key considerations absent) or full (all key considerations present) information has been provided for each component. The set of key considerations is defined to ensure a full assessment of the impact that missingness may have on the final conclusions of the analysis with respect to all three components. However, providing a certain level of information on one component (e.g. full information on Description) typically has a different impact on the results with respect to providing the same level of information on another component (e.g. full information on Limitations).

Based on this, we suggest computing a numerical score that weights each component by the impact that it may have on the final results to summarise the overall information provided on missingness. Table 2.2 shows the proposed scores for each of the three components by level of the information content.

Score / Content	Description	Method	Limitations
Full (F)	6	4	2
Partial (P)	3	2	1
Null (N)	0	0	0

Table 2.2: Numerical scores associated with the level of the information content (Full, Partial or Null) provided in each of the three components (Description, Method and Limitations). The level of information content is defined according to the number of key considerations satisfied for each component (Table 2.1). The scores are computed by weighting the components using a ratio 3:2:1 (Description, Method and Limitations).

Different score values are calculated based on whether full, partial or null information content is provided in each component and by weighting the three components in a ratio of 3:2:1 (Description: Method: Limitations). This weighting scheme has been chosen according to the impact that each component is likely to have on the final conclusions based on assumptions that we deemed to be reasonable. Specifically, the Limitations component typically has the least importance among the three because of its limited impact on the conclusions. In the same way, the

Description component has potentially a higher impact on the results than the Method component as it generally drives the choice for the initial assumptions about the missingness.

Finally, the relevance of the scores in terms of decision analysis is mainly associated with a qualitative assessment of the articles. Therefore, we suggest converting the scores into ordered grades (A-E) to evaluate the studies based on the overall information reported on the handling of the missing data. Studies that are graded in the top categories should be associated with a higher degree of confidence in their results, whereas more caution should be given in the consideration of results coming from studies that are graded in the bottom categories. When qualitatively assessing the articles, the different grading assigned to each of them could be an indication of a lack in the robustness of the conclusions provided due to missingness uncertainty. The complete list of the grades (and the associated scores) can be interpreted as follows:

A (12) The highest quality judgement, identified by the upper thicker blue path in Figure 2.1, including only those studies that simultaneously provide all the key considerations for all the components. It is the benchmark for a comprehensive explanation and justification of the adopted missing data method.

B (9-11) Includes studies providing full details for either the description or the method and at least partial information for the other components. Studies with no information about the limitations are only included in this category if full detail is provided for both the other components.

C (6-8) Studies for which information about missingness is not well-spread across the components. All key considerations are provided either for the description or the method, but with only a partial or no content in the other components. Alternatively, we can have partial content for description and method, and partial or full content for limitations.

D (3-5) Indicates a greater lack of relevant information about missingness. Despite possibly including key considerations on any of the components, the information provided will at most be partial for the description in which case it will be combined with a total lack of content on either the method or the limitations.

E (0-2) The worst scenario where the overall information about the missing data is considered to be totally unsatisfactory. No description is given and we can observe at most only some of the key considerations for the method.

With respect to the quality assessment of the studies, the aggregation of the quality scores on the components of the analysis (Description, Method and Limitations) into ordered grades could lead to some loss of information compared with the direct use of the quality scores on each component. However, merging the scores into a fewer number of categories ensures a relatively easy comparison of the quality of the information provided across the three analysis components and provides a useful indication about the different degree of confidence to assign to the results obtained by each study.

Figure 2.1 shows a visual representation of the grade (and score) assignment in the quality evaluation scheme. Although the importance between the different components is subjective, the chosen structure represents a reasonable and relatively straightforward assessment scheme.

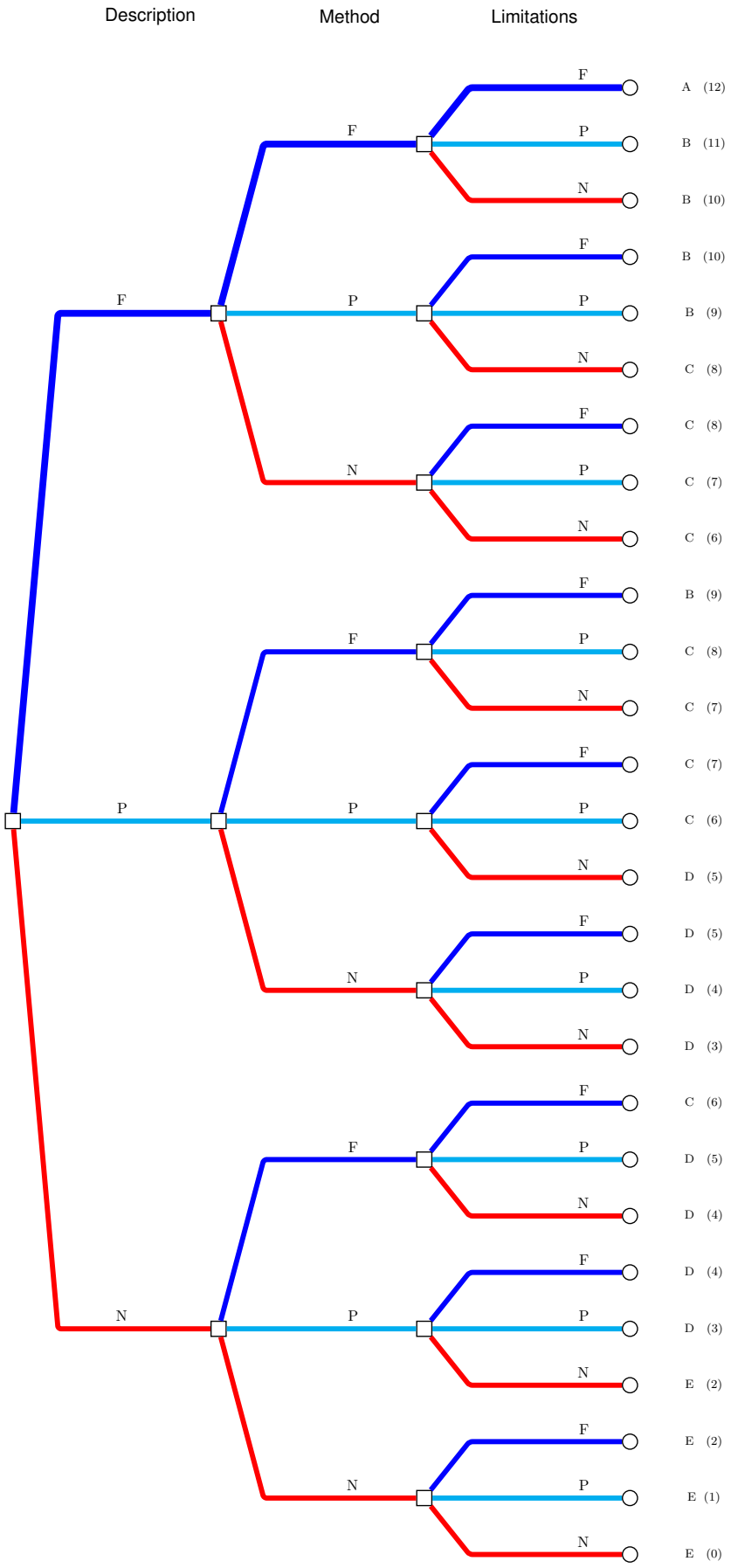


Figure 2.1: Diagram representation for the quality score grades (E-A) based on final scores (0 - 12). Branches colour represents the different way information can be provided: Red=Null information (N), Light Blue=Partial information (P), Blue=Full information (F).

2.2 Review

Noble et al. (2012), reviewed the methods used to handle missing cost measures in 88 articles published during the period 2003-2009. The review is extended to include missing effects, while using the same strategy to identify papers in the subsequent period, 1 April 2009 to 31 December 2015. A list of all the articles reviewed is available in Appendix E.

The choice of the on-line databases from which to extract the articles and the eligibility criteria used to select the studies to include in the review for the period 2009-2015 were based on those used by Noble et al. (2012). This would ensure a fair comparison of the conclusions based on the articles between the two periods of the review. The search engines of three on-line full-text journal repositories were used: 1) `Science-Direct.com`, 2) `bmj.com`, and 3) The Database of Abstracts of Reviews of Effects and NHS Economic Evaluation Database. The searches of the databases identified 1129 articles most of which were duplicates. After abstract review, 128 articles were considered, of which 81 fulfilled the eligibility criteria.

Articles were considered eligible for the review only if they were cost-effectiveness analyses within RCTs, used individual patient-level data and mentioned `missing data` in the text. The key words used in the search strategy were (`cost effectiveness` OR `economic evaluation`) AND `missing data` AND `trial` AND (`randomised` OR `randomized`). One author examined the abstract of each article and any article that was not an economic evaluation of a randomised controlled trial was excluded, as were articles that did not use individual patient data. Data were recorded on an Excel spreadsheet by one author. Data were extracted on type of economic evaluation. The following information was also extracted for those articles, which were cost-effectiveness analyses of randomised controlled trials that used individual patient data: the methods of dealing with missing cost and effect data, the year of publication, the number of complete cases used in the cost-effectiveness analysis and the overall sample size.

The articles reviewed for the two periods are presented and compared by type of analysis performed. First, the base-case methods are considered, i.e. those used in the main analysis. Second, any alternative methods in these analyses are discussed; when present, these assess the robustness of the results obtained in the main analysis against departures from the initial assumptions on missingness.

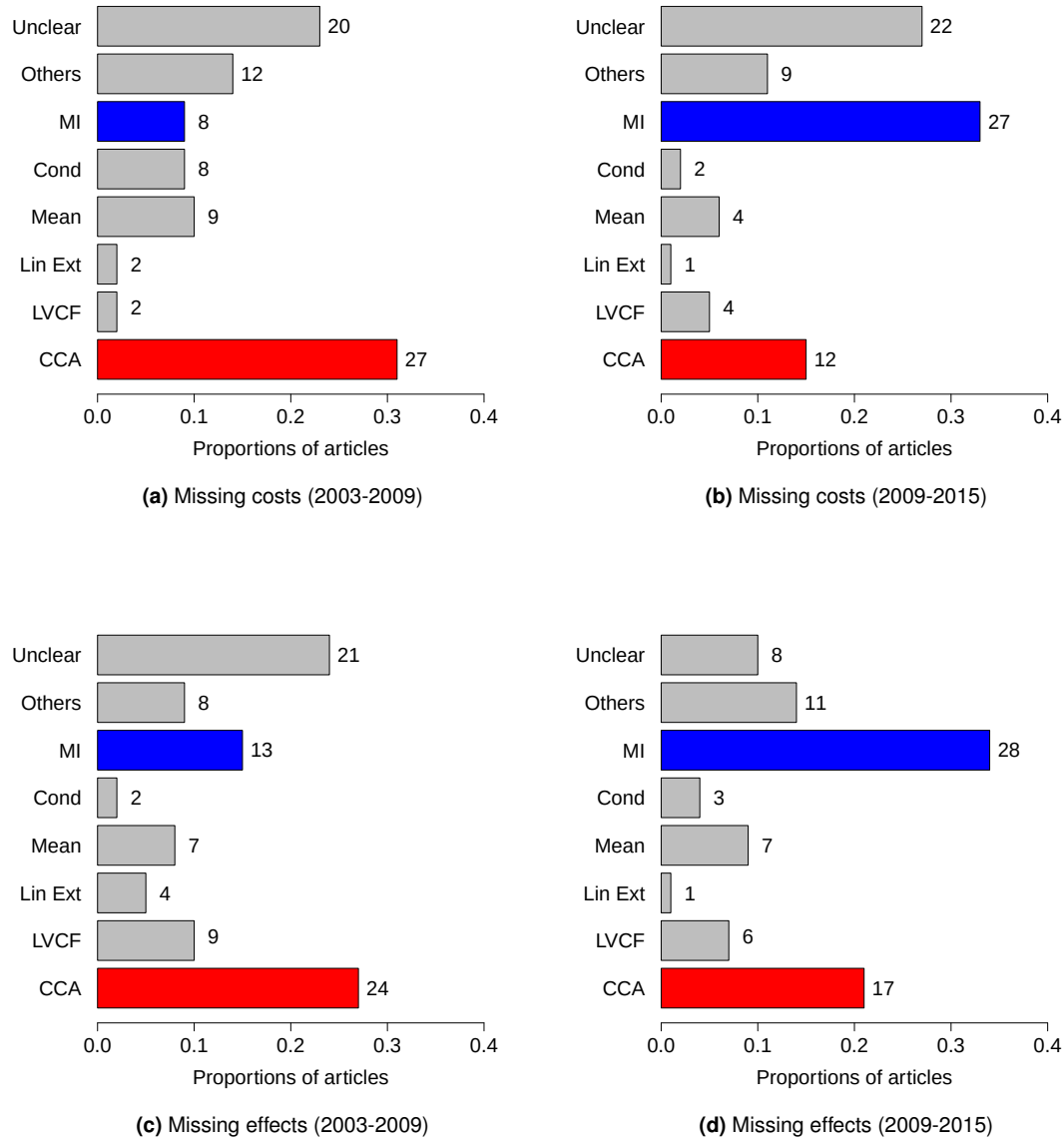
2.2.1 Base-case Analysis

The numbers of reviewed articles using a different base-case method are displayed in Figure 2.2. In the graphs, the methods associated with the largest number of articles for both outcomes between 2003-2009 and 2009-2015 are denoted with red (CCA) and blue (MI) bars, respectively.

As shown in Figure 2.2 (a), Noble et al. (2012) found that CCA was the most popular base-case method, used in 31% of the papers; 23% were unclear about the technique adopted. Different single imputation methods were adopted; among the most popular, mean imputation and conditional imputation were used in 10% and 9% of the articles respectively. MI was found in 9% of the articles. Our analysis of the methods for missing effectiveness measures shows a similar pattern in Figure 2.2 (c). CCA was used in 27% of the cases and with a sizeable proportion of papers unclear about the technique adopted (24%). Single imputation methods are here dominated by last value carried forward (10%), while a slightly higher proportion uses MI (15%).

In 2009-2015, MI replaces CCA as the most frequently used base-case method in both costs and effects, at 33% and 34% respectively – Figures 2.2 (b) and 2.2 (d). However, CCA is still the method of choice in many papers (15% for costs and 21% for effects). The proportion of papers that are unclear about the chosen method is similar over the two time periods for costs, but it is reduced by 50% in the later period for effects. This may be due to the fact that, in general,

clinical effectiveness measures and the estimate of treatment effect are typically the main focus of the analyses, whereas costs are less frequently included as a primary outcome in the study research questions. This will in turn translate into a more careful and reasoned examination of the missingness problem in the effect compared to the cost analysis.



(a) Missing costs (2003-2009)

(b) Missing costs (2009-2015)

(c) Missing effects (2003-2009)

(d) Missing effects (2009-2015)

Figure 2.2: Review of the base-case methods used to handle missing cost and effect data between 2003-2009 and 2009-2015. Legend: Complete Case Analysis (CCA), Last Value Carried Forward (LVCF), Linear Extrapolation (Lin Ext), Mean Imputation (Mean), Conditional Imputation (Cond), Multiple Imputation (MI), any other method present in less than 4 articles (Others), unspecified method (Unclear). The numbers to the right of the bars in the graphs are the number of papers including the corresponding method in the base-case analysis.

2.2.2 Robustness Analysis

With the term "robustness analysis" we refer to any analysis using different missing data methods compared to those in the base-case analysis, which is implemented without an appropriate justification, and which therefore is unlikely to provide a plausible approach to handle missingness in the context considered. By contrast, with the term "sensitivity analysis" we refer to the concept intro-

duced in Section 1.6.2 and which can be thought of as structurally varying the assumptions about the missingness model, whose plausibility is justified in light of known information. The difference between the two types of analysis consists in whether the additional methods implemented to assess the robustness of the results to the missing data assumptions can be considered plausible (sensitivity analysis) or not (robustness analysis) in the given context based on the available information.

However, in practice, even robustness analyses are rarely performed in CEAs. This poses an important question related to the reliability of the findings, as they may be affected by the specific assumptions about the missing data. From both review periods it seems that a robustness analysis is infrequently used and typically involves only one alternative scenario. This is not likely to be an optimal choice as the main objective of this analysis is to explore as many plausible alternative missing data assumptions as possible.

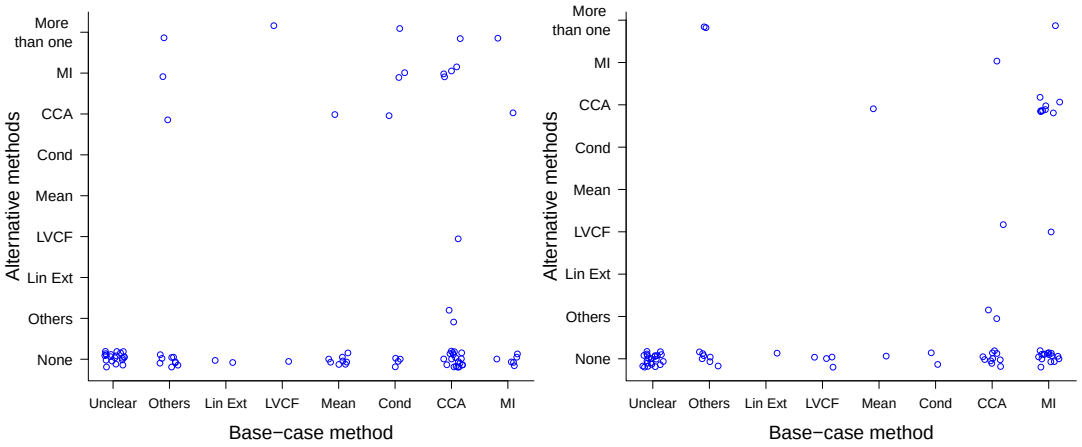
Noble et al. (2012) found that 75% of the articles did not include any robustness analysis, with the remaining papers typically performing an analysis by comparing CCA and MI. Similar findings apply to missing effects, with about 76% of the studies lacking any alternative missing data method. Similarly in the 2009-2015 review, we observe no robustness analysis in the majority of the articles for both costs (75%) and effects (70%).

Figure 2.3 provides a pictorial overview of the alternative methods used for cost and effect data. For costs, shown in Figure 2.3 (a) and Figure 2.3 (b), most articles describe no alternative analysis. In the earlier period, the choice of alternative missingness methods seems well-spread across CCA (4 cases), MI (7 cases) and the use of more than one method (5 cases), with a slightly more frequent adoption of MI. By contrast, in the later period, CCA (9 cases) is by far the most used robustness method, followed by more than one method (3 cases). Figures 2.3 (c) and 2.3 (d) describe the effects, with most of the articles not reporting any robustness analysis and with a decrease in MI (from 6 to 2 cases) used for robustness, opposed to an increase in CCA (from 4 to 7 cases), between the two periods.

Finally, the costs and effects graphs show a similar change with respect to the methods used in combination between the base-case and robustness analyses in the two periods. Specifically, in the earlier period, MI is the most used robustness method in combination with CCA as the most used base-case method (4 cases for the costs and 3 for the effects); conversely, in the later period, the situation is reversed, with CCA being used as the robustness method in combination with MI as the base-case method (8 cases for the costs and 7 cases for the effects). Two-way frequency tables are included in Appendix C to provide detailed information about the number of the studies that belong to each combination of base-case and robustness methods shown in Figure 2.3 for costs and effects in both periods.

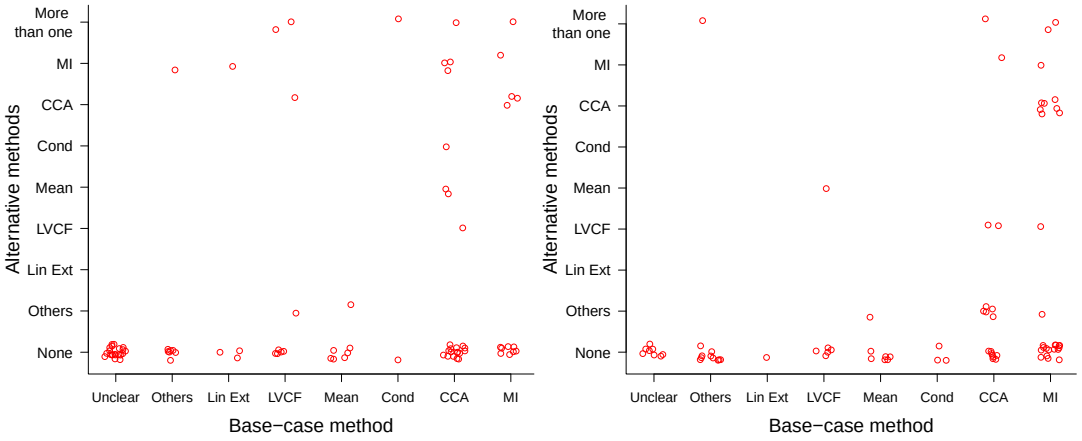
2.3 Application of the quality evaluation scheme to the reviewed articles

Comparing the information provided in the reviewed articles to the list of considerations given in Table 2.1 allows to assess the quality of the missingness handling in the studies. More specifically, the articles are classified from the perspective of the strength of the assumptions about the missingness mechanism. This is related to the choice of method, since each is underpinned by some specific missing data assumption. It is possible to view the quality judgement and strength of assumptions as two dimensions providing a general mapping of how the missingness problem is handled. This applies to both the level of knowledge about the implications of a given missingness assumption on the results and how these are translated into the chosen method. Details about the evaluation of both aspects are provided next, starting with the strength of assumptions.



(a) Robustness Analyses Costs (2003-2009)

(b) Robustness Analyses Costs (2009-2015)



(c) Robustness Analyses Effects (2003-2009)

(d) Robustness Analyses Effects (2009-2015)

Figure 2.3: Comparison of methods used in the base-case analysis (x axis) and those used as alternatives in a robustness analysis (y axis) for the articles between 2003-2009 and 2009-2015 for missing costs and effects. Legend: unspecified methods (Unclear), other methods (Others), Linear Extrapolation (Lin Ext), Last Value Carried Forward (LVCF), Mean Imputation (Mean), Conditional Imputation (Cond), Complete Case Analysis (CCA), Multiple Imputation (MI). Jittering is used in all graphs to avoid overplotting and visualise all the points.

Methods are grouped into five categories, broadly ordered according to the strength of the associated missingness assumptions. These are: *Single Imputation* (SI); *Complete Case Analysis* (CCA); *Multiple Imputation* (MI); and *Unknown* (UNK), a residual group in which we classify studies that do not explicitly mention the method used. We associate this class with the strongest level of assumptions, since the lack of any method description may implicitly suggest (over)confidence in a small effect of missingness on the results. By contrast, we define *Sensitivity Analysis* (SA) as the least restrictive approach, which can assess the robustness of the results to different alternative missing data assumptions.

The suitability of each method to handle missingness is very dependent on the specific problem considered and the available information. For example, as noted in Section 1.5.3, when partially-observed covariates are included in the analysis, the validity of CCA does not fit neatly into Rubin's

categories as it is not only valid under MCAR but also under MAR and MNAR under certain conditions (see Appendix A.3). Thus, the broad distinction between the methods in terms of their strength about the missing data assumptions is based on some simplifying assumptions. Specifically, we assume that SI methods are valid only under very restrictive missing data assumptions (e.g. some implausible MNAR scenarios such as that of LVCF) and that there are no covariates in the analysis model, which makes CCA valid only under MCAR. By contrast, the possible inclusion of auxiliary variables in the imputation model allows MI to be valid under MAR.

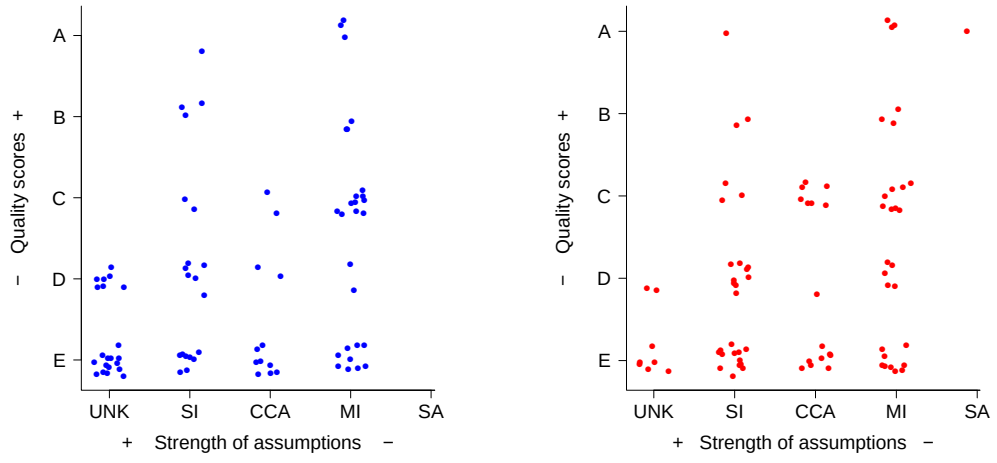
Figure 2.4 gives a graphical representation of both aspects for the articles reviewed between 2009-2015 in terms of the assumptions and justifications (quality scores) on missingness. In both graphs, more studies lie in the lower than in the upper part, indicating that fewer studies can be classified as high quality in terms of the considerations about missingness. This is highlighted by a greater concentration of points at the bottom of the figures (grade E). As we move along the vertical axis, this tends to reduce up to the top level (grade A), where there are only 4 and 5 cases for the cost and effect analyses, respectively. Of particular interest is the (almost) total absence of articles that performed a sensitivity analysis, clearly indicating very slow uptake of this technique. Detailed information about the number of the studies that belong to each combination of quality score and missing data assumption categories for both the costs and effects shown in Figure 2.4 is provided using two-way frequency tables, which are available in Appendix C.

A shift along the vertical axis in the graphs indicates an increase in the level of understanding about the implications on the results for different choices of the missing data assumptions. Therefore, an upward movement in the plot will always improve the justification of a specific assumption. However, to be able to follow this path we may have to rely on more sophisticated methods that can match the given missingness assumption, i.e. if we think our data are MNAR, then CCA assumptions are less likely to hold. The aim of an optimal analysis should be to select a method that can be fully justified by matching the description of the missing data problem to the assumptions underpinning the chosen method, i.e. map onto the upper section of the graphs.

As a concrete example about the importance of exploring different missingness assumptions in terms of the impact they may have on the CEA, we consider one of the reviewed studies that has been graded as "A" by our scheme (Pickard et al., 2015). The authors provide an assessment of the probability of accepting a given treatment against a comparator (CEAC) for different willingness to pay thresholds and missingness methods. For all the thresholds considered, substantial uncertainty is reflected by significant variations in the CEAC values according to the different missing data methods used. Specifically, the incorporation of external information leads to missingness assumptions that significantly affect the uncertainty in the results, producing a decrease in the value of the CEAC, at the target willingness to pay threshold, from 71% in the base-case (CCA) to 53% in one alternative scenario (MI). This example should encourage authors to recognise the importance that a comprehensive examination of missingness via sensitivity analysis may have on the uncertainty around CEA conclusions.

2.4 Summary of the Findings

Our review is based on a sample of recently published studies and should therefore provide a picture of current missing data handling in within-trial CEAs. However, the quality assessment of the articles is based on the information reported in the articles. It is possible that authors had assessed the robustness of their conclusions to the missing data using alternative approaches that were not reported in the published version because of space limitations in journals. In these cases, it is important that on-line appendices and supplementary material are used to report these alternatives. In our literature review, information about missing data information and meth-



(a) Missing Cost Analyses (2009-2015)

(b) Missing Effect Analyses (2009-2015)

Figure 2.4: Joint assessment, in the reviewed articles between 2009-2015, for missing costs and effects, of two components. The x-axis is the missingness method assumptions: Unknown (UNK), Single Imputation (SI), Complete Case Analysis (CCA), Multiple Imputation (MI) and Sensitivity Analysis (SA). The y-axis is the ordered classification for the quality judgement (grades) to support these assumptions: E, D, C, B, A. Jittering is used in all graphs to avoid overplotting and visualise all the points.

ods was available from 4 and 9 on-line supplementary materials for the period 2003-2009 and 2009-2015, respectively. Both the larger number of on-line materials and more detailed information reported about missingness handling in the analyses indicate an increased use of this tool in the later period (2009-2015) compared to the first period (2003-2009).

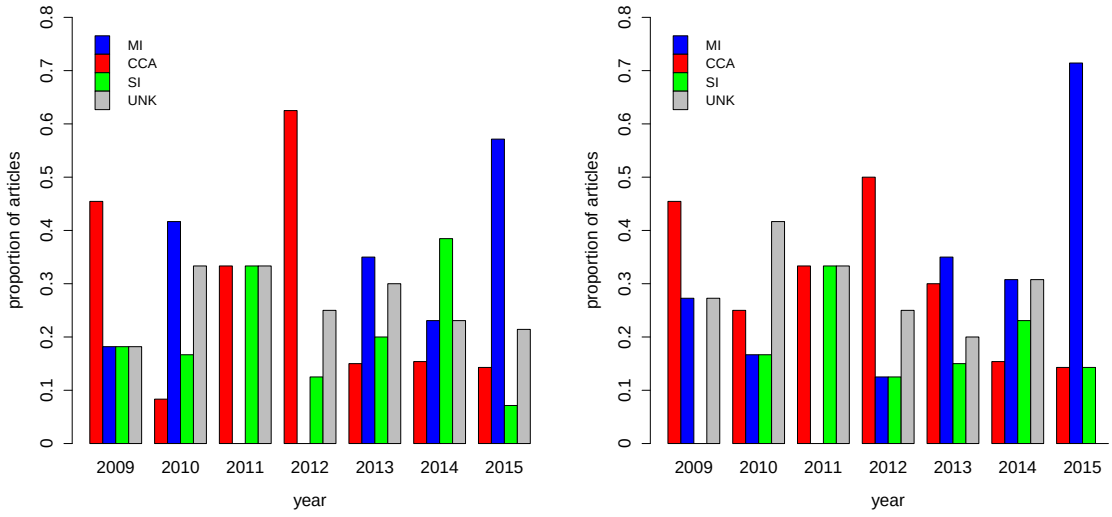
2.4.1 Descriptive review

From the comparison of the base-case methods used for the costs and effects between 2009 and 2015 (Figure 2.2), a marked reduction is observed in the number of methods not clearly described for the effects, compared to those for the costs. A possible reason for this is that, while clinical effectiveness measures are often collected through self-reported questionnaires, which are naturally prone to missingness, cost measures rely more on clinical patient files which may ensure a higher completeness rate. It was not possible to confirm this interpretation in the reviewed studies due to the high proportions of articles not clearly reporting the missing rates in both 2003-2009 and 2009-2015 periods, for effects ($\approx 45\%$ and $\approx 38\%$) and costs ($\approx 50\%$ and $\approx 62\%$). In addition, clinical outcomes are almost invariably the main objective of RCTs and are usually subject to more advanced and standardised analyses. Arguably, costs are often considered as an add-on to the standard trial: for instance, sample size calculations are almost always performed with the effectiveness measure as the only outcome of interest. Consequently, missing data methods are less frequently well thought through for the analysis of the costs. However, this situation is likely to change as cost data from different perspectives (e.g. caregivers, patients, society, etc.) are being increasingly used in trials, leading to the more frequent adoption of self-report cost data which may start to exhibit similar missingness characteristics to effect data.

The review identified only a few articles using more than one alternative method (Figure 2.3). In addition, these analyses are typically conducted without any clear justification about their underlying missing data assumptions and may therefore not provide a concrete assessment of the impact of missingness uncertainty. This situation indicates a gap in the literature associated with an under-implementation of sensitivity analysis, which may significantly affect the whole decision-making process outcome, under the perspective of a body who is responsible for providing rec-

ommendations about the implementation of alternative interventions for health care matters.

To assess whether the number and type of missing data methods used in the reviewed papers vary across the years in the most recent period (2009-2015), an additional analysis by year of publication is performed. The results from this analysis for the different types of base-case methods, grouped into four classes (UNK, SI, CCA and MI) are displayed in Figure 2.5. The proportions of



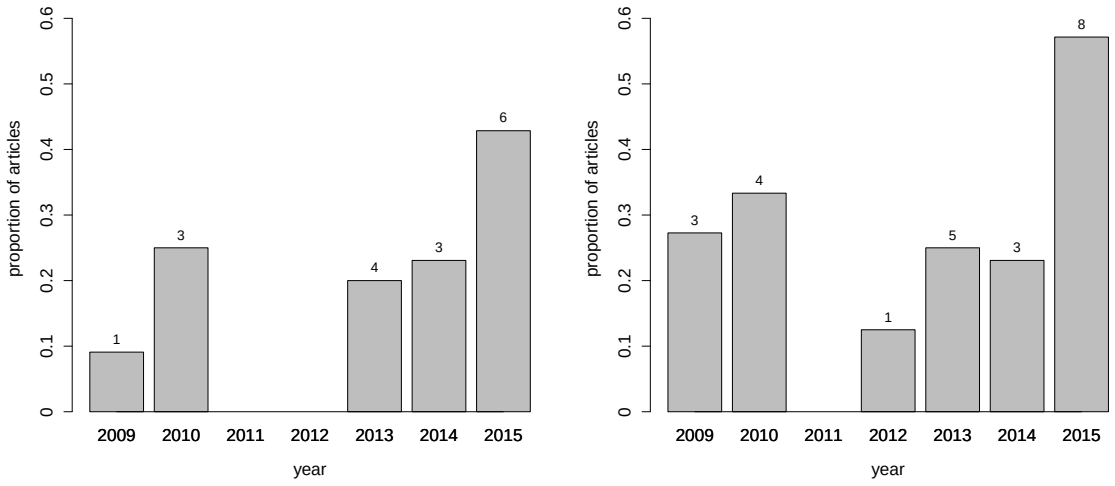
(a) Missing costs (2009-2015)

(b) Missing effects (2009-2015)

Figure 2.5: Proportions of base-case methods used to handle missing cost and effect data between 2009-2015, presented by type of method and year of publication. Legend: Multiple Imputation (MI), Complete Case Analysis (CCA), Single Imputation (SI) and unspecified method (UNK).

the different types of base-case methods show a similar pattern over the years between missing costs (panel a) and effects (panel b). The proportions of unclear (UNK) and single imputation (SI) methods remain roughly constant over time, with an exception in 2015 for the missing effects for which the proportion of UNK methods disappear while SI methods are substantially reduced (especially for the missing costs). The proportions of CCA methods are generally the highest until 2012 for both outcomes (with the exception of the missing costs in 2010), after which they are halved while those of MI raise and achieve their highest in 2015 (above 0.5 for both outcomes). The results suggest that up to 2012 CCA was the reference base-case method, but is then replaced by MI, whose frequency of implementation has rapidly increased and achieved its highest in 2015 for both missing costs and effects. Figure 2.6 shows the results for the proportions of any robustness methods used in the reviewed articles by year of publication. No robustness method was found for the articles in 2011 (for both costs and effects) and 2012 (only for the costs), possibly due to the low number of articles that were reviewed from those years (3 and 8, respectively). Across almost all the years the proportion of robustness methods is low and generally below 0.3. In 2015, however, the proportions show an marked increase for both missing costs (6/14) and, especially, for missing effects (8/14). These results suggest a general lack of implementation of robustness analyses for most of the articles until 2014, with a an uptake of this technique for the articles in 2015.

Limiting the assessment of missingness assumptions to a single case is unlikely to provide a reliable picture of the underlying mechanism. This, in turn, may have a significant impact on the CEA and mislead its conclusions, suggesting the implementation of non-cost-effective treatments. Robustness analyses assess the sensitivity of the results to alternative missing data methods but do not justify the choice of these methods and their underlying assumptions about missingness which may therefore be inappropriate in the specific context analysed. By contrast,



(a) Missing costs (2009-2015)

(b) Missing effects (2009-2015)

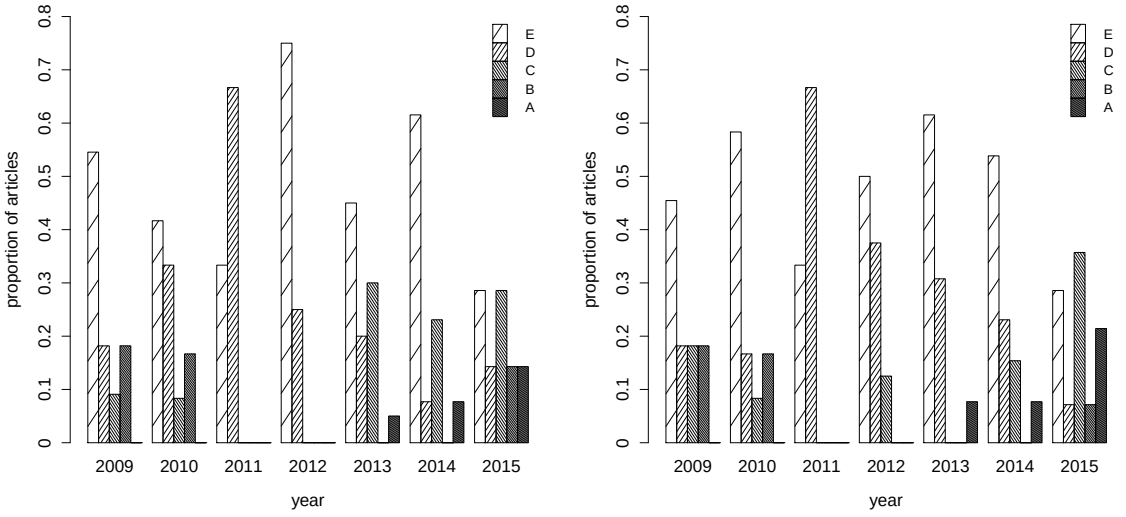
Figure 2.6: Proportions of articles using robustness methods to handle missing cost and effect data between 2009-2015 by year of publications. The numbers at the top of the bars in the graphs are the absolute numbers of papers performing robustness analysis.

sensitivity analyses, which rely on external information to explore plausible alternative methods and missingness assumptions, represent an important and more appropriate tool to provide realistic assessments of the impact of missing data uncertainty on the final conclusions.

2.4.2 Quality assessment

Generally speaking, most of the reviewed papers achieved an unsatisfactory quality score under the Quality Evaluation Scheme (Figure 2.4). Indeed, the benchmark area on the top-right corner of the graphs is barely reached by less than 7% of the articles, both for cost and effect data. Figure 2.7 shows the results from the application of the Quality Evaluation Scheme to the reviewed studies, disaggregated by year of publication. Overall, the proportions of the studies associated with the lowest category (E) prevails in the majority of the years, with a similar pattern over time between missing costs and effects. All the articles that are associated with the top category (A) belong to the period 2013-2015, with the highest proportions of articles falling in this category being observed in 2015 for both outcomes. The opportunity of reaching such a target might be precluded by the choice of the method adopted, which may not be able to support less restrictive assumptions about missingness, even when this would be desirable. As a result, when simple methods cannot be fully justified it is necessary to replace them with more flexible ones that can relax assumptions and incorporate more alternatives. In settings such as those involving MNAR, sensitivity analysis might represent the only possible approach to account for the uncertainty due to the missingness in a principled way. However, due to the lack of studies either performing a sensitivity analysis or providing high quality scores on the assumptions, missingness is not adequately addressed in most studies. This could have the serious consequence of imposing too restrictive assumptions about missingness and affect the outcome of decision making.

Table 2.1 provides a convenient tool to check that all relevant information on missing data is taken into account in determining the assumptions in the analysis. All the key considerations should be fully satisfied, if possible. These criteria are grouped by type of components in the analysis and summarise previously published missing data recommendations from various settings, drawing them together within a general, simple and easy-to-read checklist table.



(a) Missing costs (2009-2015)

(b) Missing effects (2009-2015)

Figure 2.7: Proportions of articles across the categories of the quality evaluation scheme (from E to A) for both missing cost and effect data between 2009-2015, divided by year of publication.

The assignment of scores, based on the weighting of the components, is an appealing feature of the Quality Evaluation Scheme which allows a ranking of the studies based on the quality of the analysis in terms of missing data methods used and information reported. The robustness of the scheme to two different weight choices has been tested on the review's articles to assess the sensitivity of the score assignment and articles' classification across the quality scores (A-E). In one version all components (Description: Method: Limitations) are given exactly the same weight using a ratio 1:1:1, while in the other version the difference in weights between the components is increased using a ratio 6:3:1. The results of these comparisons in terms of weight allocation, scoring system, and articles' classification are provided in Appendix C.1.1. In general, the results do not show any substantial changes in the classification of the articles between the three versions and suggest a general robustness of the scheme to alternative weight allocations.

Finally, the grouping of the studies into ordered categories (Figure 2.1) is potentially a valuable tool for meta-analysis. The proposed quality evaluation scheme could be used by analysts to assign scores and grade individual studies based on their overall quality level in terms of missingness handling. These grades could then be taken into account in assigning different weights to the individual studies within the meta-analysis framework in order to reflect a different degree of confidence in their results.

2.5 Conclusions

Our review shows, over time, a significant change from more to less restrictive methods in terms of the assumptions on the missingness mechanism. This is an encouraging movement towards a more suitable and careful missing data analysis. The results from the disaggregated analysis by year of publication in the later period (2009-2015) indicates the rise of a better and more transparent approach to handle missingness in the latest years of the review, especially in 2015. In particular, compared to the previous years, the articles reviewed from 2015 are associated with a higher proportion of MI methods used in the base-case analysis, a substantial increase in the number of robustness methods implemented, and a better quality score assignment.

Nevertheless, improvements are still needed as, overall, only a small number of articles pro-

vide transparent information about the missing data and almost no study performs a sensitivity analysis. These failings are probably due to the fact that the implications of using methods that do not handle missingness in a principled way are not well-known among practitioners. In addition, the choice of the missing data methods may also be guided by their ease of implementation in standard software packages rather than methodological reasons. This is a potentially serious issue for bodies such as the NICE who use these evaluations in their decision making, thus possibly leading to incorrect policy decisions about the cost-effectiveness of new treatment options.

The Quality Evaluation Scheme represents a valuable tool to improve missing data handling. By carefully thinking about each component in the analysis we are forced to explicitly consider all the assumptions we make about missingness and assess the impact of their variation on final conclusions. The main advantage is a more comparable formalisation of the uncertainty as well as a better indication of possible issues in assessing the cost-effectiveness of new treatments.

In the next chapter, we present the CEA data from the two case studies analysed in this thesis and describe the limitations of the "standard" statistical approach used by practitioners in routine analyses. In the rest of the thesis we then present our missing data strategy, which overcomes the limitations of the standard approach and can be implemented in a relatively easy way using freely available software.

Key points of this chapter:

- The Quality Evaluation Scheme assembles guidelines of best practice about missing data handling. These recommendations are summarised in terms of three dimensions associated with the task of analysing missing data: description, method and limitations. For each dimension we define a list of considerations that should be satisfied. According to the number of considerations satisfied and the importance given to each dimension, quantitative scores can be assigned. These scores are then grouped into categories that reflect the amount and quality of information about missingness provided in a given analysis.
- The review summarises the types of methods used to handle missing effect and cost data in trial-based CEAs between 2003-2015. MI seems to have replaced CCA as the reference method in base-case analyses. A general drop in the use of more restrictive SI methods is also observed. However, in many cases, the lack of information about the details of method implementation or the absence of any robustness analysis severely undermines the confidence that can be attributed to the results of the studies.
- The application of the Quality Evaluation Scheme to the reviewed studies reveals a generally poor picture about the quality of missingness handling in CEA. Most of the studies fall in the categories associated with the lowest scores with only a single article performing a sensitivity analysis.

Chapter 3

Case Studies and Standard Approach

After having summarised the current situation in terms of the missing data handling in trial-based CEAs in Chapter 2, in this chapter we present two case studies to illustrate the characteristics of the typical dataset in CEA. The two studies are: the Men's Safer Sex (MenSS; Bailey et al., 2016) and the Positive Behaviour Support (PBS; Hassiotis et al., 2018) trials. We also describe the "standard" statistical approach used in routine analyses and highlight some of its pitfalls and limitations that may lead to some bias in the results or mislead the final cost-effectiveness assessment.

3.1 Case Studies

3.1.1 The MenSS trial

The MenSS trial is a pilot RCT whose purpose was to establish the feasibility and optimal design of a full-scale RCT to test the effect of the Men's Safer Sex (MenSS) intervention website. The MenSS website is an interactive digital intervention which provides information and tailored advice on sexual well-being and barriers to condom use. The website was offered to heterosexual men in the waiting rooms of NHS sexual health clinics, with the aim of increasing condom use and reducing the acquisition of sexually transmitted infections.

The aim of the health economic evaluation was to assess the feasibility of the economic analysis and to inform the methods for the collection of future cost and outcome data alongside a full-scale RCT and to indicate whether or not the intervention seems promising with regard to cost-effectiveness. Given the pilot nature of the trial, results and cost-effectiveness conclusions derived from the analysis of these data should be treated very cautiously and only provide a preliminary economic assessment about the new intervention. Specifically, both the design and the methods used to collect the data could be modified in a future full-scale trial, which has implications on the economic assessment. For example, the length of the follow-up period of one year in the pilot trial may be considered insufficient to assess the true cost-effectiveness of the new intervention. Nevertheless, the findings from the MenSS trial provide preliminary evidence about the potential cost-effectiveness of the website intervention, which could be used to guide further research.

Individuals enrolled in the study ($n = 159$) were men aged 16 or over who reported female sexual partners and recent unprotected sex or suspected acute sexually transmitted infections. Participants were randomised to receive the MenSS website plus usual clinic care (reference

intervention, $n_1 = 84$), or usual clinic care only (comparator, $n_2 = 75$). Sexual health related utilities u_{ij} were calculated for all participants at baseline ($j = 0$) and at 3, 6 and 12 months follow-ups ($j = 1, 2, 3$) using the EQ-5D 3 level in combination with the preference-based tariff system of Dolan and Gutex (1995), which provides the utility scores for each possible health state from the EQ-5D based on the implementation of the time-trade off algorithm (Section 1.2) to a representative sample of the general UK population. Sexual health related costs c_{ij} (in £) were collected for each participant via responses to resource use questionnaires at the three follow-ups (not collected at $j = 0$).

Table 3.1 shows the number and proportion of *available case* (AC) at each time point in the trial, i.e. the set of cases comprising the *complete cases* (CC) and any other observed value at that time point, for both utilities and costs by treatment group. The number of CC in each treatment $t = 1, 2$ is also reported at the bottom of the table. Baseline costs were not collected,

Time	**Type of outcome**	**Control** ($n_1 = 75$)	**Intervention** ($n_2 = 84$)
		observed (%)	observed (%)
$j = 0$	utilities	72 (96%)	72 (86%)
$j = 1$	utilities and costs	34 (45%)	23 (27%)
$j = 2$	utilities and costs	35 (47%)	23 (27%)
$j = 3$	utilities and costs	43 (57%)	36 (43%)
complete cases	utilities and costs	$n_1^{cc} = 27$ (36%)	$n_2^{cc} = 19$ (23%)

Table 3.1: Number and proportion of observed cases at each time point for the utility and cost data (self-recorded questionnaires), presented by trial group (baseline data only related to the utilities). The number of individuals having valid data at each time point (complete cases) is also reported at the bottom of the table. Over the trial period both drop-out and intermittent missingness occur; at each time point only unit-nonresponse is observed.

while across the other follow-ups utility and cost data were either both observed or both missing. The average proportion of missing responses across follow-ups is 50% for the control $(t = 1)$ and 32% for the intervention $(t = 2)$. The proportions of observed data are systematically lower in the intervention compared with the control group at each time point. This pattern could indicate the existence of an informative missingness mechanism in which individuals in the intervention are less likely to report their utility and cost data with respect to those in the control. Specifically, given the relatively high QALYs values associated with the individuals in the trial (Figure 3.2), it is plausible that individuals in the intervention group would experience health states closer to full-health compared with those in the control and decide not to report these. The sensitivity of the cost-effectiveness conclusions from the trial to alternative informative missingness assumptions is assessed in Section 5.3.1.

Baseline utilities have the largest number of observed values. Specifically, the proportion of AC for u_{i0} is 96% ($n_1^{ac} = 72$) in the control and 86% ($n_2^{ac} = 72$) in the intervention. Figure 3.1 compares the empirical distributions of CC and AC baseline utilities in the MenSS trial.

The empirical distribution of the CC is systematically different from that of the AC. Specifically, the utilities show mean differences of -0.038 in the control (panel a) and of 0.037 in the intervention (panel b) group between the AC and the CC. Although these changes are relatively small, when adjusting the estimates of the QALYs for the potential imbalance between treatments in the baseline utilities (e.g. using regression methods – see Section 3.3), their magnitude is large enough to have substantial implications on the estimation of the mean QALYs differentials. This can be seen by looking at the distribution and mean values of the QALYs (e_{it}) and total cost (c_{it}) variables in the MenSS trial, which are shown in Figure 3.2. Since the time horizon of the trial is 1 year, the range of the QALYs coincide with that of the utilities and is defined between $[-0.594; 1]$, so that variations in the two variables are directly comparable. The empirical QALY means in the control (0.904) and intervention (0.902) group of the trial are almost identical with a mean differ-

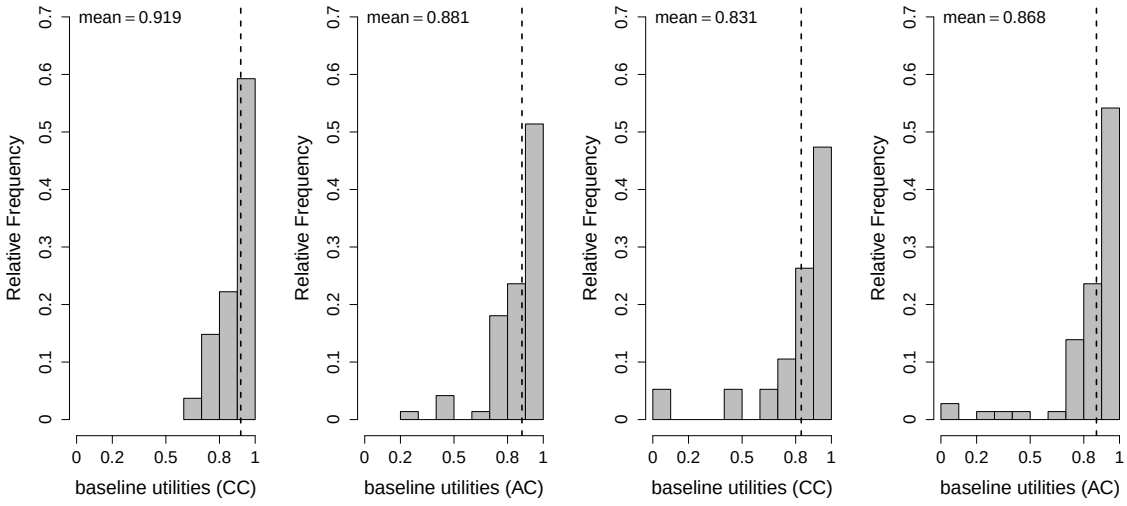


Figure 3.1: Empirical distributions for the baseline utilities in the control (panel a) and intervention (panel b) of the MenSS trial computed either on the AC or CC. A dashed line is drawn in correspondence of the mean for each variable and the value reported in the plots.

ential of -0.002. Therefore, when the potential imbalance in the baseline utilities is accounted for in the estimation of mean QALYs differential, even small variations in the mean baseline utilities between the CC and AC can lead to QALY differentials of different sign with opposite implications in terms of cost-effectiveness conclusions.

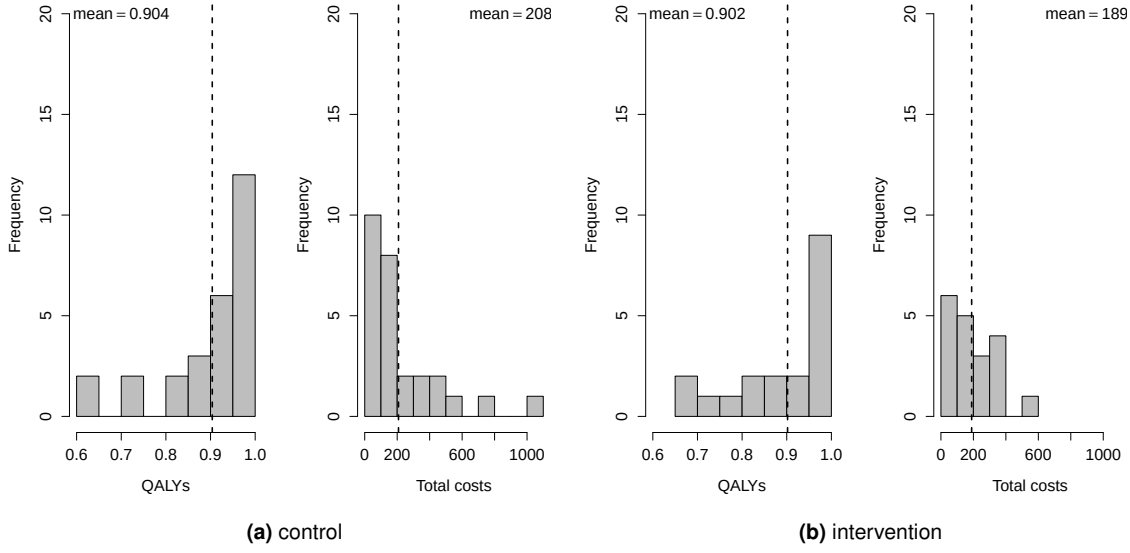


Figure 3.2: QALYs and total cost distributions for the control (panel a) and intervention (panel b) groups in the MenSS trial. A dashed line is drawn in correspondence of the mean for each variable and the value reported in the plots. Costs are expressed in £.

Both outcome data shown in Figure 3.2 clearly present some of the features described in Section 1.2. A relatively high degree of skewness characterises the empirical distributions of QALYs and total costs in both treatment groups. In particular, the substantial proportion of individuals incurring a perfect health status ("structural ones") observed in both the control (33%) and intervention (42%) effectively induces spikes at 1 in the QALYs. Finally, a large proportion of

missingness characterises both variables.

In addition to the outcomes, at baseline, data on three fully-observed covariates were collected for each individual in the study. These are age, ethnicity and employment status, which are briefly summarised in Table 3.2 The main target population in the MenSS trial was represented by young

name	description	details
age	age at baseline	continuous – median = 27, range = (16, 67)
ethn	ethnicity	categorical – 14 levels
empl	employment	categorical – 6 levels

Table 3.2: Description of the available covariates in the MenSS trial

men $\geq$ 16 attending sexual health clinics in England which serve a diverse range of patients in terms of age, socio-economic status and ethnicity but who were all at risk of sexually transmitted infections. However, no upper bound for the age of the participants was imposed in the inclusion criteria defined in the protocol of the trial (Bailey et al., 2016) and one middle-aged individual in the control (51 years-old) and two individuals in the intervention (59 and 67-years old) were recruited in the trial.

In the original analysis (Bailey et al., 2016), the economic evaluation was carried out using independent linear regressions for the QALYs and cost data, adjusting for differences in baseline utilities using a regression approach, to derive the mean incremental cost per QALY gained (i.e. ICER) of the website intervention compared with the control over the 12-month duration of the trial. The analysis was repeated 1000 times using bootstrap sampling from the data collected and the results of the bootstrap were used to construct CEACs for a range of values of the willingness to pat for a QALY gained. The results reported indicate a 88% probability that the website is cost-effective compared with current practice at a threshold of $\pounds 20,000$ per QALY gained.

3.1.2 The PBS trial

The PBS trial is a cluster RCT involving community intellectual disability services and service users with mild to severe intellectual disability and challenging behaviour. The purpose of the trial is to evaluate the clinical outcomes of adults with challenging behaviour and intellectual disabilities (e.g. carer-reported ratings of challenging behaviour), who are treated by staff who have received manual-assisted face-to-face staff training in Positive Behaviour Support (PBS). PBS is a multi-component intervention which is designed to foster prosocial actions and enhance the person's quality of life and his/her integration within the local community.

The primary aim of the economic evaluation was to assess the cost-effectiveness of the intervention compared with the control from a health care perspective. Cluster RCTs raise additional challenges for statistical methods for cost-effectiveness analysis, which should address the specific characteristics of the trials. More specifically, in cluster RCTs the unit of randomization is the cluster (the community intellectual disability service in the PBS), not the patient. Thus, individuals within a cluster are likely to be somewhat similar in their characteristics and the care they receive, and therefore, individual outcomes or costs within the same cluster tend to be more homogeneous than those in different clusters. Thus, statistical methods are required to allow for both individual and cluster level correlations between QALYs and costs to avoid biased results.

Participants ($n = 244$) were enrolled from a total of $S = 23$ sites. 12 sites were allocated to staff teams trained to deliver PBS in addition to treatment as usual (reference intervention, $n_2 = 208$), and 11 sites to staff teams trained to deliver treatment as usual alone (comparator, $n_1 = 136$). Utilities u_{ij} were derived for all participants at baseline ($j = 0$) and at 6 and 12 months follow-ups ($j = 1, 2$) using the EQ-5D 3 level in combination with the utility scores associated with each

health state that were obtained from the tariff system of Dolan and Gutex (1995) (Section 1.2). Health related costs c_{ij} (in £) were collected for each participant via family and paid carer records at each time point. Among the objectives of the study there was an economic evaluation aimed at examining the costs and effectiveness of staff training in PBS.

Table 3.3 reports the missingness patterns in each treatment group as well as the number of individuals and the observed mean responses within each pattern. In the table, $\boldsymbol{r} = (\boldsymbol{r}_0; \boldsymbol{r}_1; \boldsymbol{r}_2)$ denotes the observed missingness patterns in the study, where each pattern is associated with different values for the pairs of missing utility and cost indicators $\boldsymbol{r}_j = (r_j^u, r_j^c)$. For example, the pattern $\boldsymbol{r} = \mathbf{1}$ corresponds to the completers pattern.

	control ($t = 1$)							**intervention** ($t = 2$)						
	u_{i0}	c_{i0}	u_{i1}	c_{i1}	u_{i2}	c_{i2}	$n_{\boldsymbol{r}1}$	u_{i0}	c_{i0}	u_{i1}	c_{i1}	u_{i2}	c_{i2}	$n_{\boldsymbol{r}2}$
$\boldsymbol{r} = \mathbf{1}$	1	1	1	1	1	1	108	1	1	1	1	1	1	96
mean	0.678	1546	0.684	1527	0.680	1520		0.726	2818	0.771	2833	0.759	2878	
$\boldsymbol{r} = 2$	0	1	1	1	1	1	7	0	1	1	1	1	1	5
mean	–	1310	0.704	1440	0.644	1858		–	2573	0.780	2939	0.849	2113	
$\boldsymbol{r} = 3$	1	1	0	1	1	1	4	1	1	0	1	1	1	1
mean	0.709	1620	–	1087	0.737	851		0.467	9649	–	4828	0.259	4930	
$\boldsymbol{r} = 4$	1	1	1	1	0	1	2	1	1	1	1	0	1	1
mean	0.564	640	0.648	512	–	286		0.817	3788	0.884	0	–	0	
$\boldsymbol{r} = 5$	1	1	0	0	1	1	4	1	1	0	0	1	1	1
mean	0.716	2834	–	–	0.634	679		0.501	3608	–	–	0.872	4781	
$\boldsymbol{r} = 6$	1	1	0	0	0	0	4	1	1	0	0	0	0	4
mean	0.434	1528	–	–	–	–		0.760	3086	–	–	–	–	
$\boldsymbol{r} = 7$	0	1	0	1	1	1	2	0	1	0	1	1	1	0
mean	–	595	–	397	0.483	69		–	–	–	–	–	–	
$\boldsymbol{r} = 8$	1	1	1	1	0	0	2	1	1	1	1	0	0	0
mean	0.743	1434	0.705	1606	–	–		–	–	–	–	–	–	
$\boldsymbol{r} = 9$	1	1	0	1	0	1	3	1	1	0	1	0	1	0
mean	0.726	1510	–	432	–	976		–	–	–	–	–	–	

Table 3.3: Missingness patterns for the outcome $\boldsymbol{y}_{ij} = (u_{ij}, c_{ij})$ in the PBS study. For each pattern and treatment group, the number of subjects ($n_{\boldsymbol{r}t}$) and the observed mean responses at each time $j = 0, 1, 2$ are reported. We denote the absence of response values or individuals within each pattern with –.

The number of observed patterns is relatively small in both the control ($R_1 = 9$) and intervention ($R_2 = 6$) groups. Missingness is mostly nonmonotone and, with the exception of the completers ($\boldsymbol{r} = \mathbf{1}$), the patterns are quite sparse. At each time point, when the costs are missing, the utilities are always observed; but, when the utilities are missing, the costs may or may not be observed. Table 3.4 shows the number and proportion of AC at each time point in the trial for both utility and cost data by treatment group. The number of CC is also reported at the bottom of the table.

Time	**Control** (n_1=136)		**Intervention** (n_2=108)	
	observed (%)		observed (%)	
	utilities	costs	utilities	costs
$j = 0$	127 (93%)	136 (100%)	103 (95%)	108 (100%)
$j = 1$	119 (86%)	128 (94%)	102 (94%)	103 (95%)
$j = 2$	125 (92%)	130 (96%)	103 (95%)	104 (96%)
complete cases	$n_1^{cc} = 108$ (79%)		$n_2^{cc} = 96$ (89%)	

Table 3.4: Number and proportion of observed cases at each time point for the utility (self-recorded questionnaires) and cost (clinic records) PBS data, presented by trial group. The number of individuals having valid data at each time point (complete cases) is also reported at the bottom of the table. Over the trial period both drop-out and intermittent missingness occur.

The average proportions of missing utilities and costs across the follow-ups are 11% and 5% for the control $(t = 1)$ and 5% and 6% for the intervention $(t = 2)$ group respectively. While the proportions of observed costs between the two treatment groups are roughly equal at each

time point, the proportions of observed utilities are systematically lower in the control compared with the intervention group. This may be due an informative missingness mechanism where the individuals in the control are associated with worse health states with respect to those in the intervention, which also seems to be supported by the generally lower mean QALY value in the control compared with the intervention (Figure 3.4). The impact of alternative informative missingness assumptions on the cost-effectiveness conclusions of the PBS trial is assessed in Section 6.3.2.

Baseline costs are the only fully observed variables, while for the baseline utilities the proportion of AC is 93% ($n_1^{ac} = 127$) in the control and 95% ($n_2^{ac} = 103$) in the intervention. Figure 3.3 compares the empirical distributions of CC and AC baseline utilities and costs in the PBS trial.

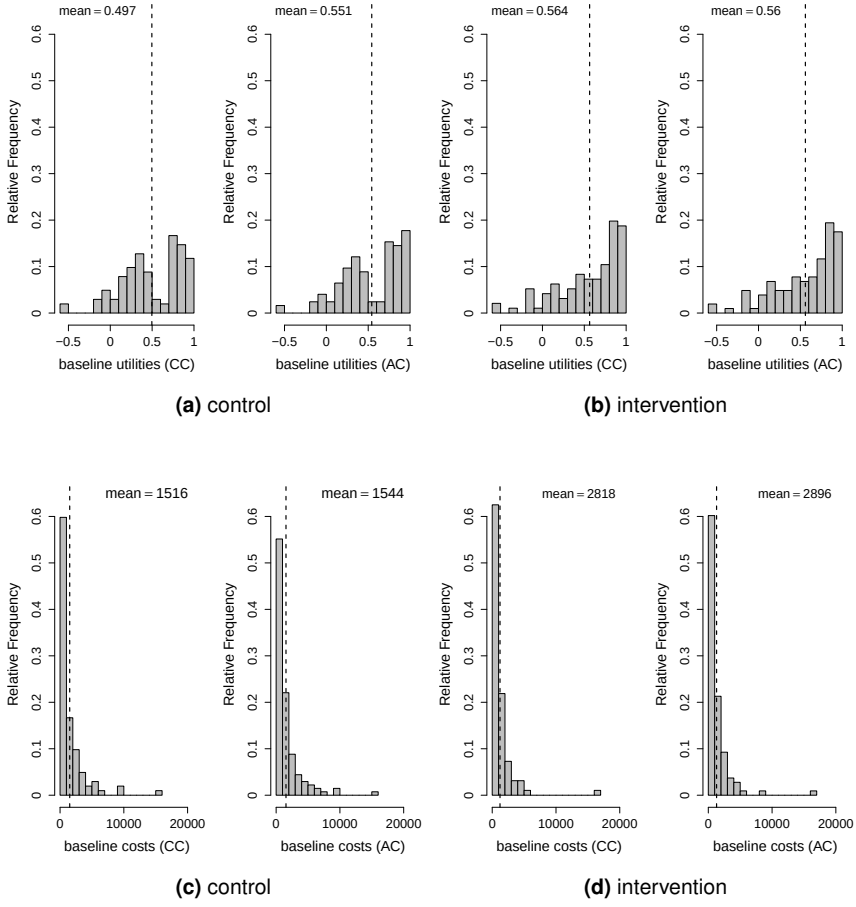
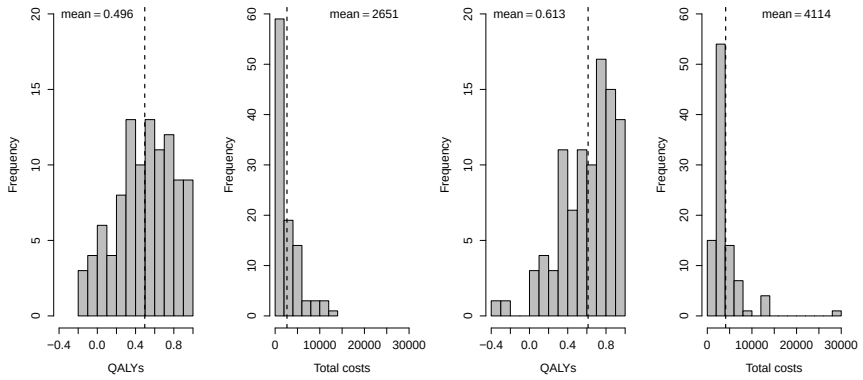


Figure 3.3: Empirical distributions for the baseline utilities and costs in the control (panels a and c) and intervention (panels b and d) group of the PBS trial computed either on the AC or CC. A dashed line is drawn in correspondence of the mean for each variable and the value reported in the plots.

The baseline utilities show a mean variation of 0.054 and 0.004 between the AC and the CC in the control and intervention group, respectively (Figure 3.3, panel a-b). These differences are relatively small in terms of both baseline utility scores and QALYs, which share the same range of $[-0.594; 1]$ due to the 1 year time horizon of the PBS trial. However, compared with the MenSS trial, the magnitude of these changes is less likely to have a substantial impact on the estimation of the adjusted mean QALYs differentials. This is due to the fact that the empirical mean QALYs differential between the two groups in the PBS trial is equal to 0.117, which is substantially larger compared with the differences between the AC and CC in the baseline utilities. Similar considerations hold for the baseline cost variables which show variations of £28 and £78 between the AC and the CC in the control and intervention (Figure 3.3, panels c-d), against a mean total

cost differential of $£1463$ between the two groups. The distribution and mean values of the QALYs (e_{it}) and total cost (c_{it}) variables in both groups of the PBS trial are shown in Figure 3.4.



(a) control **(b)** intervention

Figure 3.4: QALYs and total cost distributions for the control (panel a) and intervention (panel b) groups in the PBS trial. A dashed line is drawn in correspondence of the mean for each variable and the value reported in the plots. Costs are expressed in £.

Few individuals are associated with a perfect health status in both treatment groups, i.e. unit QALYs, while cost distributions show a relatively high degree of skewness, especially in the control. The dataset also includes a few fully-observed covariates that were collected at baseline. These variables are summarised in Table 3.5.

name	description	details
age	age at baseline	continuous – median = 37, range = (18, 76)
ethn	ethnicity	categorical – 6 levels
liv	living condition	categorical – 3 levels (1 = with others, 2 = alone, 3 = with parents)
dis	level of disability	categorical – 3 levels (1 = mild, 2 = moderate, 3 = severe)
carer	type of carer	categorical – 2 levels (1 = family, 2 = paid)
gender	sex	categorical – 2 levels (1 = male, 2 = female)
marital	marital status	categorical – 3 levels (1 = single, 2 = married, 3 = widow)
site	site	categorical – 23 levels

Table 3.5: Description of the available covariates in the PBS trial.

The economic evaluation in the original analysis (Hassiotis et al., 2018) was carried out using independent linear regressions for the QALYs and cost data, adjusting for differences in baseline utilities/costs using a regression approach and accounting for clustering using site-specific intercept terms. The mean incremental QALYs, cost and ICER of the new intervention compared with the control were then computed over the 12-month duration of the trial. Bootstrapping methods were used to construct CEACs for a range of values of the willingness to pat for a QALY gained. The results reported indicate a 60% probability that the website is cost-effective compared with current practice at a threshold of $£20,000$ per QALY gained.

3.2 Standard Approach to Economic Evaluation

Trial-based routine analyses typically rely on a "standard" approach to perform the economic evaluation and assess the cost-effectiveness of the treatment options being compared. The methods that belong to this class are identified by reviewing the methods used in the trial-based economic evaluations for the articles included in the literature review in Chapter 2 and the approach used in the original analyses of the MenSS and PBS studies (see Section 3.1.1 and Section 3.1.2).

The statistical approaches used in the primary CEA analysis (and their frequency of use) across the 81 studies included in the review between 2009-2015 are summarised in Table C.10, which is provided in Appendix C.1.3. In particular, across the 81 articles included in the review for the period 2009-2015, the majority of studies (53) used independent linear regression methods to derive the mean incremental cost and effectiveness parameters between the intervention groups (30 of these used bootstrap methods to account for parameter uncertainty). Of the remaining 28 studies,15 do not clearly report the methods used in the economic analysis, while the use of other methods is uniformly distributed across 13 studies.

The majority of the reviewed studies used a frequentist statistical framework, where individual QALYs and total costs are modelled independently (often implicitly) assuming normality and linearity, and by controlling for baseline values (Manca et al., 2005; Van Asselt et al., 2009; Hunter et al., 2015). Using e_i and c_i to indicate the QALYs and total costs as before, the model is

$$\begin{aligned} e_i &= \alpha_0 + \alpha_1 u_{i0} + \alpha_2 t_i + \varepsilon_{ei}\,[+\ldots], \qquad & \varepsilon_{ie} &\sim \text{Normal}(0, \sigma_e) \\ c_i &= \beta_0 + \beta_1 c_{i0} + \beta_2 t_i + \varepsilon_{ci}\,[+\ldots], & \varepsilon_{ic} &\sim \text{Normal}(0, \sigma_c), \end{aligned} \tag{3.1}$$

where t_i is a treatment indicator variable, and ε_{ie} and ε_{ic} are independent error terms associated with the QALYs and total costs, respectively. The notation $[+\ldots]$ indicates that other terms (e.g. quantifying the effect of relevant baseline covariates) may or may not be included in the model. Common examples are demographic factors, e.g. age, gender, ethnicity or some other prognostic factors, e.g. stage, size or location of the disease (Hoch et al., 2002; Willan et al., 2004; Vazquez Polo et al., 2005; Nixon and Thompson, 2005). For simplicity, here we assume that only the baseline utilities/costs are included in the regression models.

Once the parameter estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ from Equation 3.1 are derived, e.g. using maximum likelihood estimates, then the population means are estimated as

$$\begin{aligned} \hat{\mu}_{et} &= \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 \overline{u}_0 \\ \hat{\mu}_{ct} &= \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 \overline{c}_0, \end{aligned} \tag{3.2}$$

where $\overline{u}_0$ and $\overline{c}_0$ are the sample means for the baseline utility and cost variables, respectively.

Nonparametric bootstrapping techniques are typically used to quantify the uncertainty around the estimates of Equation 3.2 (Rascati et al., 2001; Thompson and Nixon, 2005). A large number of samples are usually drawn with replacement from the original data and confidence intervals for $(\hat{\mu}_{et}, \hat{\mu}_{ct})$ are then computed based on the distribution of average costs and QALYs across the repeated samples (Briggs et al., 2003; Willan and Briggs, 2006; Ng et al., 2013).

When confronted with a multilevel structure, e.g. when utilities and costs are collected from $s = 1, \ldots, S$ different sites, some structured or random effects are typically incorporated into the regression models in Equation 3.1 to account for clustering (Grieve et al., 2010; Gomes et al., 2012b; Ng et al., 2016). For example, site-specific intercept terms α_{0s} and β_{0s} can be included, which are typically assumed to be normally distributed with zero means and variances estimated from the data, i.e. $\alpha_{0s} \sim \text{Normal}(0, \sigma^2_\alpha)$ and $\beta_{0s} \sim \text{Normal}(0, \sigma^2_\beta)$. In this way, the precision of the site-specific estimates is improved since information is "borrowed" from other sites.

The popularity of the "standard" approach is mostly due to its ease of implementation in popular statistical software, such as `STATA` or `R`, through some built-in functions with only a limited customisation from the user. While this may favour the spread and accessibility of this method to a wider audience, it can also lead to some careless use of the method, especially when the underlying assumptions are likely to be unrealistic in the context analysed.

The original analyses for both the MenSS and PBS studies were performed using the approach described above (Bailey et al., 2016; Hassiotis et al., 2018). In the MenSS trial, since baseline

costs were not collected, the baseline regression adjustment was implemented only for the QALYs. In the PBS trial, the multilevel structure of the data was taken into account by including site-specific random intercepts in both the QALYs and cost regression models.

3.3 Pitfalls and Issues of the Standard Approach

In this section we first describe a pitfall of the "standard" approach which may lead to an incorrect computation of the population mean QALYs and cost parameters. Then, we discuss how the "basic" modelling framework described in Section 3.2 can be extended to deal with the typical complexities of CEA data. These are: correlation, skewness, spikes at the boundaries and missing data.

Complete and Available Cases Adjustment

Since baseline data (u_{i0}, c_{i0}) are often available for all or most of the individuals in the study, the adjusted estimates $(\hat{\mu}_{et}, \hat{\mu}_{ct})$ can be computed using the baseline means $(\overline{u}_0, \overline{c}_0)$ from either the CC or the AC. This introduces an ambiguity in the way the adjustment is performed and can lead to potentially different cost-effectiveness conclusions.

There are no available guidelines in the literature about which approach to use, with standard software implementations often implicitly selecting one of the two. We demonstrate the potential consequences of this pitfall in terms of both inferences and cost-effectiveness conclusions in Chapter 4.

Correlation

The assumption of independence between costs and QALYs is often questionable. This is a recognised problem in the CEA literature and alternative methods to deal with correlation have been proposed (O'Hagan et al., 2006; Nixon and Thompson, 2005; Baio, 2012; Gomes et al., 2012b). One approach is to specify a joint bivariate normal distribution. For example, the models in Equation 3.1 can be extended by allowing the regression residuals to be correlated.

$$\begin{pmatrix} \varepsilon_{ie} \\ \varepsilon_{ic} \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \rho\sigma_e\sigma_c \\ \rho\sigma_c\sigma_e & \sigma_c^2 \end{pmatrix} \right] \tag{3.3}$$

where ρ is a parameter capturing the correlation between the variables. Within a frequentist paradigm, this approach is typically referred to as seemingly unrelated regression (Gomes et al., 2012b). Seemingly unrelated regressions consist in a system of equations that can provide estimates that are more efficient that those obtained, for example, from equation-by-equation ordinary least squares methods, because they recognise the correlation between individual costs and health outcomes in the parameter estimation (Green, 2003).

Sometimes it is more convenient to represent Equation 3.3 using conditional probabilities and factor the joint distribution $p(e, c)$ into the product of a marginal and conditional distribution (Nixon and Thompson, 2005; Baio, 2012). In this case, we can specify the model as

$$\begin{aligned} e_i &= \alpha_0 + \alpha_1 u_{i0} + \alpha_2 t_i + \varepsilon_{ei}\,[+\ldots], & \varepsilon_{ie} &\sim \text{Normal}(0, \sigma_e) \\ c_i &= \beta_0 + \beta_1 c_{i0} + \beta_2 t_i + \beta_3 e_i + \varepsilon_{ci}\,[+\ldots], & \varepsilon_{ic} &\sim \text{Normal}(0, \sigma_c), \end{aligned} \tag{3.4}$$

where the regression coefficient β_3 quantifies the association between costs and QALYs. Equation 3.4 re-expresses the joint model by keeping the regression framework of Equation 3.1 and including e_i into the cost model. Although the models in Equation 3.3 and Equation 3.4 assume

Normal distributions, alternative bivariate specifications could be considered for the joint modelling of the QALYs and costs. For example, Nixon and Thompson (2005) and Thompson and Nixon (2005) use either Normal-Gamma or Normal-LogNormal models, Diaz-Ordaz et al. (2014b) and Ng et al. (2016) consider either Normal-Gamma or Normal-Inverse Gaussian models, while Baio (2014) use Beta-Gamma and Beta-LogNormal models.

When the data have a multilevel structure, e.g. they are collected from $s = 1, \ldots, S$ different centers or clusters, methods that account only for the individual-level correlation between QALYs and costs are not adequate and can lead to biased results (Gomes et al., 2012c,b). Multivariate mixed effects models can overcome this problem by explicitly recognising the clustering in the parameter estimation through cluster-level random effects, either or both for the intercept and slope terms in the regressions in Equation 3.4, while also allowing for the incorporation of cluster and/or individual-level covariates (Nixon and Thompson, 2005). For example, assuming that only random intercepts $(\alpha_{0s}, \beta_{0s})$ are included in the regression models in Equation 3.4, these are typically modelled as

$$\begin{pmatrix} \alpha_{0s} \\ \beta_{0s} \end{pmatrix} \sim \text{Normal} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \psi\sigma_\alpha\sigma_\beta \\ \rho\sigma_\beta\sigma_\alpha & \sigma^2_\beta \end{pmatrix} \right] \tag{3.5}$$

where σ_α and σ_β are the cluster-specific variances for the random effects in the QALYs and cost regressions, while ψ is the parameter that captures the cluster-level correlations. Multivariate mixed effects models acknowledge separately individual and cluster-level correlations between the outcomes through Equation 3.3 and Equation 3.5 and can be implemented either under a frequentist or Bayesian framework, e.g. using maximum likelihood or MCMC methods (Nixon and Thompson, 2005; Gomes et al., 2012b).

Skewness

In general, when the sample size is relatively large or the degree of skewness in the data is not high, inferences about the population means are not strongly affected even if the model does not explicitly account for skewness. In trial-based economic evaluations, however, since both outcomes are typically characterised by a large degree of skewness and relatively small sample sizes, either the the normality assumption encoded in Equation 3.1 or the use of nonparametric bootstrapping (Rascati et al., 2001) is not an optimal modelling approach. More specifically, both methods rely on asymptotic assumptions which require a relatively large sample size to produce correct inferences if the data are highly skewed (O'Hagan and Stevens, 2003; Nixon and Thompson, 2004).

Alternative approaches have been proposed in the literature to account for skewness in both variables, particularly within a Bayesian approach, through the use of more appropriate skewed parametric distributions, such as Gamma, LogNormal or Inverse-Gaussian distributions for the costs Nixon and Thompson (2005); Thompson and Nixon (2005); Diaz-Ordaz et al. (2014b) or Beta distributions for the QALYs (Basu and Manca, 2012). These distributions often allow improvements in the model fit to the observed data and appropriately capture skewness (Baio, 2014).

Spikes at the Boundaries

Fourth, data may exhibit spikes at one or both of the boundaries of the range for the underlying distribution. For example, some patients 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, we may observe individuals who are associated with perfect health, i.e. unit QALY (Basu

and Manca, 2012), which makes it difficult to use a Beta distribution, defined on the open interval $(0, 1)$. A possible solution to avoid this problem is to add/subtract a small constant ϵ from the entire set of observed values, artificially re-scaling it in the desired interval. However, no clear guideline exists about the value to use for ϵ (e.g. $0.1, 0.01, \ldots$) to minimise its influence on the economic results (Cooper et al., 2003; Basu and Manca, 2012). In addition, this approach fails to recognise that the underlying data generating process which characterise the individuals with observed boundary values is different from that of the others, e.g. those associated with a perfect health state may be associated with lower costs with respect to those with an impaired state.

A more efficient solution is the application of *hurdle models* (Ntzoufras, 2009; Mihaylova et al., 2011; Baio, 2014). 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. Hurdle models have been discussed and applied in CEA mainly for handling zero costs (Tooze et al., 2002; Harkanen et al., 2013). In particular, Baio (2014) uses bivariate CEA models that combine hurdle models to handle structural zeros with alternative parametric specifications for the distribution of the QALYs (Beta distributions) and costs (Gamma or LogNormal distributions) to account for skewness in the empirical distributions (the authors also provide an `R` package (Baio, 2013) to implement these methods). We show how to incorporate and extend these approaches within a flexible Bayesian modelling framework that jointly accounts for correlation, skewness and spikes, while also dealing with missing data, in both QALYs and costs in Chapter 5.

Missing Data

Finally, routine analyses are conducted on (e_{it}, c_{it}), which are typically derived using the data from the completers in the study. This, however, is an inefficient approach that discards some observations and is also likely biased unless the completers are a random sample of all study participants. Alternative approaches that account for missing data uncertainty (e.g. MI) have become increasingly popular in trial-based CEAs, as shown by our review in Chapter 2. However, these analyses typically assume that the missing data mechanism is ignorable and do not conduct any sensitivity analysis to alternative missingness assumptions. In Chapter 6, we propose a Bayesian longitudinal model that extends the standard approach using all observed values in the study, accounts for the complexities of the data and facilitates a sensitivity analysis to plausible nonignorable missing data assumptions.

Key points of this chapter:

- Data from two RCTs are analysed in this thesis: The MenSS and PBS trials. The motivating question in both studies is to assess the cost-effectiveness of new treatment interventions compared with the standard of care.
- In both studies utility and cost data are derived from self-reported questionnaires or patient records and show the typical complexities that affect individual-level CEA data: skewed empirical distributions, spikes at one for the utility and at zero for costs, and missingness in both outcomes.
- The original analyses of the studies were performed using a "standard approach" that could lead to biased results. Potential pitfalls and issues involve the ambiguity with respect to the way mean baseline adjustment is implemented, the failure to account for most of the complexities of the data and the assumption that the missingness mechanism is MAR.
- Sensitivity analysis to assess the impact of plausible nonignorable missing data departures is typically not conducted. This undermines the confidence we may place on both the inferences and cost-effectiveness conclusions derived from the analysis.

Chapter 4

A Pitfall in Mean Baseline Utility/Cost Adjustment

Having introduced our data and the "standard" approach in CEA, we use the two studies as motivating examples to demonstrate a drawback in the implementation of mean baseline utility/cost regression adjustment. Specifically, when dealing with partially-observed data, mean baseline values can be computed in alternative ways, which have distinct implications in terms of missing data assumptions. This, in turn, can lead to different estimates and impact the cost-effectiveness assessment. A synthesised version of this chapter in the form of a research article has been submitted for publication in *Health Economics*.

4.1 Complete versus Available Cases

Routine CEAs conducted alongside RCTs are typically performed using the statistical approach described in Section 3.2, where the target estimates are derived using the baseline variables evaluated at their sample means. When some individuals fail to follow-up and have missing outcome values at some time points, analyses are typically performed using only the completers (CCA). However, since baseline values are often available for all or most of the individuals, the mean values $\bar{u}_0$ and $\bar{c}_0$ can be effectively calculated using either the CC or the AC.

To our knowledge, there is no current guideline about which approach to use and standard software implementations often implicitly select one of the two. This is problematic as analysts may be unaware of how the adjustment is calculated and that alternatives exist. For example, in `STATA`, adjusting for baseline utilities at means can be performed using the following commands:

```
reg QALY treatment_group baseline_utility
margins treatment_group, atmeans
```

The first line specifies a linear regression with `QALY` as the response variable and `treatment_group` and `baseline_utility` as covariates. The second line uses the `margins` function to apply the adjustment by treatment group at the mean of the baseline variable in the model. In `R`, a similar implementation of the regression adjustment can be performed using the commands:

```
reg = lm(QALY ~ treatment_group + baseline_utility)
predict(reg)
```

The first line computes the linear regression, while the second line uses the function `predict` to obtain the adjusted QALYs estimates by treatment group, evaluated at the mean of the baseline variable in the model. By default both functions discard all missing values when fitting the regression and make the adjustment by calculating the mean of the baseline variables on the AC.

However, analysts may be unaware of the fact that if only the subset of the complete cases is retained for the variables, then the adjustment will be computed using the mean of the CC.

The inferences under both approaches are obtained from the information contained in the observed data, i.e. under MAR (Section 1.5.2). However, the appropriateness of the assumptions about the missing data mechanism may differ according to which set of observed values is used in the adjustment. When the additional observations from the AC are systematically different from the CC, then the MAR assumption based on the complete case dataset is untrue and potentially biased values for the population mean QALYs and total costs can be estimated. Thus, the application of mean regression adjustment is characterised by some ambiguity and can lead to different cost-effectiveness conclusions depending on how the mean baseline values are computed. This pitfall holds even when model complexity is increased to account for framework-specific issues such as the multilevel structure in the data or the correlation between outcomes.

4.2 Implementation

We demonstrate the potential issue associated with the implementation of mean baseline adjustment using the data from the MenSS and PBS trials. In both studies, differences between the empirical distributions of the CC and AC for the baseline variables (see Figure 3.1 and Figure 3.3) suggest a potential impact on the final cost-effectiveness assessment depending on whether the mean of the CC or AC is used to derive the target quantities. We compare the two approaches for a set of models of varying complexity. We fit these models using a Bayesian approach, which allows to extend the model structure in a relatively easy way to account for the issues that are relevant to each dataset.

4.2.1 Models

Table 4.1 summarises the different types of models implemented for the economic analysis of the MenSS and PBS studies in this work (Bayesian approach), comparing their structures with those from the models used in the original analyses (frequentist approach).

For the MenSS trial, we compare the model of Bailey et al. (2016) with an analysis based on both the CC and AC for the baseline utility adjustment, assessing the results under independent and joint models. For the PBS study, we compare the model of Hassiotis et al. (2018) with a set of models of increasing complexity. First, the CC and AC baseline utility/cost regression adjustment are considered. The model is then extended to incorporate three additional baseline covariates: living condition, level of disability and type of carer (see Section 3). Finally, the multilevel structure is accounted for by assuming structured regression coefficients for the baseline utility/cost and for the intercept terms (varying-intercept/slope model). For each of these models we compare the impact on the inferences of independence/joint assumptions about the QALYs and cost distributions and the use of the CC or AC in the calculation of the mean baseline utilities/costs.

For all Bayesian models, we specify vague priors on the parameters so that inferences are based on the observed data alone (numerically similar to a frequentist approach). Specifically, we choose normal priors centred at 0 with a standard deviation of 1000 for all the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Uniform distributions between $(-5, 10)$ are assigned to standard deviation parameters on the log scale. Prior sensitivity to alternative specifications for all parameters (e.g. vague Half-Cauchy or Half-Normal priors for the standard deviations on the natural scale) suggested that these choices were adequate in this setting.

Study	model	baseline utilities/costs (CC)	baseline utilities/costs (AC)	baseline covariates	multilevel structure	correlation
MenSS	Bailey et al. (2016)	✗	✓	—	—	✗
	Utility model CC (ind)	✓	✗	—	—	✗
	Utility model CC (joint)	✓	✗	—	—	✓
	Utility model AC (ind)	✗	✓	—	—	✗
	Utility model AC (joint)	✗	✓	—	—	✓
PBS	Hassiotis et al. (2018)	✓	✓	✗	✓	✗
	Utility/cost model CC (ind)	✓	✗	✗	✗	✗
	Utility/cost model CC (joint)	✓	✗	✗	✗	✓
	Utility/cost model AC (ind)	✗	✓	✗	✗	✗
	Utility/cost model AC (joint)	✗	✓	✗	✗	✓
	Covariate model CC (ind)	✓	✗	✓	✗	✗
	Covariate model CC (joint)	✓	✗	✓	✗	✓
	Covariate model AC (ind)	✗	✓	✓	✗	✗
	Covariate model AC (joint)	✗	✓	✓	✗	✓
	Multilevel model CC (ind)	✓	✗	✓	✓	✗
	Multilevel model CC (joint)	✓	✗	✓	✓	✓
	Multilevel model AC (ind)	✗	✓	✓	✓	✗
	Multilevel model AC (joint)	✗	✓	✓	✓	✓

Table 4.1: List of the different models compared in the analysis of the MenSS and PBS data. The models used in the original analyses are indicated with the author's papers, while specific names are assigned to the models implemented in this work according to the different types of complexities that are accounted for. Different symbols are used to indicate whether the corresponding complexity is addressed (✓), ignored (✗) or not relevant (–).

4.2.2 Software

We fitted the models using `JAGS`, which is interfaced with the freely available statistical software `R` using the package `R2jags` (Su and Yajima, 2015). Samples from the posterior distribution of the parameters of interest are then saved to the `R` workspace and used for producing relevant statistics and plots. We ran two chains with 30,000 iterations per chain, using a burn-in of 15,000, for a total sample of 30,000 iterations for posterior inference.

For each variable in the model, convergence of the MCMC sampler was assessed using diagnostic measures, such as the potential scale reduction factor and visual inspection of the density, trace and autocorrelation plots, as well as measures to assess the adequacy of the posterior sample, such as the effective sample size. In particular, all the runs discussed in this chapter were assumed to have converged when the value of the potential scale reduction factor was below 1.05 and the effective sample size was at least 20,000 for all model parameters. The `JAGS` code used to implement the model for the MenSS trial is available in Appendix B.1.

4.3 Results

Results are reported and compared for all models in Table 4.1 in terms of the estimates and credible intervals of key parameters of interest for assessing the cost-effectiveness of the interventions (Section 1.4.1). These include the mean QALYs and total costs in both intervention groups of the trials (μ_{et}, μ_{ct}) and the mean QALY and total cost increments (Δ_e, Δ_c). The latter are also used to compute summary incremental measures, such as the incremental net benefit and the ICER, which provide information about the relative cost-effectiveness of the two interventions compared. In particular, the results associated with the net benefits (evaluated at the NICE's recommended

willingness to pay value of $k = £20,000$) have been included for all the analysis scenarios explored for both the MenSS and PBS studies.

4.3.1 The MenSS study

Table 4.2 shows the posterior results for the two treatment groups of the MenSS trial. In this particular case, small variations in the inferences are observed with respect to using a joint model, and therefore only the results under the latter are presented. Since the time horizon of the trial is

Parameter	Utility model CC (joint)		Utility model AC (joint)	
	Mean	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$				
mean QALY (μ_{e1})	0.904	(0.873;0.935)	0.874	(0.841;0.907)
mean cost (μ_{c1})	207	(104;307)	207	(104;307)
Intervention $(t = 2)$				
mean QALY (μ_{e2})	0.902	(0.859;0.943)	0.915	(0.873;0.959)
mean cost (μ_{c2})	189	(111;266)	189	(111;266)
Incremental				
QALY differential (Δ_e)	-0.002	(-0.054;0.05)	0.041	(-0.013;0.094)
Cost differential (Δ_c)	-18	(-146;110)	-18	(-146;110)
IB (at $k = 20000$)	-23	(-1063;1042)	835	(-241;1928)
ICER	8822		-449	

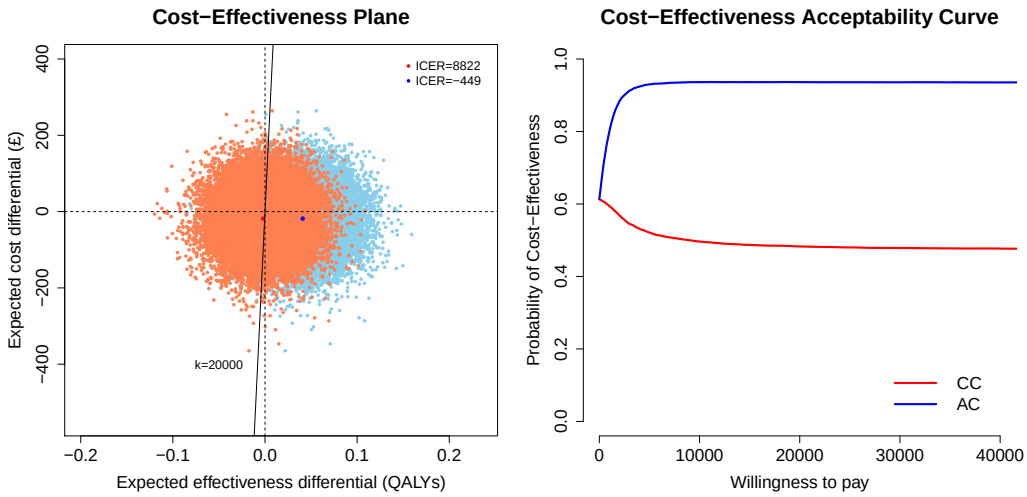
Table 4.2: Posterior means and 95% credible intervals of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the MenSS trial. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Two models are considered: Utility model CC (joint) and Utility model AC (joint). Cost values are expressed in £.

one year, the theoretical range of the QALYs coincide with that of the utility scores, i.e. between -0.594 and 1 (see Section 1.2). Thus, for example, the mean QALYs under both the Utility model CC (joint) and Utility model AC (joint) are approximately 0.9, which is close to the upper bound of full health (corresponding to 1 QALY) and indicates that individuals in both treatment groups have on average a relatively good health status.

Changes in the QALYs estimates are observed between the two types of baseline adjustment. In particular, going from the CC to the AC, the mean QALYs has an average decrease of 0.03 in the control group and an average increase of 0.013 in the intervention group. Although these differences are relatively small in absolute terms, they have a substantial impact on the estimate of the QALYs differential which, at the average value, changes its sign from negative to positive. Because the cost differential is on average negative, this implies that the intervention dominates the control, i.e. lower costs and higher QALYs.

Figure 4.1 shows a graphical representation of the CEP and CEAC based on the posterior samples of the mean parameters from the models shown in Table 4.2. Results related to the CC and AC are respectively indicated with red and blue dots and lines. The graphs provide a clear picture about the impact of the two approaches on the final cost-effectiveness conclusions. At a willingness to pay threshold of $k = £20,000$, the CEP (panel a) shows a much larger proportion of samples that fall in the sustainability area for the model based on the AC (Utility model AC (ind)) compared with the model based on the CC (Utility model CC (joint)). In the CEAC (panel b), the CC are associated with a low probability of cost-effectiveness for almost all k values (red line), while the AC shows a curve which consistently settles at values close to certainty (blue line).

Similar cost-effectiveness conclusions are obtained by considering the Expected Incremental Benefit and Incremental Benefit distribution for the models based on the CC and AC, displayed

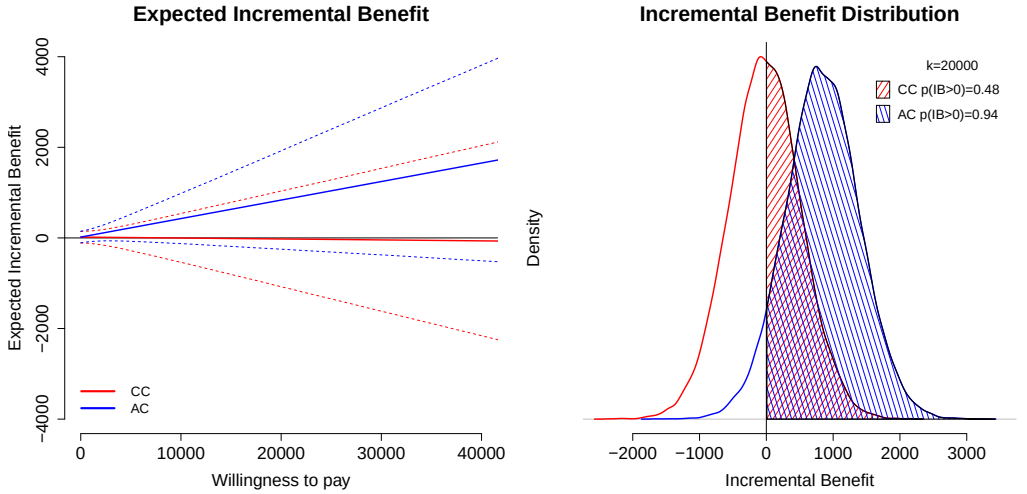


(a) Utility model CC and AC

(b) Utility model CC and AC

Figure 4.1: CEP (panel a) and CEAC (panel b) associated with the Utility model CC (ind) (red dots and lines) and Utility model AC (ind) (blue dots and lines) in the MenSS study.

in Figure 4.2. For most values of k, the EIB (panel a) for the Utility model CC (red solid line) is



(a) Utility model CC and AC

(b) Utility model CC and AC

Figure 4.2: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with the Utility model CC (joint) (red solid and dashed lines) and Utility model AC (joint) (blue solid and dashed lines) in the MenSS study.

approximately 0 with a substantial degree of uncertainty indicated by the inclusion of both negative and positive values within the lower and upper bound estimates (dashed red lines). Conversely, for all values of k the EIB for the Utility model CC (blue solid line) is above 0 (with a positive slope) as well as most of the values falling between the lower and upper bound estimates (dashed blue lines). These results are reflected in the distribution of the IB (panel b), evaluated at $k = £20,000$, with a probability of cost-effectiveness for the Utility model AC (blue shaded area) that is almost twice that of the Utility model CC (red shaded area).

4.3.2 The PBS study

Table 4.3 shows the posterior results for the two treatment groups of the PBS trial. Like the MenSS

Parameter	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
	Utility/cost model CC (joint)		Utility/cost model AC (joint)		Utility/cost model CC (ind)		Utility/cost model AC (ind)	
Control $(t = 1)$								
mean QALY (μ_{e1})	0.491	(0.455;0.527)	0.523	(0.489;0.562)	0.491	(0.455;0.527)	0.526	(0.489;0.562)
Mean cost (μ_{c1})	3076	(2211;3945)	3085	(2234;3966)	3076	(2126;3960)	3090	(2138;3974)
Intervention $(t = 2)$								
mean QALY (μ_{e2})	0.613	(0.571;0.652)	0.611	(0.569;0.650)	0.613	(0.573;0.654)	0.611	(0.571;0.651)
Mean cost (μ_{c2})	4648	(3805;5494)	4700	(3883;5538)	4500	(3653;5365)	4559	(3723;5402)
Incremental								
QALY differential (Δ_e)	0.12	(0.07;0.16)	0.08	(0.03;0.13)	0.12	(0.06;0.17)	0.08	(0.03;0.13)
Cost differential (Δ_c)	1572	(545;2784)	1615	(598;2618)	1423	(132;2640)	1469	(195;2678)
IB (at $k = 20000$)	857	(-740;2526)	82	(-1482;1760)	1007	(-731;2577)	230	(-1448;1841)
ICER	12942		19025		11711		17286	
	Covariate model CC (joint)		Covariate model AC (joint)		Covariate model CC (ind)		Covariate model AC (ind)	
Control $(t = 1)$								
mean QALY (μ_{e1})	0.485	(0.439;0.535)	0.516	(0.467;0.564)	0.485	(0.437;0.534)	0.517	(0.467;0.565)
Mean cost (μ_{c1})	3072	(2227;3955)	3081	(2237;3965)	3081	(2150;3978)	3094	(2164;3991)
Intervention $(t = 2)$								
mean QALY (μ_{e2})	0.580	(0.531;0.628)	0.578	(0.528;0.625)	0.580	(0.531;0.629)	0.578	(0.529;0.627)
Mean cost (μ_{c2})	4649	(3825;5495)	4701	(3887;5525)	4500	(3654;5372)	4559	(3716;5402)
Incremental								
QALY differential (Δ_e)	0.09	(0.05;0.14)	0.06	(0.02;0.11)	0.09	(0.03;0.14)	0.06	(0.01;0.11)
Cost differential (Δ_c)	1577	(567;2595)	1620	(620;2629)	1419	(188;2667)	1465	(241;2696)
IB (at $k = 20000$)	307	(-1251;1945)	-396	(-1954;1228)	463	(-1144;2154)	-243	(-1893;1388)
ICER	16231		25288		15078		23979	
	Multilevel model CC (joint)		Multilevel model AC (joint)		Multilevel model CC (ind)		Multilevel model AC (ind)	
Control $(t = 1)$								
mean QALY (μ_{e1})	0.493	(0.441;0.546)	0.521	(0.469;0.577)	0.493	(0.439;0.546)	0.522	(0.467;0.576)
Mean cost (μ_{c1})	3054	(2122;3997)	3055	(2124;4000)	3055	(2110;3967)	3057	(2098;3958)
Intervention $(t = 2)$								
mean QALY (μ_{e2})	0.585	(0.531;0.637)	0.583	(0.529;0.635)	0.584	(0.532;0.639)	0.583	(0.531;0.637)
Mean cost (μ_{c2})	5449	(4673;6217)	5456	(4693;6221)	5442	(4693;6224)	5449	(4683;6199)
Incremental								
QALY differential (Δ_e)	0.09	(0.03;0.15)	0.06	(0.01;0.12)	0.09	(0.03;0.15)	0.06	(0.01;0.12)
Cost differential (Δ_c)	2395	(1172;3592)	2401	(1155;3567)	2386	(1203;3593)	2392	(1226;3600)
IB (at $k = 20000$)	-551	(-2220;1108)	-1166	(-2876;452)	-557	(-2231;1108)	-1170	(-2837;518)
ICER	25985		38871		26091		39174	

Table 4.3: Posterior means and 95% credible intervals of the mean QALYs and cost parameters for the control ($t = 1$) and the intervention ($t = 2$) group in the PBS trial. Mean QALYs and cost differentials and ICERs are also reported. Twelve different models are considered: Utility/cost model CC/AC (ind), Covariate model CC/AC (ind), Multilevel model CC/AC (ind) and their corresponding joint versions. Cost values are expressed in £.

study, the time horizon in the PBS study is one year, which means that QALYs are defined on the same range of the utility scores, i.e. between -0.594 and 1. Therefore, compared with the MenSS trial, the individuals in the PBS study are on average associated with worse health states (mean QALYs approximately equal to 0.5 compared with 0.9 in the MenSS study).

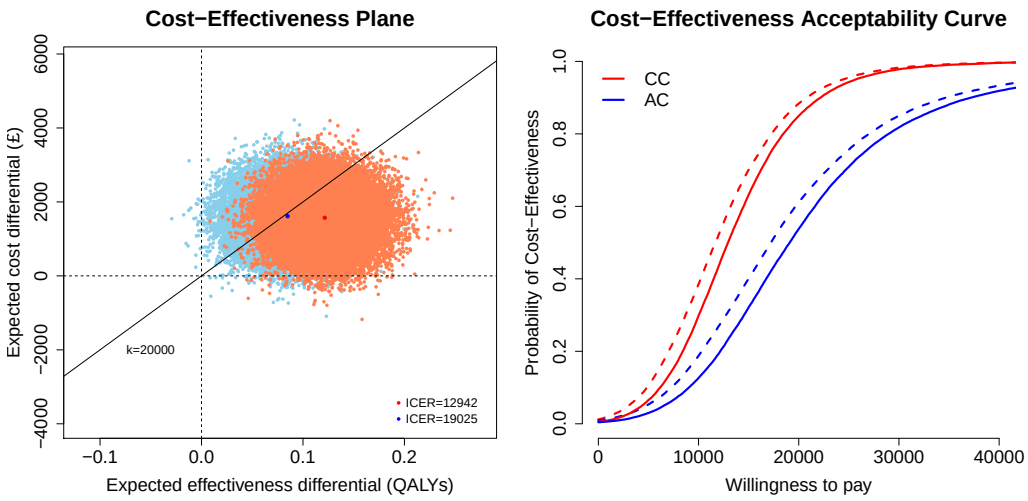
Changing the model structure produces variations in the economic results. More specifically, incorporating the covariates (Covariate model) leads to an average decrease in the mean QALYs of 0.033 in the intervention group with respect to the simpler baseline utility/cost adjustment (Utility/cost model). This discrepancy is similar for both the independent and joint models based on the CC and AC. Given that all the other quantities barely change, this induces a reduction in the QALYs differential and a less favourable cost-effectiveness assessment for the new intervention. When the multilevel structure is accounted for (Multilevel model), a substantial average increase of $£800$ is observed in the mean cost estimates of the intervention group for both the CC and AC versions compared with the other models. Relatively smaller differences are observed in the mean cost incremental estimates between the AC and CC versions of Multilevel model (≈ 7) compared with those of Utility/cost or Covariate models (≈ 50). These are due to the different assumptions about the impact that the baseline costs have on the mean total costs, which is modelled either by accounting for clustering (Multilevel model CC/AC) or ignoring it (Utility/cost model CC/AC or Covariate model CC/AC). Correlation assumptions have a limited impact on mean QALYs estimates, which remain almost unchanged across all models. Conversely, mean cost estimates are lower for the Utility/cost model (ind) and Covariate model (ind) compared to the Utility/cost model (joint) and Covariate model (joint).

Compared with the other models, the results associated with the multilevel models are more robust to correlation assumptions. This may be due to the fact that, in the PBS trial, the magnitude of the correlations between QALYs and costs at the cluster level (Spearman correlations of -0.55 and -0.65 in the control and intervention group) is substantially greater than that at the individual level (Spearman correlations of -0.37 and -0.28 in the control and intervention group). Thus, it is possible that, by accounting for both levels of correlations, the discrepancy between the joint and independent version of the model is different for the multilevel models compared with those that ignore clustering. Finally, the multilevel models are only slightly less sensitive (for the mean cost estimates) to the use of the CC or AC in the baseline adjustment. This is again due to the impact of accounting for both individual and cluster level correlations, even though differences compared with those from the models that ignore clustering are relatively small (mean cost differences of $£52$ and $£59$ between the Covariate CC/AC models against $£7$ and $£7$ between the Multilevel CC/AC models).

Figure 4.3 shows the CEP and CEAC based on the inferences for all the models described in Table 4.3. The graphs distinguish between the results from using the CC (red) and AC (blue), as well as between independence (dashed lines) and joint (solid lines) models. CEPs are reported only for their joint versions for simplicity.

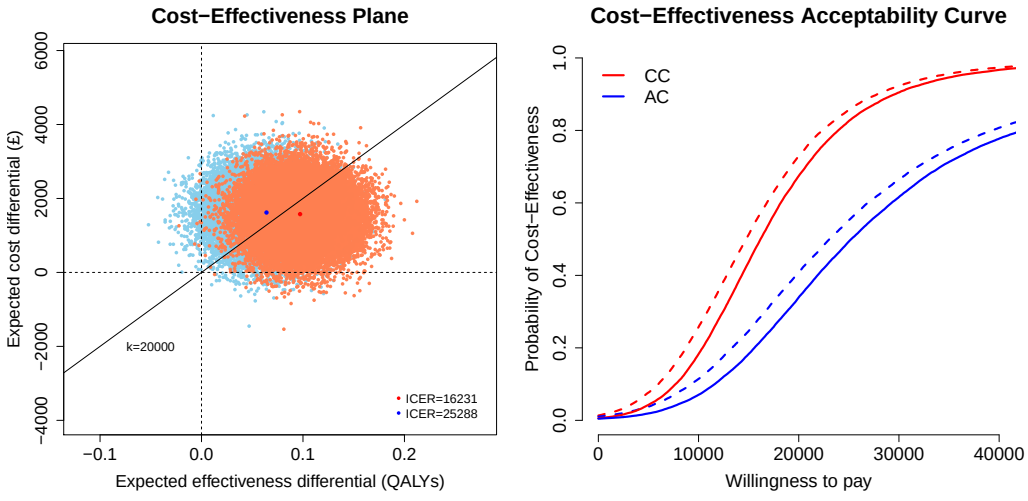
At a willingness to pay threshold of $k = £20,000$ the ICERs for all the models indicate a more cost-effective intervention compared to the control. However, the magnitude of the assessment substantially changes between the models. When the multilevel structure is ignored (panels a-d), inferences are sensitive to correlation assumptions. Both the Utility/cost model (ind) and Covariate model (ind) are characterised by CEACs that are shifted upwards by 10% compared to the Utility/cost model (joint) and Covariate model (joint). A substantial decrease of the curves is observed for the Multilevel model (panel f) compared to the others for values of k below $£20,000$. In this case, however, differences between the independent and joint models almost disappear, which suggests that adjusting for the clustering in the data may also substantially capture the correlation at the individual level.

Finally, we focus on the multilevel models, and look at the differences in terms of Expected



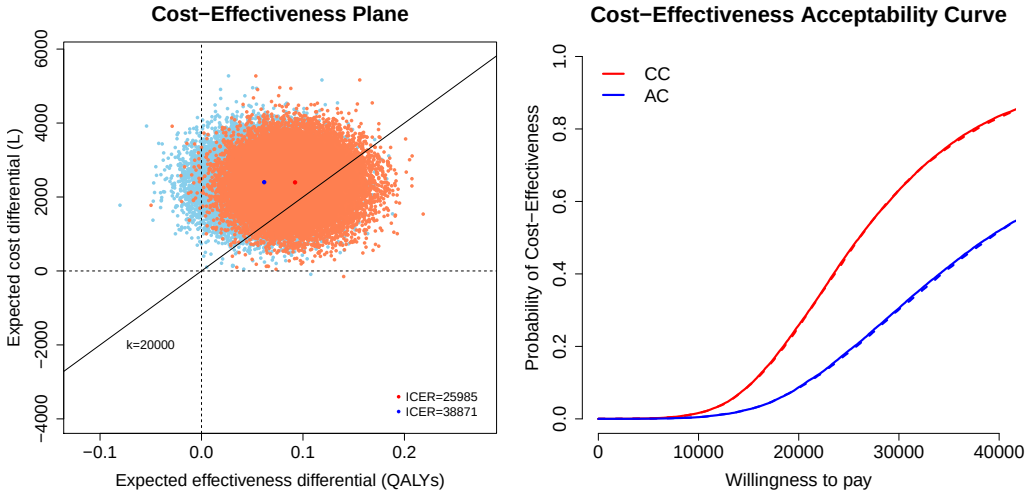
(a) Utility model CC and AC

(b) Utility model CC and AC



(c) Covariate model CC and AC

(d) Covariate model CC and AC

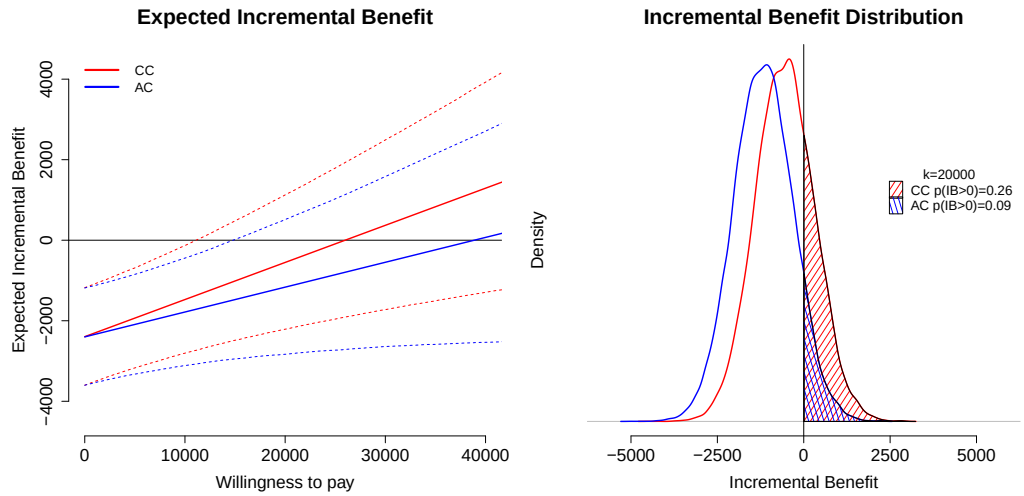


(e) Multilevel model CC and AC

(f) Multilevel model CC and AC

Figure 4.3: CEPs (panels a, c, e) and CEACs (panels b, d, f) associated with the models shown in Table 4.3 for the PBS study. The results based on the CC (red dots and lines) and the AC (blue dots and lines) are indicated with different colours while correlation assumptions are associated with different line types (solid and dashed lines for joint and independence models). In the CEPs only the results from the joint models are shown.

Incremental Benefit and Incremental Benefit distribution between the models estimated using the CC and AC baseline utilities/costs, shown in Figure 4.4. For almost all values of the willingness



(a) Multilevel model CC and AC

(b) Multilevel model CC and AC

Figure 4.4: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with the Multilevel model CC (joint) (red solid and dashed lines) and Multilevel model AC (joint) (blue solid and dashed lines) in the PBS study.

to pay k the EIB for both the CC and AC models is negative (red and blue solid lines, panel a), therefore indicating a poor cost-effectiveness assessment for the intervention compared with the control. The Multilevel model CC is associated with a higher slope of the EIB compared with the Multilevel model AC and a higher proportion of positive EIB values, as also indicated by the lower and upper bound estimates (dashed lines). The incremental benefit distribution, evaluated at $k = £20,000$, is mostly below 0 for both models, with a probability of cost-effectiveness that is approximately 0.3 for Multilevel model CC and 0.1 for Multilevel model AC (shaded red and blue areas).

The results between the CC and AC show discrepancies that persist almost regardless of the complexity of the model. This is indicated by the stable gap between the CEACs associated with the two approaches for all models. More specifically, results based on the CC indicate a more cost-effective intervention compared with those based on the AC for most values of k. Similarly, in the CEPs, the largest proportion of samples falling in the sustainability area is associated with the use of the CC for all models.

4.4 Discussion

Baseline regression adjustment is considered the reference approach to deal with baseline utility/cost imbalance in trial-based CEA. However, when some of the participants fail to follow-up and the economic evaluation is performed using a CCA, this method is subject to the pitfall of whether the mean of the CC or the AC should be computed for the baseline variables in the procedure. Many statistical software packages have built-in functions that by default perform the adjustment using one of two approaches. This is undesirable as analysts may be unaware of the type of adjustment the software implements, which in turn may affect the inferences and the decision-making.

We compared the results from the baseline CC and AC adjustment obtained under the Bayesian framework implemented in this chapter with those obtained from other frequentist methods that allow for either the individual-level correlation (seemingly unrelated regressions – Section 3.3) or both the individual and cluster level correlation (multivariate linear mixed effects models – Section 3.3) for the analysis of the MenSS and PBS trials, respectively. For each treatment group in

both case studies, the estimates for the mean outcome parameters and other incremental quantities (e.g. net benefit) are almost identical between the Bayesian and frequentist methods for both the CC and AC baseline adjustment. Inferences from these methods are also compared in terms of mean QALYs, mean costs and key incremental quantities (e.g. net benefit) with those derived from the original analyses of the two trials. These comparisons are reported in Figure C.2 and Table C.11 (for the MenSS study) and in Figure C.3, Figure C.3 and Table C.12 (for the PBS study) in Appendix C.2.

In the MenSS trial, accounting for the individual-level correlation between QALYs and costs does not lead to sizeable differences in the estimates compared with assuming independence (i.e. with respect to the results from the "standard approach" – only performed using the AC). In the PBS trial, accounting for both levels of correlation (either under a Bayesian or frequentist approach) leads to substantially different estimates for both mean QALYs and costs compared with those from the "standard approach" . These results suggest that, regardless of whether the CC or AC baseline adjustment is used, failure to account for the relevant levels of correlations in the data may lead to incorrect inferences.

While the two baseline adjustment approaches could lead to similar results, our two motivating examples demonstrate that this is not always the case. For the MenSS trial, the cost-effectiveness conclusions derived from the two types of adjustments are completely opposite; for the PBS trial, the differences in the results based on the CC and AC are almost unaltered regardless of the complexity of the model considered. In both studies baseline variables show considerably different mean values between the CC and AC, which lead to different model estimates and, crucially, cost-effectiveness conclusions.

When this occurs, there is a clear indication that the MAR assumption, when using the complete case dataset, is implausible and the results based on CCA are likely to be biased. By including the additional data contained in the AC, rather than just the subset of the CC, may provide sufficient information to eliminate the bias associated with a CCA and obtain valid inferences. However, it is possible that neither the CC nor the AC lead to valid inferences as MAR can never be verified from the data at hand. Thus, it should be more reasonable to avoid CCA and use a method that retains the full sample by imputing the missing values, while also assessing the robustness of the results to different missing data assumptions (including MNAR). Bayesian methods are well-suited to accomplish this task through the specification of suitably-defined (informative) prior distributions, which allow to incorporate external evidence about missingness (e.g. expert opinion) into the analysis. However, the lack of a sensitivity analysis to missingness is only one of the issues that question the validity of the conclusions derived from routine analyses. As we discussed in Section 3.3, both outcome data are typically affected by a series of complexities, such as skewness and spikes, that are ignored by the standard approach and that may bias the results.

In the next chapter we present a flexible Bayesian framework that improves the standard approach by jointly accounting for these complexities, while simultaneously imputing the missing values. In addition, the framework accommodates a sensitivity analysis to MNAR assumptions that can be implemented in a relatively easy way.

Key points of this chapter:

- In routine analyses mean baseline regression adjustment can be implemented either using the mean of the CC or AC for the baseline variables. Standardised functions in popular statistical software (often implicitly) perform the adjustment using one of the two approaches. This is potentially dangerous as analysts may not be aware that the two approaches could lead to different results.
- Using the data from two case studies, we showed that, when there are systematic differences between the CC and AC for the baseline utilities/costs, mean baseline regression adjustment under the two approaches can lead to substantially different cost-effectiveness conclusions. In addition, for both studies, the discrepancy between the inferences associated with the two methods was generally unaffected by the level of complexity of the model, i.e. differences between the CC and AC scenarios remained roughly constant after accounting for correlation (for the MenSS trial) or for correlation, covariates and multilevel structure (for the PBS trial).
- Discrepancies between the empirical distributions of the baseline CC and AC indicate that those individuals whose follow-up values are missing are associated with systematically different baseline utility/cost values compared with the completers. Therefore, even when missingness is MAR, computing the mean utility/costs in the regression adjustment using only the CC is inefficient (i.e. it discards the additional information contained in the AC) and is also more likely to lead to biased results than using the AC because the adjusted QALYs/total cost estimates do not account for the systematic difference between completers and non-completers in the baseline variables. However, in general, both approaches can lead to biased results since assumptions on the missing values can never be checked from the data at hand; thus, methods that explicitly account for missingness uncertainty by retaining the full sample should be preferred to assess the robustness of the results to departures from MAR.

Chapter 5

A General Bayesian Framework for Health Economic Evaluation

In this chapter we propose a unified Bayesian framework that jointly accounts for the typical complexities of the data discussed in Section 3.3 (i.e. correlation, skewness, spikes at the boundaries and missingness), and that can be implemented in a relatively easy way. We demonstrate the benefit of using our approach using the MenSS trial as motivating example (Section 3.1.1). We then show how the framework can be flexibly adapted to accommodate the characteristics of the data in the PBS trial (Section 3.1.2). A simplified version of this chapter in the form of a research article is published in *Statistics in Medicine*.

5.1 Modelling Framework

Consider the usual cross-sectional bivariate outcome formed by the QALYs and total cost variables (e_{it}, c_{it}) calculated for the $i-$th person in group t of the trial. To simplify the notation, unless necessary, we suppress the treatment indicator t. Following Nixon and Thompson (2005), we specify the joint distribution $p(e_i, c_i)$ as

$$p(e_i, c_i) = p(c_i)p(e_i \mid c_i) = p(e_i)p(c_i \mid e_i), \tag{5.1}$$

where, for example, $p(e_i)$ is the *marginal* distribution of the QALYs and $p(c_i \mid e_i)$ is the *conditional* distribution of the costs given the QALYs. Note that, although the two factorisations shown in Equation 5.1 are mathematically equivalent, the choice of which to use has different practical implications. From a statistical point of view, the factorisations require the specifications of different statistical models, e.g. $p(e_i)$ or $p(e_i \mid c_i)$, which may have different approximation errors. From a clinical point of view, the two versions make different assumptions about the casual relationships between the outcomes, i.e. either e_i determines c_i or vice versa. We describe our analysis under the assumption that the costs are determined by the effectiveness measures and therefore we specify the joint distribution $p(e_i, c_i)$ in terms of a marginal distribution for the QALYs and a conditional distribution for the costs.

For each individual we consider a marginal distribution $p(e_i \mid \boldsymbol{\theta}_e)$ indexed by a set of parameters $\boldsymbol{\theta}_e$ comprising a *location* ϕ_{ie} and a set of *ancillary* parameters $\boldsymbol{\psi}_e$ typically including some measure of *marginal* variance σ_e^2. We can model the location parameter using a generalised linear structure, e.g.

$$g_e(\phi_{ie}) = \alpha_0 \; [+ \ldots], \tag{5.2}$$

where α_0 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 Equation 5.2. In the absence of covariates or assuming that a centered version $x_i^* = (x_i - \bar{x})$ is used, the parameter $\mu_e = g_e^{-1}(\alpha_0)$ represents the population average QALYs.

For the costs, we consider a conditional model $p(c_i \mid e_i, \boldsymbol{\theta}_c)$, which explicitly depends on the QALYs, as well as on a set of quantities $\boldsymbol{\theta}_c$, again comprising a location ϕ_{ic} and ancillary parameters $\boldsymbol{\psi}_c$. For example, when normal distributions are assumed for both $p(e_i \mid \boldsymbol{\theta}_e)$ and $p(c_i \mid e_i, \boldsymbol{\theta}_c)$, i.e. bivariate normal on both outcomes, the ancillary parameters $\boldsymbol{\psi}_c$ include a *conditional* variance τ_c^2, which can be expressed as a function of the marginal variance σ_c^2 (Nixon and Thompson, 2005; Baio, 2012). More specifically, the conditional variance of $p(c_i \mid e_i, \boldsymbol{\theta}_c)$ is a function of the marginal effectiveness and cost variances and has the closed form $\tau_c^2 = \sigma_c^2 - \sigma_e^2\beta^2$, where $\beta = \rho\frac{\sigma_c}{\sigma_e}$ and ρ is the parameter capturing the correlation between the variables (see Equation 3.3).

The location can be modelled as a function of the QALYs as

$$g_c(\phi_{ic}) = \beta_0 + \beta_1(e_i - \mu_e)\ [+\ldots]. \tag{5.3}$$

Here, $(e_i - \mu_e)$ is the centered version of the QALYs, while β_1 quantifies the correlation between costs and QALYs. Assuming other covariates are either also centered or absent in Equation 5.3, $\mu_c = g_c^{-1}(\beta_0)$ is the estimated population average cost.

Figure 5.1 shows a graphical representation of the general modelling framework described above. The QALYs and cost distributions are represented in terms of combined "modules" — the

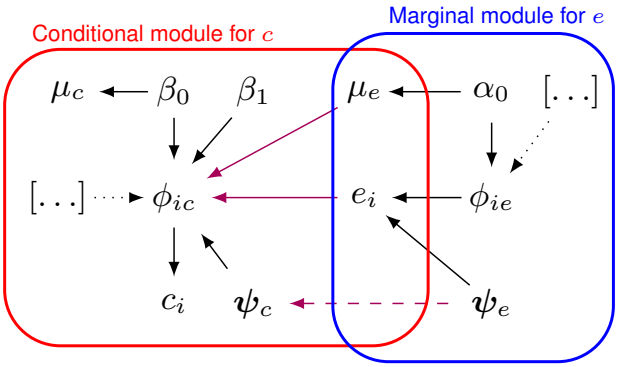


Figure 5.1: Joint distribution $p(e_i, c_i)$, expressed in terms of a marginal distribution for the QALYs and a conditional distribution for the costs, respectively indicated with a solid red and blue line. The solid black and magenta arrows show the dependence relationships between the parameters within 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 QALYs parameters. The dots enclosed in the square brackets indicate the potential inclusion of other covariates at the mean level for both modules.

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.

The proposed framework allows jointly tackling of the different complexities that affect the data in a relatively easy way by means of its modular structure and flexible choice for the distributions of the QALYs and cost variables. Using the MenSS trial as motivating example, we start from the original analysis and expand the model using alternative specifications that progressively account for an increasing number of complexities in the outcomes. We specifically focus on appropriately modelling spikes at the boundary and missingness, as they have substantial implications in terms of inferences and, crucially, cost-effectiveness results.

Three model specifications are considered: 1) Normal marginal for the QALYs and Normal conditional for the costs (which is identical to a Bivariate Normal distribution for the two outcomes); 2) Beta marginal for the QALYs and Gamma conditional for the costs; and 3) Hurdle Model. The

choice of the Beta distribution for modelling the QALYs is considered reasonable given the specific characteristics of the MenSS trial: individuals are associated with relatively high utility and QALY values (always above 0) and the time horizon of the evaluation is 1 year, which ensures that the maximum QALYs value is equal to 1 (see Section 1.2). However, in general, QALYs may be either negative or above 1, for example when the individuals are associated with very poor health states (utilities below 0) or the time horizon of the analysis is longer than 1 year, respectively. In these situations, the choice of the Beta distribution may be inadequate and alternative approaches should be considered; these include the rescaling of the QALYs to ensure they fall in the range $[0, 1]$ (see Section 5.7) or the choice of alternative parametric distributions which explicitly allow for these values, e.g. Normal distributions (Ng et al., 2016).

First, we present each assuming a "complete cases" scenario and then extend the structure to an "all cases" scenario. The latter includes the complete cases and additionally imputes the outcome values for all the remaining individuals in the trial, either under MAR (for all models) or alternative MNAR scenarios (for the Hurdle Model only). In particular, the sensitivity of the results to MNAR is not assessed using a proper nonignorable model (see Section 1.6). This is due to the fact that a cross-sectional framework is not ideal to handle missingness in trial-based CEAs because it does not allow to incorporate the information from all partially-observed utility and cost data into the model, which instead is likely to provide at least some information for imputing the missing values. However, for all analyses in this chapter, we use a cross-sectional framework because it represents the standard framework practitioners are familiar with in routine analyses and which is currently recommended by NICE for trial-based economic evaluations (NICE, 2013). We overcome the limitations associated with this framework for handling missingness in Chapter 6, where a longitudinal modelling framework is used to efficiently incorporate the information from all partially-observed data to impute the missing values and a pattern mixture approach is implemented to conduct sensitivity analysis to MNAR (see Chapter 6).

For the analyses in this chapter, the alternative MNAR scenarios (Section 5.3.1) are chosen according to the specific features of the QALY data in the MenSS study, where some of the individuals in both groups are associated with a perfect health status at one or more time points ($u_{ij} = 1$) but have unobserved measurements at some other time points. Thus, it is plausible to assume that at least some of these individuals are associated with a perfect utility score at all time points, which would result in a perfect health status over the trial period (i.e. $e_i = 1$). We exploit the modelling structure of the Hurdle Model to assess the robustness of the results to differing assumptions about the proportions of these individuals that could be potentially observed in both treatment groups and the impact that these assumptions may have on the final conclusions.

5.2 Complete Cases Scenario

5.2.1 Bivariate Normal

The first specification jointly models the two outcomes assuming bivariate normality, which in our framework is factorised into marginal and conditional Normal distributions for e_i and $c_i \mid e_i$. The model is similar to the "standard" approach described in Section 3.2 with the difference that correlation between the variables is explicitly captured.

In line with the original analysis of the MenSS trial, we adjust for the baseline utilities — using a centered version $u^{\star}_{i0} = u_{i0} - \bar{u}_0$. We model $e_i \mid \boldsymbol{\theta}_e \sim$ Normal(ϕ_{ie}, σ^2_e), using an identity link function for the location parameter

$$g_e(\phi_{ie}) = \phi_{ie} = \alpha_0 + \alpha_1 u^{\star}_{i0}. \tag{5.4}$$

Here, the parameter α_1 quantifies the impact of the centered baseline utilities on the QALYs, while $\mu_e = \alpha_0$ and σ_e^2 represent the marginal (population level) mean and variance, respectively.

As for the costs, we model $c_i \mid e_i, \boldsymbol{\theta}_c \sim \text{Normal}(\phi_{ic}, \tau_c^2)$, where the conditional mean and variance are defined as

$$g_c(\phi_{ic}) = \phi_{ic} = \beta_0 + \beta_1(e_i - \mu_e) \qquad \text{and} \qquad \tau_c^2 = \sigma_c^2 - \sigma_e^2\beta_1^2. \tag{5.5}$$

This conditional specification of the model corresponds to the multivariate structure in Equation 3.3 (Nixon and Thompson, 2005; Baio, 2012), where the parameters indexing the distribution of the costs can be re-expressed as $\beta_1 = \frac{\sigma_c}{\sigma_e}\rho$ and $\tau_c^2 = \sigma_c^2(1-\rho^2)$.

The sets of the QALYs and cost parameters from the models specified in Equation 5.4 and Equation 5.5 are thus $\boldsymbol{\theta}_e = (\alpha_0, \alpha_1, \sigma_e^2)$ and $\boldsymbol{\theta}_c = (\beta_0, \beta_1, \mu_e, \sigma_c^2, \sigma_e^2)$ — note that the marginal mean and variance of the QALYs link the two modules and therefore feature in both sets of parameters. The model is completed by assigning suitable prior distributions to the elements of $\boldsymbol{\theta} = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_c)$. We specify independent Normal$(0, 1000)$ priors for the regression parameters, while Uniform$(0, 1000)$ priors are assigned on the scale of the standard deviations.

A limitation of the Bivariate Normal model is that it fails to capture skewness in both outcomes. This may introduce some bias in the estimates, especially when the sample size is small and the degree of asymmetry in the data is relatively high.

5.2.2 Beta-Gamma

The second model assumes a Beta marginal for the QALYs and a Gamma conditional for the costs. We choose this specification as, in addition to the correlation between e_i and c_i, it allows to capture skewness in both outcome variables.

We parameterise the Beta distribution in terms of the mean ϕ_{ie} and the precision parameter $\tau_{ie} = \left(\frac{\phi_{ie}(1-\phi_{ie})}{\sigma_e^2} - 1\right)$ as $e_i \mid \boldsymbol{\theta}_e \sim \text{Beta}\left(\phi_{ie}\tau_{ie}, (1-\phi_{ie})\tau_{ie}\right)$. We model the location using a logit link function and include $u_{i0}^\star$ in the QALYs regression

$$g_e(\phi_{ie}) = \text{logit}(\phi_{ie}) = \alpha_0 + \alpha_1 u_{i0}^\star. \tag{5.6}$$

The costs are modelled as $c_i \mid e_i, \boldsymbol{\theta}_c \sim \text{Gamma}\left(\phi_{ic}\tau_{ic}, \tau_{ic}\right)$, where the shape parameter is defined as the product of the location ϕ_{ic} and the rate τ_{ic}. We use a logarithmic link function for the location and include the centered QALYs in the cost regression

$$g_c(\phi_{ic}) = \log(\phi_{ic}) = \beta_0 + \beta_1(e_i - \mu_e). \tag{5.7}$$

The marginal means for the QALYs and total costs can then be obtained using the respective inverse link functions

$$\mu_e = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)} \qquad \text{and} \qquad \mu_c = \exp(\beta_0). \tag{5.8}$$

The model is again completed using Normal$(0, 1000)$ priors on the regression coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and a Uniform$(0, 1000)$ prior on the standard deviation σ_c. As for σ_e, a little more care is needed in defining suitable prior distributions. In fact, by the mathematical properties of the Beta distribution, the variance is bounded by a function of the mean such that $\sigma_e^2 \leq \mu_e(1-\mu_e) = v$. Consequently, we can place an informative prior on the standard deviation $\sigma_e \sim \text{Uniform}(0, \sqrt{v})$, which coupled with a prior for μ_e induces a suitable prior for τ_{ie} as well. We choose this parameterisation of the Beta distribution because it is generally easier from an interpretation perspective to specify priors on the standard deviations rather than on the precision or variance parameters, for which different types of weakly informative priors may have different impacts on σ_e^2.

Notice that, in comparison to the Bivariate Normal, the Beta-Gamma model reflects more closely the range of the observed data, i.e. between $(0, 1)$ for e_i and between $(0, +\infty)$ for c_i. Nevertheless, the model fails to directly account for the structural values, e.g. unit QALYs or zero costs, which do not belong to the support of the Beta and Gamma distributions. Since in the MenSS trial, some individuals are associated with $e_i = 1$ and $c_i = 0$, it is necessary to rescale the observed data before fitting the model. We therefore apply the Beta and Gamma distributions to $e_i^* = e_i - \epsilon$ and $c_i^* = c_i + \epsilon$ and we assessed the sensitivity of the results to different choices of ϵ. The mean QALYs and costs in both intervention groups remain almost unaffected by the choice of ϵ and suggest a general robustness of the posterior estimates (Figure C.4 in Appendix C).

5.2.3 Hurdle Model

To overcome the limitations of the Beta-Gamma model in terms of the structural ones, we expand it to a hurdle version. This is achieved through the incorporation of a new module into the framework, which is linked to the marginal model for e_i and allows to explicitly handle the values at the boundary of the QALYs range.

Specifically, for each subject in the trial we define an indicator variable d_{ie} taking value 1 if the $i-$th individual is associated with a unit QALYs ($e_i = 1$) and 0 otherwise ($e_i < 1$). This variable defines the module for the structural ones into the framework and is modelled as

$$\begin{aligned} d_{ie} := \mathbb{I}(e_i = 1) &\sim \text{Bernoulli}(\pi_{ie}) \\ \text{logit}(\pi_{ie}) &= \gamma_0 + \gamma_1 u_{i0}^\star \; [+ \ldots], \end{aligned} \tag{5.9}$$

where π_{ie} is the individual probability of unit QALYs, which is estimated on the logit scale as a function of a baseline parameter γ_0 and the centred baseline utilities $u_{i0}^\star$, whose effect is quantified by the parameter γ_1. Similarly to the QALYs and cost models, other covariates can be additively included in the model of d_{ie}. We specifically distinguish the baseline utilities from any other covariate as they are likely to be particularly informative in predicting whether an individual is associated with a structural one in the QALYs. All the logistic regression parameters in Equation 5.9 are given vague Normal$(0, 1000)$ priors.

Within this module, we can apply the inverse logit function on γ_0 to retrieve the quantity

$$\bar{\pi}_e = \frac{\text{exp}(\gamma_0)}{1 + \text{exp}(\gamma_0)}, \tag{5.10}$$

which represents the estimated marginal probability of unit QALYs. Depending on the value of d_{ie}, we can partition the observed data on the QALYs into two subsets. In the first subset, defined as the n^1 subjects for whom $d_{ie} = 1$, we define a variable $e_i^1 = 1$. Conversely, the second subset consists of the $n^{<1} = (n - n^1)$ subjects for whom $d_{ie} = 0$ and for these individuals we define a variable $e_i^{<1}$.

Because the individuals associated with $e_i^{<1}$ have QALY values that are less than 1, we can model this variable directly using a Beta distribution, which is characterised by an overall mean $\mu_e^{<1}$. Next, using the estimated value for $\bar{\pi}_e$ from Equation 5.10, we can compute the overall population average QALYs measure μ_e as the linear combination

$$\mu_e = (1 - \bar{\pi}_e)\mu_e^{<1} + \bar{\pi}_e, \tag{5.11}$$

where the parameters $\bar{\pi}_e$ and $(1 - \bar{\pi}_e)$ in effect represent the weights used to mix the means of the two components $\mu_e^1 = 1$ and $\mu^{<1}$.

It is also possible to extend the cost model to account for the individuals associated with a zero cost. However, because of the small number of $c_i = 0$ in the MenSS trial, the inclusion of a hurdle

module for the structural zeros has almost no impact on the posterior results for c_i compared with the Beta-Gamma model. We therefore keep the same specification for the cost model as described in Equations 5.7 and 5.8.

Figure 5.2 shows a graphical representation of the Hurdle Model. In addition to the modules

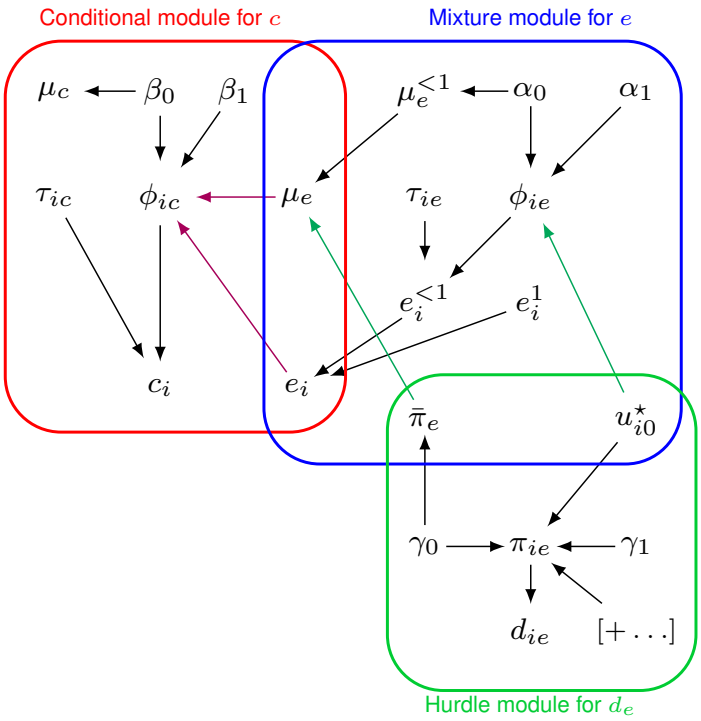


Figure 5.2: Three modules form the framework. The first two are the marginal distribution of e_i and the conditional distribution of $c_i \mid e_i$, respectively indicated by the blue and red box. The third, indicated by a green box, is the Hurdle module for d_{ie}, which separates the structural (e_i^1) and non-structural ($e_i^{<1}$) values in e_i. The solid black arrows show the dependence relationships between the parameters within the modules, while the magenta (green) arrows show the dependence between the parameters of the QALYs and costs (hurdle) modules. The dots enclosed in the square brackets indicate the potential inclusion of other covariates at the mean level in the model for π_{ie}.

associated with the QALYs (blue box) and cost (red box) variables, the Hurdle Model is formed by a third module related to the structural ones (green box), which allows to split e_i into the two components e_i^1 and $e_i^{<1}$. Using the modular structure of the framework it is possible to increase model complexity in a relatively easy way to account for the presence of structural values in either or both the outcomes, while simultaneously capturing both correlation and skewness.

5.3 All Cases Scenario

Each of the three models specified under the "complete cases" scenario can be easily extended to what we term the "all cases" scenario by additionally imputing the values for all individuals with unobserved data.

Specifically, when missingness occurs in the QALYs and/or cost variables, no change to the models is required under a MAR assumption for both the Bivariate Normal and Beta-Gamma specifications. For the Hurdle Model, when e_i is missing, it is not possible to directly define the value for d_{ie}. However, unit QALYs can only be observed if $u_{ij} = 1$ for all time points $j = 0, \ldots, J$. Consequently, we can use the observed values for u_{ij} to inform the corresponding value for d_{ie}; if an individual i is such that u_{ij} is missing at some time point j and $u_{ij} \neq 1$ at any other time point, then by necessity $d_{ie} = 0$. Notice that for all individuals having $u_{ij} = 1$ at all observed time points, but with at least one missing value at some other time point, d_{ie} is unknown.

When present in the analysis, incomplete covariates need to be explicitly modelled to impute their missing values. Within the framework, this corresponds to adding a new module for each of these variables. In the MenSS trial, the only partially-observed covariates are the baseline utilities u_{i0}, which are linked to e_i (all models) and d_{ie} (only Hurdle Model) to account for their impact on the inferences. The missing values in u_{i0} under the Bivariate Normal and the Beta-Gamma formulations are handle by assuming the same distribution of the outcome e_i, i.e. Normal and Beta, respectively. In general, if other partially-observed covariates are available, the same approach can be extended to handle the missing values in those variables.

As for the Hurdle model, it is possible to apply the same approach used for e_i to handle the structural ones in the baseline utilities through the inclusion of a module for $u_{i0} = 1$. First, a model for the individuals with a non-unit utility value $u_{i0}^{<1}$ is specified. Again, a simple solution is to base this on the same distribution assumed for $e_i^{<1}$, i.e. Beta. Second, the probability of observing a structural one in the baseline utilities π_{iu} is estimated as

$$\begin{aligned} d_{iu} := \mathbb{I}(u_{i0} = 1) &\sim \text{Bernoulli}(\pi_{iu}) \\ \text{logit}(\pi_{iu}) &= \kappa_0 \ [+ \ldots], \end{aligned} \tag{5.12}$$

where d_{iu} is the indicator variable for observing u_{i0}^1 in the baseline utilities and κ_0 is the intercept term of the logistic regression of π_{iu}. Similarly to the QALYs model, we can include other terms in Equation 5.12 to capture the impact of relevant covariates on π_{iu}.

5.3.1 Sensitivity analysis (MNAR)

Finally, we expand the Hurdle Model under the "all cases" scenario to assess the robustness of the results to some departures from MAR. Specifically, hurdle models offer a convenient setting for performing a simple type of sensitivity analysis to the missingness assumptions.

The analysis involves two groups of cases: **a)** the individuals for whom utility values are missing throughout the follow up, i.e. $u_{ij} = \texttt{NA}$ for all $j = 1, \ldots, J$; **b)** the individuals for whom all the observed utilities are equal to 1, but with at least one time point j at which $u_{ij} = \texttt{NA}$.

For both these groups, it is impossible to compute the value of the indicator d_{ie} according to the information from the observed data and, under MAR, these cases are assigned by the model to either e_i^1 or $e_i^{<1}$. However, given that for these individuals no utility score < 1 is ever observed, it is plausible to assume that at least some of them remain in full health for the entire duration of the trial and are therefore belong to e_i^1. We can then use the structure of the Hurdle Model and arbitrarily set the value of d_{ie} to either 1 or 0 using different configurations, e.g. by varying the number of structural values potentially observed in a given scenario. Since these configurations are based on assumptions about the missing values that cannot be verified from the data at hand (but are in fact arbitrarily set by the experimenter), they effectively represent a way to assess the robustness of the results to some departures from MAR.

In the MenSS trial, there are $n^* = 13$ (12%) individuals in the control and $n^* = 22$ (26%) in the intervention group who fall within group **a** or **b**. Thus, we perform sensitivity analysis by defining a set of alternative MNAR scenarios for these individuals and assess the robustness of the results across them. The four different scenarios considered are summarised in Table 5.1. We choose these scenarios in order to assess the impact of different "extreme" combinations of the number of potential structural ones in the intervention and control groups on the inferences and cost-effectiveness conclusions compared with the "all cases" scenario under MAR. Although these scenarios are associated with deterministic MNAR values, they can be incorporated in the framework at no extra cost in terms of model complexity and are meant to provide a broad assessment about the impact that missingness uncertainty may have on the final conclusions and

Scenario	**Control** ($n^* = 13$)	**Intervention** ($n^* = 22$)
MNAR1	$d_{ie} = 1$	$d_{ie} = 1$
MNAR2	$d_{ie} = 0$	$d_{ie} = 0$
MNAR3	$d_{ie} = 1$	$d_{ie} = 0$
MNAR4	$d_{ie} = 0$	$d_{ie} = 1$

Table 5.1: Alternative MNAR scenarios considered in the MenSS study for the Hurdle Model. In each scenario, individuals who are potentially associated with a unit QALYs in the control ($n^* = 13$) and intervention ($n^* = 22$) group are assigned to either the structural or non-structural components by setting the value of the indicator d_{ie} equal to 1 or 0, respectively.

how these vary with respect to the results obtained under MAR. Specifically, if the results are not robust to the departures explored, these "extreme" scenarios offer a starting point for further analyses where more advanced methods can be used to explicitly allow for the variability in the MNAR values, e.g. selection or pattern mixture models (Section 1.6). A novel modelling approach which implements these methods in trial-based economic evaluations is presented and discussed in Chapter 6.

5.4 Application to the MenSS trial

Following the original analysis in the MenSS trial, we fit all models under a "complete cases" scenario and add the baseline utilities in the module for the QALYs. In the Hurdle Model, we also include three fully-observed covariates, namely age, ethnicity and employment status (see Section 3.1.1) in the linear predictor of Equation 5.9 to improve the estimation of the probability of structural ones. We then extend the models to the "all cases" scenario by including the module for u_{i0} into the framework and imputing all missing outcome data under MAR. In the Hurdle Model, the baseline utilities are explicitly modelled using the hurdle approach and the estimation of the probability of structural ones is again improved through the inclusion of age, ethnicity and employment status in Equation 5.9 (for the QALYs) and Equation 5.12 (for the baseline utilities).

5.4.1 Software

We fitted all models to the MenSS data using `JAGS`, which is interfaced with `R` through the package `R2jags` (Su and Yajima, 2015). We ran two chains with 20,000 iterations per chain, using a burn-in of 10,000, for a total sample of 20,000 iterations for posterior inference. For each unknown quantity in the model, we assessed convergence and autocorrelation of the simulations through visual inspection of the density, trace and autocorrelation plots and other MCMC diagnostic measures. These include the potential scale reduction factor (below 1.05 for all parameters) and the effective sample size (at least 10,000 for all parameters).

Alternative prior distributions were considered to check that we were not incorporating any unintended information into the models through the priors. We specified Uniform$(0, 10000)$ and Half-Normal$(0, 1000)$ priors for the standard deviations or chose different values for the variance of normally distributed regression parameters, e.g. Normal$(0, 100000)$. Figure C.6, available in Appendix C.3.3, shows the robustness of the posterior inferences of each model fitted to the MenSS data to the three alternative priors considered. Overall, the results were robust to these specifications.

The Hurdle Model as described in Section 5.2 can be implemented in `JAGS` using a simple "coding trick". Appendix B.2 provides the full `JAGS` script for the Hurdle Model, while Appendix C.3.2

shows the technical details of the coding trick for implementing the model and compares the posterior results for the MenSS trial with respect to alternative configurations.

5.4.2 Model Assessment

We compare the fit of the different models using the DIC (Section 1.3.2). We consider a DIC based on the observed data under an ignorability assumption for the missing data (i.e. integrated over the missing data). The reason for the need to calculate the DIC for a specific model using the observed data alone is that model fit should never be assessed with respect to the imputed data (Daniels and Hogan, 2008). Indeed, the fit of the model to the imputed data depends on the specific missingness assumptions made to derive those imputations, which can never be verified from the available data, therefore making it difficult to use any measure of fit for the purpose of model assessment. A sample algorithm for the computation of the DIC based on the observed data is provided in Appendix A.2.1. In our analyses we calculated the observed data DIC only for the modules that are in common between the models, i.e. excluding the contribution from the structural indicators for the Hurdle Model.

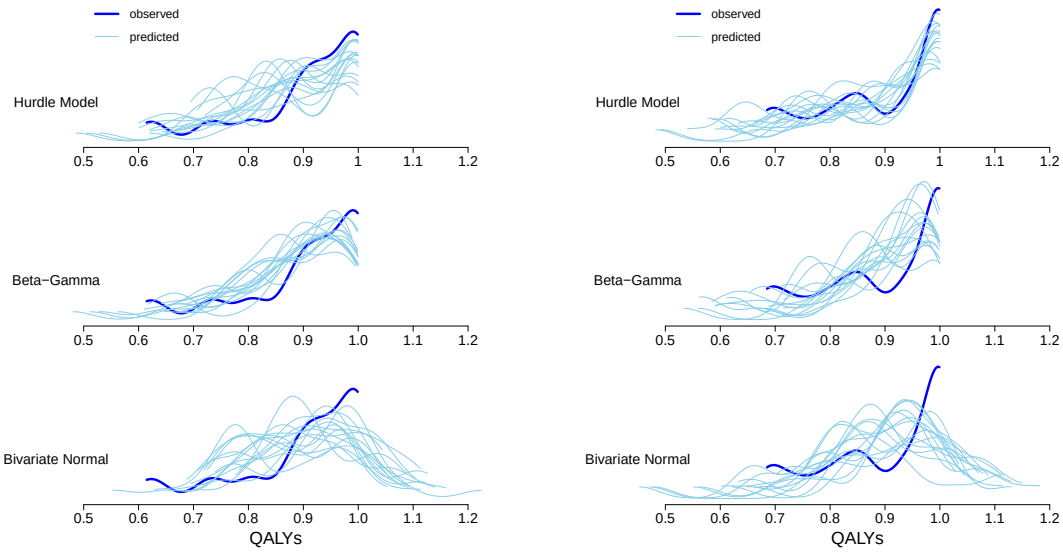
Table 5.2 compares the values of the DIC and the effective number of parameters p_D computed from the observed data likelihood for each variable across the three models under the "all cases" scenario. The total values for the DIC and p_D for each model are also reported at the bottom of the table, while values of the actual number of parameters for each variable and for the whole model are reported in brackets. The number of parameters is similar in each specification, as indicated

	Bivariate Normal		Beta-Gamma		Hurdle Model	
variable	DIC	p_D	DIC	p_D	DIC	p_D
u_0	-92	3.7 (4)	-322	4.3 (4)	-2474	4 (4)
$e \mid u_0$	-90	7.1 (6)	-135	6.7 (6)	-572	7.6 (6)
$c \mid e$	627	6.4 (6)	517	6 (6)	517	5.9 (6)
Total	445	17.2 (16)	60	17 (16)	*-2419*	17.5 (16)

Table 5.2: DIC and p_D based on the observed data likelihood for each variable in the Bivariate Normal, Beta-Gamma and Hurdle Model fitted to the MenSS data. Total DIC and p_D are reported at the bottom of the table, while the actual number of parameters for each variable and for the whole models are reported in brackets. The lowest DIC values are shown in italics.

by the almost identical values of p_D; by contrast, the DIC values show considerable differences between the models. The Bivariate Normal model is always associated with the highest DIC for all variables. The Beta-Gamma and, especially the Hurdle model substantially improve the model fit to the observed data. The baseline utilities are the variables associated with the largest DIC decrease moving from either the Beta-Gamma or Bivariate Normal to the Hurdle Model. This is expected as the Hurdle Model is better aligned with the empirical data distributions compared with the other models. Similar conclusions are obtained when comparing the total DIC value for the Bivariate Normal (536), Beta-Gamma (386) and Hurdle Model (-50) under a "complete cases" scenario.

We also assess the fit of the models to the observed data using different types of posterior predictive checks. Figure 5.3 shows the posterior predictive QALYs densities for the complete cases from the Bivariate Normal, Beta-Gamma and Hurdle Model (light blue lines). These are compared with the empirical distributions of the complete cases (dark blue lines) in both treatment groups of the MenSS trial. When Normal distributions are used (Bivariate Normal model), the replicated QALYs poorly fit the observed data in both the control and intervention groups. In addition, the densities generated from the model fall outside the permitted range of the QALYs, exceeding the upper threshold of 1. The Beta distributions (Beta-Gamma model) improve the



(a) control **(b)** intervention

Figure 5.3: Posterior predictive QALYs densities for Bivariate Normal, Beta-Gamma and Hurdle Model (light blue lines) compared with the empirical distributions of the complete cases (dark blue lines) in the control (panel a) and intervention (panel b) in the MenSS trial. For each model, 1000 replications of the QALYs data are generated, of which only 15 are displayed in each graph for visualisation purposes.

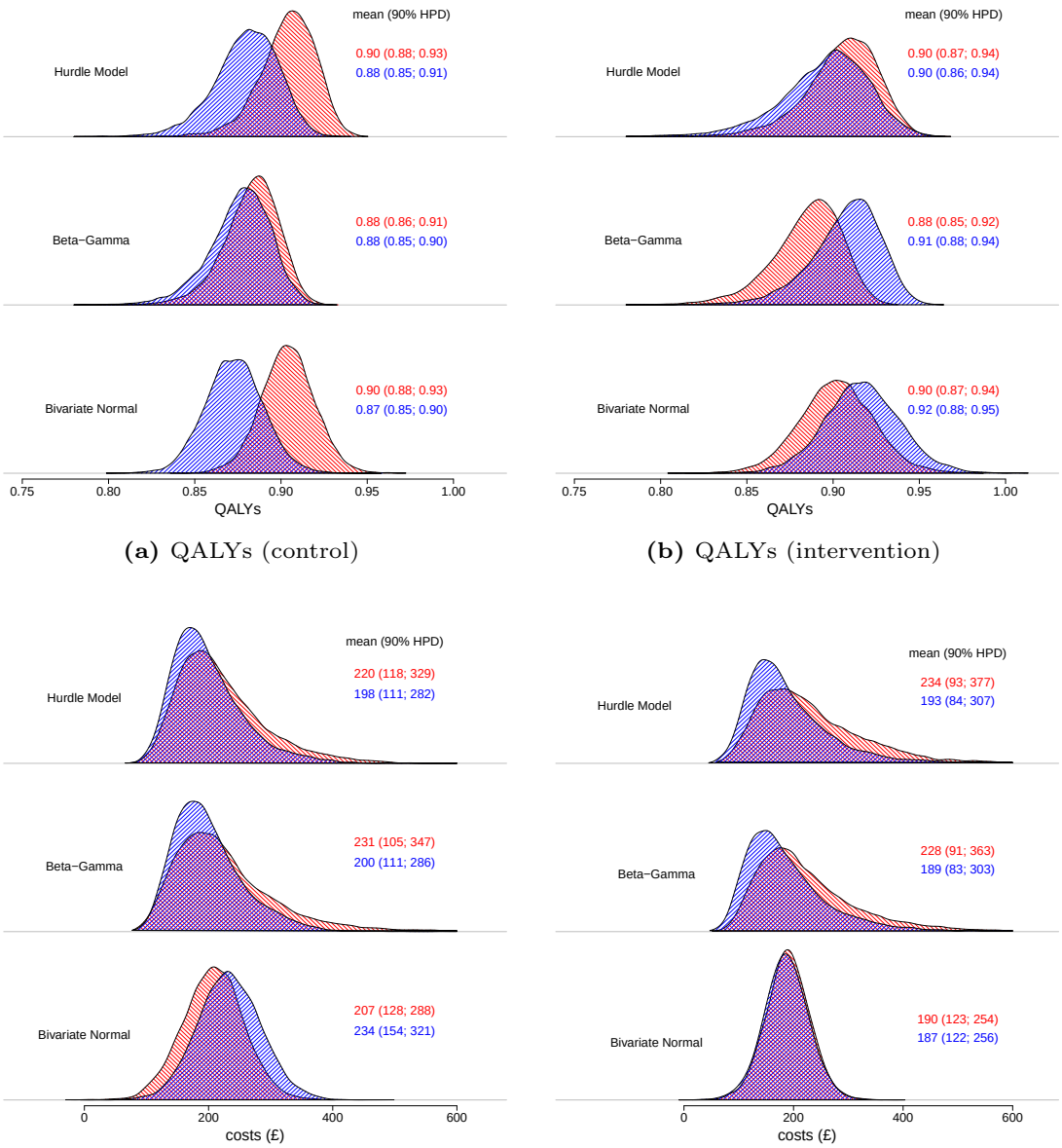
fit to the observed data compared with the Normal distributions but, especially in the intervention group, fail to capture the unit QALYs. Finally, extending the Beta distributions to a hurdle approach for the structural ones (Hurdle Model) provides the best predictive performance among the models assessed and most closely fits the empirical distributions in both treatment groups. The results from other posterior predictive checks, which evaluate the predictive ability of the Hurdle Model with respect to the QALYs variables are presented in Appendix C.3.5.

5.5 Results

5.5.1 Complete and All Cases Scenarios (MAR)

Figure 5.4 shows the posterior distributions of the mean QALYs and costs for both the control and intervention group under a “complete cases” (red) and “all cases” (blue) scenarios for each model under MAR. There are some discrepancies in the posterior distributions of the mean QALYs (panels a-b) between the “complete cases” and “all cases” scenarios, with the magnitude varying according to the treatment group and model considered. In general, the estimates under all cases are lower in the control group and higher in the intervention group in comparison to those obtained using the complete cases.

As for the mean costs (panels c-d), the estimates from a Gamma distribution are substantially more skewed and are typically associated with higher estimates compared with those from a Normal distribution. In addition, unlike the Bivariate Normal, both the Beta-Gamma and Hurdle Model typically lead to mean cost estimates that are systematically lower under the “all cases” scenario compared with the “complete cases” scenario.



(a) QALYs (control) **(b)** QALYs (intervention)

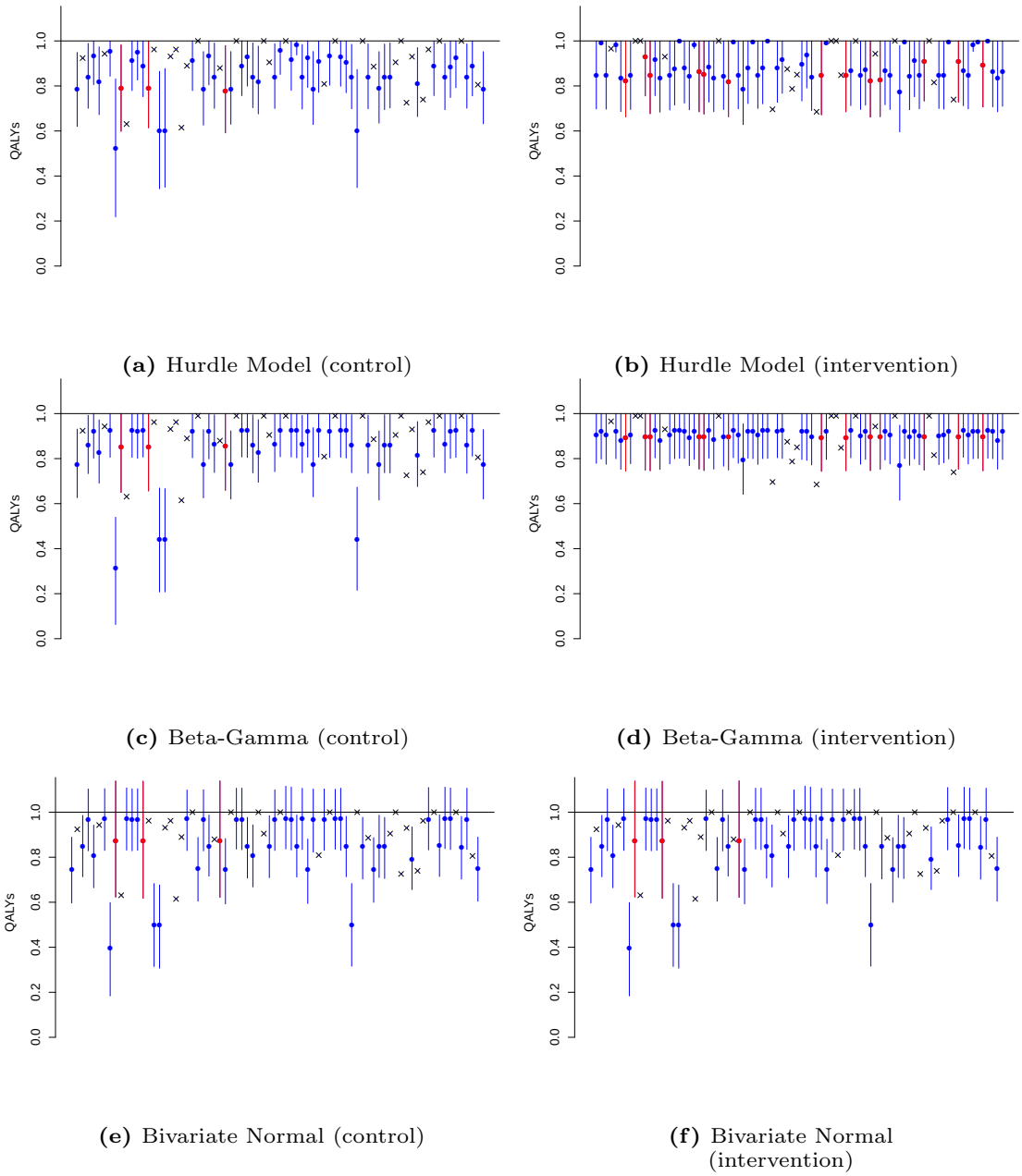
(c) costs (control) **(d)** costs (intervention)

Figure 5.4: Posterior distributions for the marginal mean parameters of the QALYs (panels a-b) and cost variables (panels c-d), expressed in £, in each group of the trial either under a "complete cases" (red) or "all cases" (blue) scenario under MAR. The posterior results are presented for all model specifications considered (Bivariate Normal, Beta-Gamma and Hurdle Model) and for each of these the posterior mean estimates and associated 90% HPD interval bounds are reported.

5.5.2 Imputations under MAR

Figure 5.5 shows the observed QALYs in both treatment groups (indicated with black crosses) as well as summaries of the posterior distributions for the imputed values, obtained from each model. Imputations are distinguished based on whether the corresponding baseline utility value is observed or missing (blue or red lines and dots, respectively) and are summarised in terms of posterior mean and 90% HPD intervals.

There are clear differences in the imputed values and corresponding credible intervals between the three models in both treatment groups. Neither the Bivariate Normal nor the Beta-Gamma models produce imputed values that capture the structural one component in the data. In addition, as to be expected, the Bivariate Normal fails to respect the natural support for the observed QALYs, with many of the imputations exceeding the unit threshold bound. These un-



(a) Hurdle Model (control)

(b) Hurdle Model (intervention)

(c) Beta-Gamma (control)

(d) Beta-Gamma (intervention)

(e) Bivariate Normal (control)

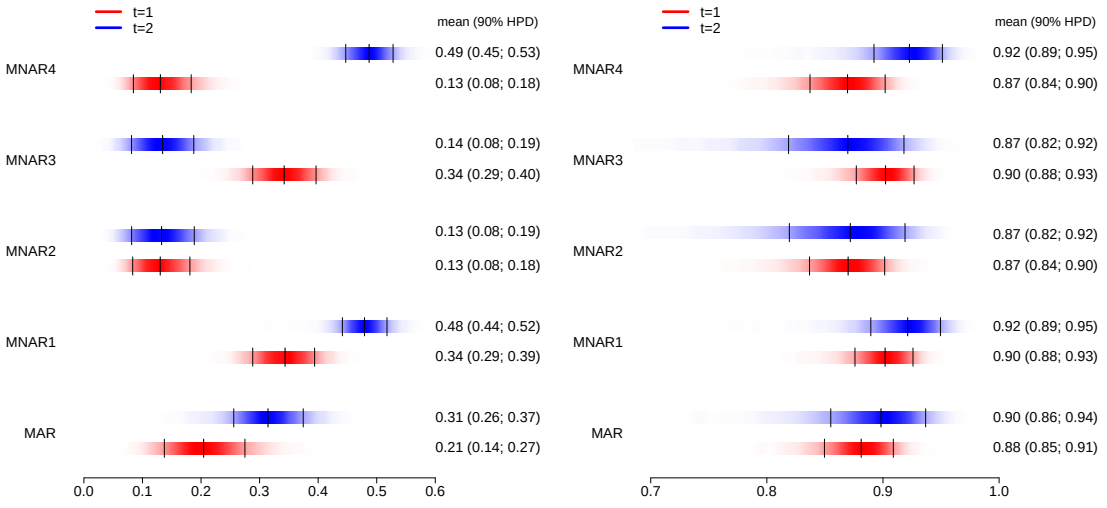
(f) Bivariate Normal (intervention)

Figure 5.5: Imputed QALYs in the control and intervention groups based on the Bivariate Normal, Beta-Gamma and Hurdle Model. Imputations are summarised in terms of posterior means and 90% HPD intervals (coloured dots and lines) while an x symbol is used to denote the observed cases. Imputed values are also distinguished according to whether the baseline utilities were either observed (blue) or missing (red). The solid black line represents the upper bound for the QALYs (calculated over one year), set at the value of 1.

realistic imputed values highlight the inadequacy of the Normal distribution for the data and may lead to distorted inferences. Conversely, imputations under the Hurdle Model are more realistic, as they can replicate values in the whole range of the observed data, including the structural ones. Imputed unit QALYs with no discernible interval are only observed in the intervention group due to the original data composition, i.e. individuals associated with a unit baseline utility and missing QALYs are almost exclusively present in the intervention group.

5.5.3 Sensitivity Analysis (MNAR)

For each of the alternative MNAR scenarios described in Section 5.3.1, as well as for the analysis under MAR, Figure 5.6 shows posterior density strips (Jackson, 2008) for the structural one probability $\bar{\pi}_e$ and the marginal mean QALYs μ_e, in the control (red) and intervention (blue) groups.



(a) Probability of structural ones

(b) Mean QALYs

Figure 5.6: Density strip plots for the posterior distributions of the probability of structural ones (panel a) and the marginal mean QALYs (panel b) under MAR and four alternative MNAR scenarios. For each scenario, results are presented for the control (red) and the intervention (blue) groups.

Estimates under MAR indicate that the new intervention is associated with a probability of observing a structural one and a mean QALYs that are on average higher compared to the control. Although the variations in the posterior mean QALYs across all scenarios do not exceed 0.05, they have a substantial impact in terms of the relative effectiveness between the two treatments because are associated with mean QALYs differentials of opposite sign. Specifically, while the results under both MAR and MNAR1 suggest that on average the new intervention is cost-effective with respect to the control ($\mathrm{E}[\Delta_e] > 0$), the estimated quantities are highly unstable across the other three MNAR scenarios. For example, under MNAR2 the probability of structural ones is substantially reduced in both groups and suggests on average an equivalent effectiveness between the two groups ($\mathrm{E}[\Delta_e] = 0$). However, under MNAR3 and MNAR4 the differences between the estimated probabilities and mean QALYs in the two groups are increased in magnitude and lead to opposite mean differentials ($\mathrm{E}[\Delta_e] < 0$ under MNAR3 and $\mathrm{E}[\Delta_e] > 0$ under MNAR4).

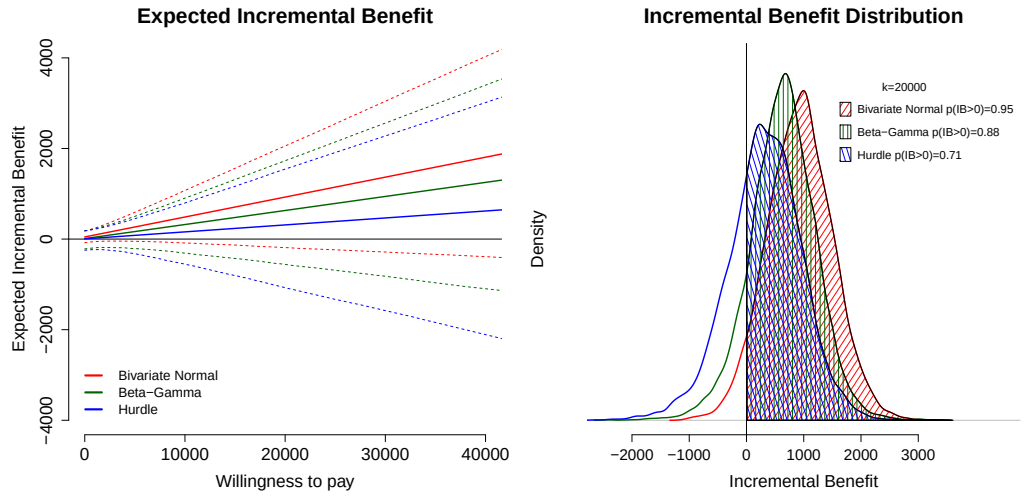
These results indicate a high sensitivity of the sign of the effectiveness differential for the two treatments in the MenSS trial with respect to the to the MNAR departures explored about the number of structural ones which can be potentially observed in both groups.

5.6 Economic Evaluation

We complete the analysis on the MenSS trial by assessing the cost-effectiveness of the new intervention with respect to the control, comparing the results of the different models under MAR and the alternative MNAR scenarios explored for the Hurdle Model. We specifically rely on the

examination of the EIB and IB distribution, together with the CEP and the CEAC to summarise the economic analysis.

We first compare the economic results between the different models estimated under MAR in terms of the impact on the Expected Incremental Benefit and Incremental Benefit distribution, shown in Figure 5.7. Overall, for all willingness to pay k values, the three models show positive



(a) Expected Incremental Benefit

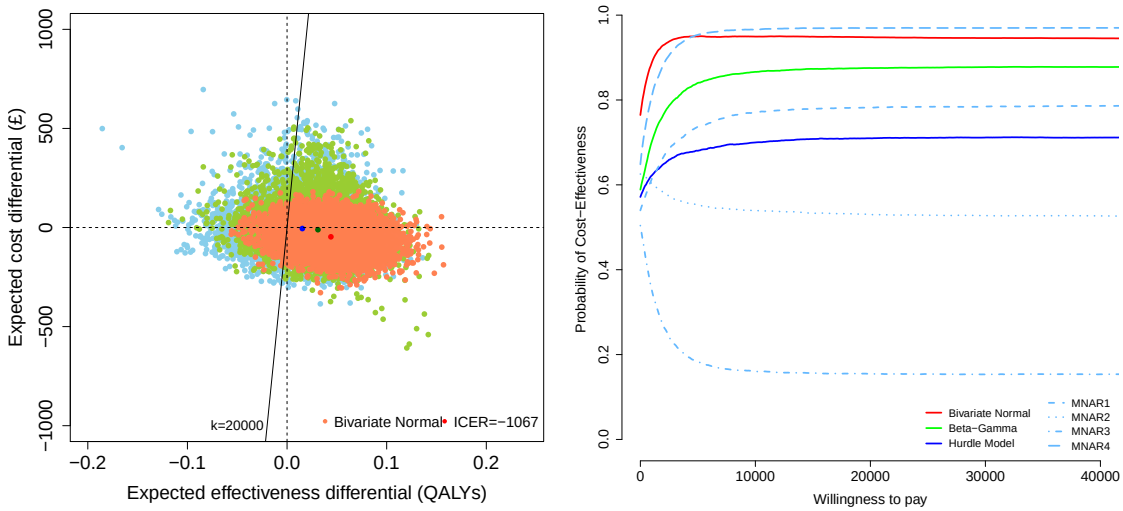
(b) Incremental Benefit Distribution

Figure 5.7: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with the Bivariate Normal (red solid and dashed lines), Beta-Gamma (green solid and dashed lines) and Hurdle (blue solid and dashed lines) models fitted to the data from the MenSS study under the "all cases" scenario.

EIB values (solid lines, panel a). However, the Bivariate Normal model is associated with the steepest slopes for the lines of the EIB and related lower/upper bounds (red solid and dashed lines), followed by the Beta-Gamma (green lines) and Hurdle model (blue lines). This is reflected in the plot of the IB (panel b), evaluated at $k = £20,000$, where the distribution under the Bivariate Normal model is shifted to the right compared with that of the Beta-Gamma and, even more, with respect to that of the Hurdle model. The highest probability of cost-effectiveness (shaded areas) is associated with the Bivariate Normal (0.95), followed by the Beta-Gamma (0.88) and the Hurdle model (0.71).

The CEP (Figure 5.8, panel a) shows that, at a willingness to pay $k = £20,000$, more than 70% of the samples for all three models (light red for the Bivariate Normal, light green for the Beta-Gamma and light blue for the Hurdle Model) fall in the sustainability area and are associated with negative ICERs, indicated by corresponding darker coloured dots. This suggests that the new intervention can be considered as cost-effective compared with the control, by producing a QALYs gain at virtually no extra costs or even saving money.

The CEAC (Figure 5.8, panel b) is evaluated up to a range for k of $£40,000$ per QALY gained. For each model, the results under MAR are reported using solid lines with different colours, i.e. red for the Bivariate Normal, green for the Beta-Gamma and blue for the Hurdle Model. In addition, the results associated with the four MNAR scenarios are reported using different types of dashed lines. Under MAR, for the Bivariate Normal and Beta-Gamma models the CEACs indicate the cost-effectiveness of the new intervention with a probability above 0.8 for most values of k. Conversely, under the Hurdle Model, the curve is shifted downward by 0.24 and 0.16 with respect to the Bivariate Normal and Beta-Gamma models, respectively, and suggests a more uncertain conclusion. Perhaps unsurprisingly, none of these results is robust to the alternative MNAR scenarios explored. The CEAC plot clearly shows a large sensitivity of the cost-effectiveness probability with



(a) Cost-Effectiveness Plane

(b) Cost-Effectiveness Acceptability Curve

Figure 5.8: 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 the four MNAR models described in Section 5.5.3 are represented with different types of dashed lines.

respect to the assumed number of structural ones in both treatment groups. Indeed, the curves span a huge probability range from 0.2 under MNAR4 to 1 under MNAR3. This implies a considerable change in the output of the decision process and severely undermines the validity of the conclusions obtained under MAR.

5.7 Application to the PBS study

We show the flexibility of our framework by adapting the models in Section 5.1 to accommodate the characteristics of the PBS data (Section 3.1.2). Specifically, we scale the QALYs to avoid negative values when using a Beta distribution, account for the multilevel structure in both outcomes and replace the Gamma distribution for the costs with a LogNormal distribution, which improves the fit to the observed data. In particular, the fit of the LogNormal distribution has a better fit compared with a Gamma distribution, with respect to both the empirical density and cumulative density functions evaluated on the observed costs in both treatment groups of the trial (see Figure C.12 in Appendix C.3.6).

We assess and compare the performance of three models. These are: 1) Bivariate Normal for the two outcomes; 2) Beta marginal for the QALYs and LogNormal conditional for the costs; and 3) Hurdle Model. All models account for the multilevel structure in the data and include three fully-observed categorical covariates (living condition, level of disability and type of carer) to estimate the mean QALYs and costs. All models are fitted and assessed using the same software configuration and diagnostic measures as in Section 5.4.1. The robustness of the posterior inferences of each model fitted to the PBS data with respect to the three alternative prior specifications described in Section 5.4.1 are shown in Figure C.7, which is available in Appendix C.3.3.

We first apply the models to the complete cases and then extend the analysis to all cases under MAR. No sensitivity analysis as in Section 5.3.1 can be performed for the PBS study because for each missing individual we observe a utility value that is lower than one at least at one time point,

i.e. none of them can be a structural one. In the next section, we present the specification of the Beta-LogNormal model to show an example of how the framework in Section 5.1 has been modified to address the characteristics of the PBS data.

5.7.1 Beta-LogNormal

The second model consists of a Beta marginal for the QALYs and a LogNormal conditional for the costs. Since the Beta distribution does not allow negative values, we scale the QALYs on $(0, 1)$ through the transformation $e_i^\star = \frac{e_i - \min(e_i)}{\max(e_i) - \min(e_i)}$ and fit the model to these transformed values. A similar scaling is applied to the baseline utilities $u_{i0}^\star$ when these are modelled in the "all cases" scenario. In both cases we use the theoretical lower (-0.594) and upper (1) bounds for the utilities and QALYs to derive the variables on the transformed scale (the range of the two variables are the same since QALYs are evaluated over one year in the PBS trial – see Section 1.2).

The Beta distribution is parameterised as in Section 5.2.2. The multilevel structure in the QALYs and cost model is accounted for through site-specific regression coefficients for the baseline utility and cost and intercept terms, respectively (i.e. varying-intercept/slope models), in accordance with the model specification assumed for the multilevel models in Chapter 4.

The QALYs are modelled as $e_i^\star \mid \boldsymbol{\theta}_e \sim \text{Beta}\left(\phi_{ie}\tau_{ie}, (1-\phi_{ie})\tau_{ie}\right)$ with location

$$\text{logit}(\phi_{ie}) = \alpha_{0s} + \alpha_{1s} u_{i0}^\star \; [+ \ldots], \tag{5.13}$$

where $\boldsymbol{\alpha}_s = (\alpha_{0s}, \alpha_{1s})$ include the structured coefficients for the intercept and $u_{i0}^\star$ which are associated with the $s = 1, \ldots, S$ sites.

The costs are modelled as $c_i \mid e_i^\star, \boldsymbol{\theta}_c \sim \text{LogNormal}\left(\phi_{ic}, \tau_c\right)$, where the mean and standard deviation parameters (ϕ_{ic}, τ_c) are defined on the log scale. The centered baseline costs $(c_{i0}^\star = c_{i0} - \bar{c}_0)$ are included in the cost model using the same multilevel specification used for u_{i0} in the QALYs model. Modelling $e_i^\star$ on a transformed scale does not allow to directly identify the marginal mean μ_e as in Equation 5.8 and therefore it is not possible to include the centered QALYs in the cost regression.

This, however, does not affect the estimates of the model because mean centering is only used to improve the efficiency of the MCMC sampling. Consequently, the cost location is

$$\phi_{ic} = \beta_{0s} + \beta_{1s}(e_i^\star) + \beta_{2s}(c_{i0} - \bar{c}_0) \; [+ \ldots], \tag{5.14}$$

where $\boldsymbol{\beta}_s = (\beta_{0s}, \beta_{1s}, \beta_{2s})$ are the site-specific baseline cost regression coefficients.

We retrieve the marginal means on the natural scale for both outcomes through Monte Carlo integration. At each iteration of the posterior distribution for the model parameters in the MCMC output, we generate a large number of samples for e_i and c_i and take the expectation over these values to obtain Monte Carlo estimates of the marginal means μ_e and μ_c.

The model is completed by specifying Normal priors for the structured coefficients $\boldsymbol{\alpha} \overset{i.i.d.}{\sim} \text{Normal}(\mathbf{0}, \boldsymbol{\sigma}_\alpha)$ and $\boldsymbol{\beta}_s \overset{i.i.d.}{\sim} \text{Normal}(\mathbf{0}, \boldsymbol{\sigma}_\beta)$ with shared standard deviations $\boldsymbol{\sigma}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\beta}}$. Finally, we choose independent $\text{Normal}(0, 1000)$ priors for the other regression coefficients in Equation 5.13 and Equation 5.14 and a $\text{Uniform}(0, 1000)$ for the standard deviations of the costs. The priors on the standard deviations of the QALYs are specified using the same approach described in Section 5.2.2.

5.7.2 Model Assessment

Similarly to the MenSS analysis, we compare each model fitted to the PBS trial using the DIC. For multilevel models, the DIC can be computed in different ways according to whether the data

likelihood is obtained by integrating out the random effects or conditional on them. Following current recommendations (Ntzoufras, 2009), for all multilevel models we compute the DIC based on the observed data likelihood under ignorability of the missing data but conditional on the random effects. For those models for which the sampling distributions of the outcomes were not available in closed form (Beta-LogNormal and Hurdle), Monte Carlo integration was used to calculate the corresponding likelihood and DIC. A more detailed discussion of how this type of DIC for multilevel models can be computed is provided in Appendix A.2.

Table 5.3 compares the DIC and p_D values computed from the observed data likelihood for each variable in the three models specified under the "all cases" scenario. Since c_{i0} are fully available in the PBS study, these variables are not directly modelled but are only included as covariates in all models. The total DIC and p_D values are reported at the bottom of the table. The

	Bivariate Normal		Beta-LogNormal		Hurdle Model	
variable	DIC	p_D	DIC	p_D	DIC	p_D
u_0	210	4.1	-103	4.1	-402	4.0
$e \mid u_0$	45	28	-124	26	-135	25
$c \mid e$	4431	18	4080	39.8	4082	41
Total	4686	50.1	3853	69.9	*3545*	70

Table 5.3: DIC and p_D based on the observed data likelihood for each variable in the Bivariate Normal, Beta-LogNormal and Hurdle Model fitted to the PBS data. Total DIC and p_D are reported at the bottom of the table, with the lowest DIC value shown in italics.

cluster-specific terms in both the QALYs and cost regressions do not allow to identify the actual number of parameters in the models. DIC values are computed conditional on the cluster, where p_D incorporates the estimated dimensionality of the random effects and takes into account the partial pooling of the coefficient estimates.

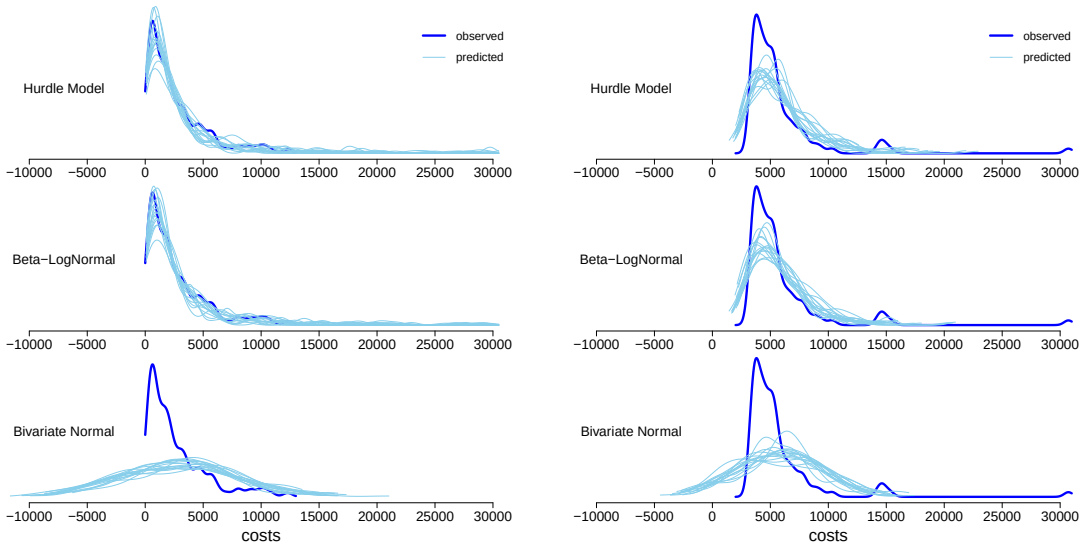
The Hurdle Model is associated with the lower DIC values for all variables compared with the alternative specifications. The DIC computed on both the modules of u_{i0} and e_{i0} show relatively large differences between the Bivariate Normal and either the Beta-LogNormal or the Hurdle Model. Cost variables are generally associated with lower DIC values when assuming LogNormal compared with Normal distributions. Similar conclusions are obtained when computing the total DIC values for the Bivariate Normal (5017), Beta-LogNormal (4538) and the Hurdle Model (3930) under the "complete cases" scenario.

Figure 5.9 compares the posterior predictive cost densities for the complete cases from the three models (light blue lines) with the empirical cost distributions (dark blue lines) in both treatment groups of the PBS trial. The replicated data under both the Beta-LogNormal and Hurdle Model more closely fit the observed data distributions compared with those from the Bivariate Normal, especially in the control group (panel a), even though no replicated samples seem to adequately capture the spike of the costs in the intervention (panel b) group. Among all replications compared, those from the Bivariate Normal have the worst fit to the observed data and lead to some unrealistic negative cost values. Appendix C.3.5 provides the results from other posterior predictive checks for the Hurdle Model, computed on the cost variables.

5.8 Results

5.8.1 Complete and All Cases (MAR)

Figure 5.10 shows the QALYs and cost mean parameters posterior densities by treatment group from the Bivariate Normal, Beta-LogNormal and Hurdle Model. The posterior distributions of the



(a) control **(b)** intervention

Figure 5.9: Posterior predictive cost densities for Bivariate Normal, Beta-LogNormal and Hurdle Model (light blue lines) compared with the empirical distributions of the complete cases (dark blue lines) in the control (panel a) and intervention (panel b) in the PBS trial. For each model, 1000 replications of the cost data are generated, of which only 15 are displayed in each graph for visualisation purposes.

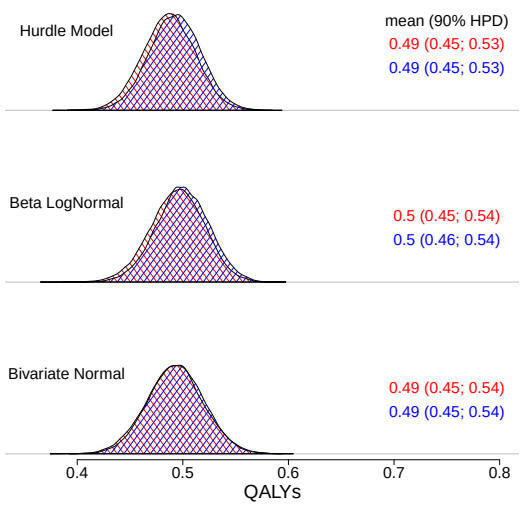
mean parameters are almost identical between the "complete cases" and "all cases" scenarios for both outcomes and treatment groups. Mean QALYs estimates show slight variations between the models (panels a-b), especially in the intervention where estimates under the Beta-LogNormal and Hurdle Model are lower compared to the Bivariate Normal. Mean cost distributions are characterised by some differences between the models in both treatment groups (panels c-d), with estimates from the Beta-LogNormal and Hurdle Model which are systematically lower compared with the Bivariate Normal.

5.8.2 Imputations under MAR

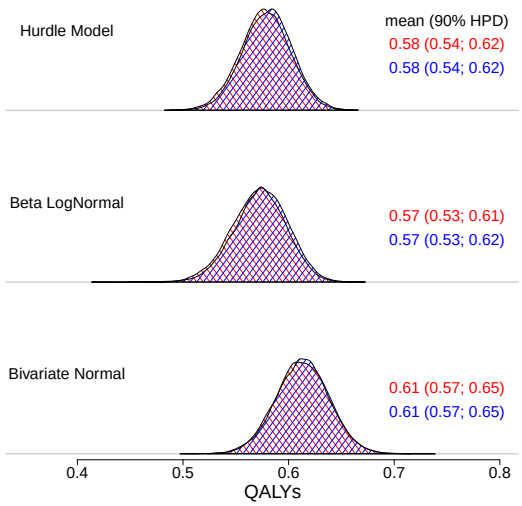
Figure 5.11 shows the imputed QALYs in both treatment groups in the PBS study under MAR for the Bivariate Normal, Beta-LogNormal and Hurdle Model. Some imputations under the Bivariate Normal exceed the upper threshold of one for the QALYs in both treatment groups. Therefore, implausible values (i.e. QALYs values that could not be observed in the PBS study) occur among the imputations. Conversely, both the Beta-LogNormal and Hurdle Model show imputed values that fall within the correct range of the data and suggest a better ability of the models to fit the observed data compared with the Bivariate Normal. No imputed value in both treatment groups under the Hurdle Model is associated with a structural one. This is due to the fact that all individuals in the PBS trial are associated with at least one observed utility value which is less than one at some time point, i.e. they cannot be associated with a unit QALYs.

5.9 Economic Evaluation

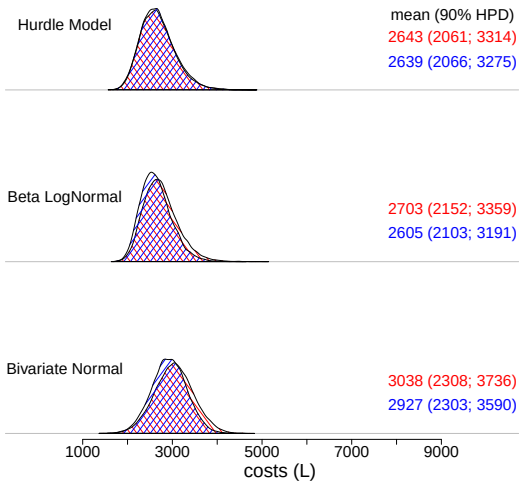
Figure 5.12 compares the estimates for the Expected Incremental Benefit and Incremental Benefit distribution obtained under the different models under MAR. The Bivariate Normal model is associated with the steepest EIB slope with positive/negative value after/before the willingness to pay value $k = £20,000$ (red solid line, panel a). The Beta-Lognormal and Hurdle model show very similar estimates for the EIB, which are negative for most of the values of k (green and blue



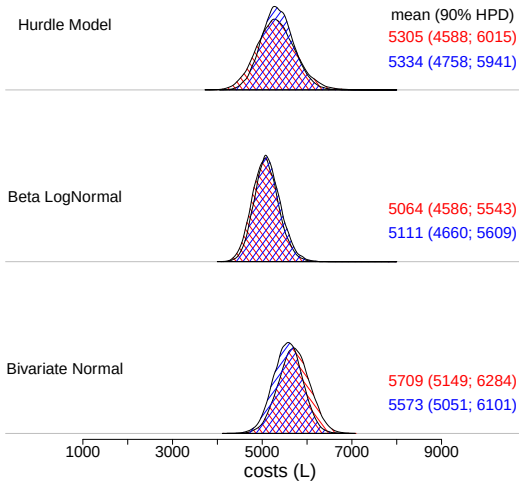
(a) QALYs (control)



(b) QALYs (intervention)



(c) costs (control)



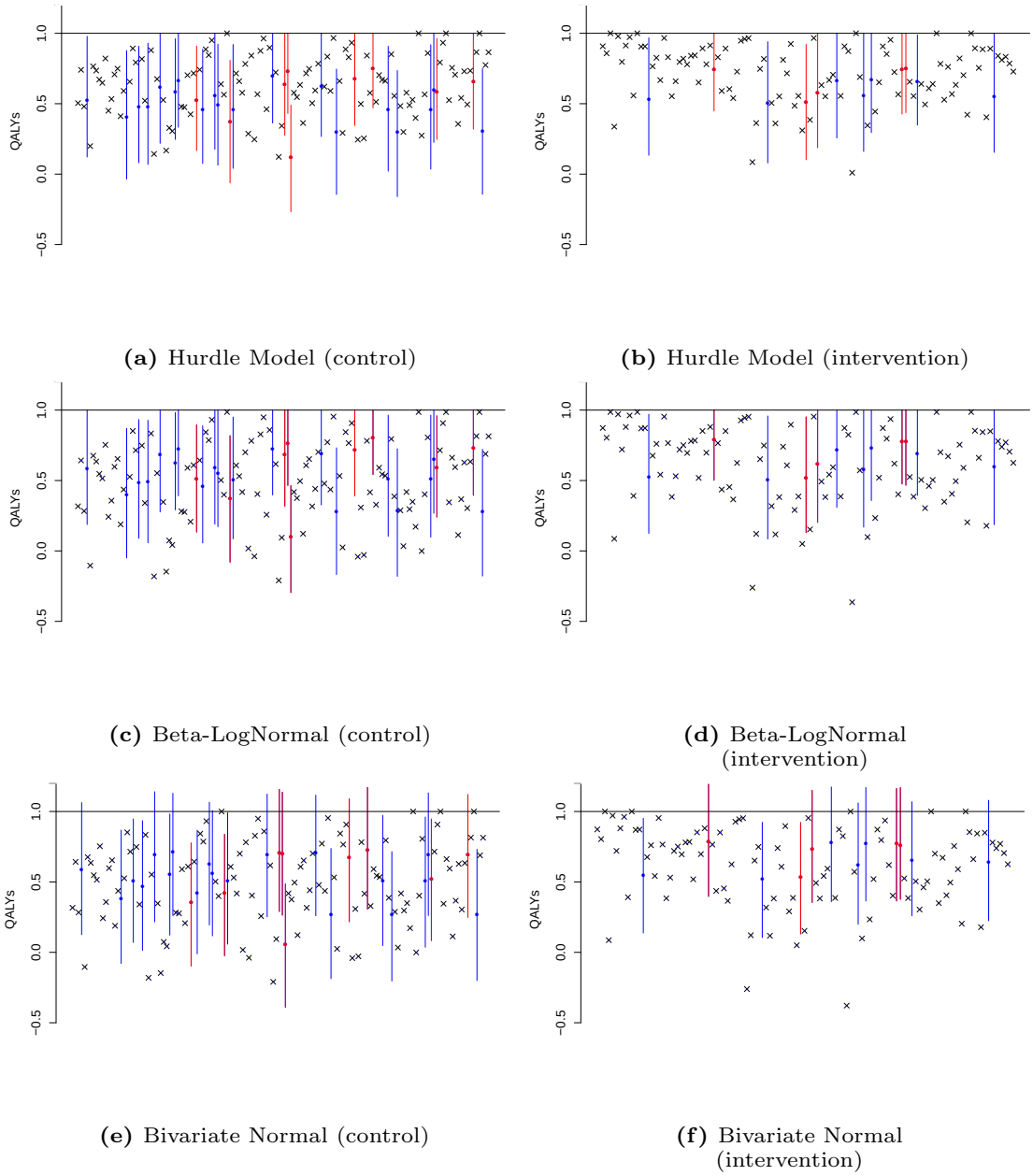
(d) costs (intervention)

Figure 5.10: Posterior distributions for the marginal mean parameters of the QALYs (panels a-b) and cost variables (panels c-d), expressed in £, in each group of the PBS trial under either a "complete cases" (red) or "all cases" (blue) scenarios. Results are presented for all model specifications considered (Bivariate Normal, Beta-LogNormal and Hurdle Model), with mean estimates and 90% HPD intervals reported in the graphs.

solid lines). Similar conclusions are obtained by looking at the EIB falling within the lower/upper bound estimates (dashed lines) which, however, contain both positive and negative values for all models. Most of the IB distribution under each model (panel b), evaluated at $k = £20,000$, falls below 0, especially for the Beta-Lognormal and Hurdle models which show a probability of cost-effectiveness (green and blue shaded areas) that are substantially lower compared with that of the Bivariate Normal (red shaded area).

Figure 5.13 shows the CEP and CEAC for the Bivariate Normal (red dots and line), Beta-LogNormal (green dots and line) and Hurdle Model (blue dots and line) for the "all cases" scenario under MAR.

For all three models, almost all samples in the CEPs (Figure 5.13, panel a) are located in the North-Eastern quadrant and most of them fall in the sustainability area at a willingness to



(a) Hurdle Model (control)

(b) Hurdle Model (intervention)

(c) Beta-LogNormal (control)

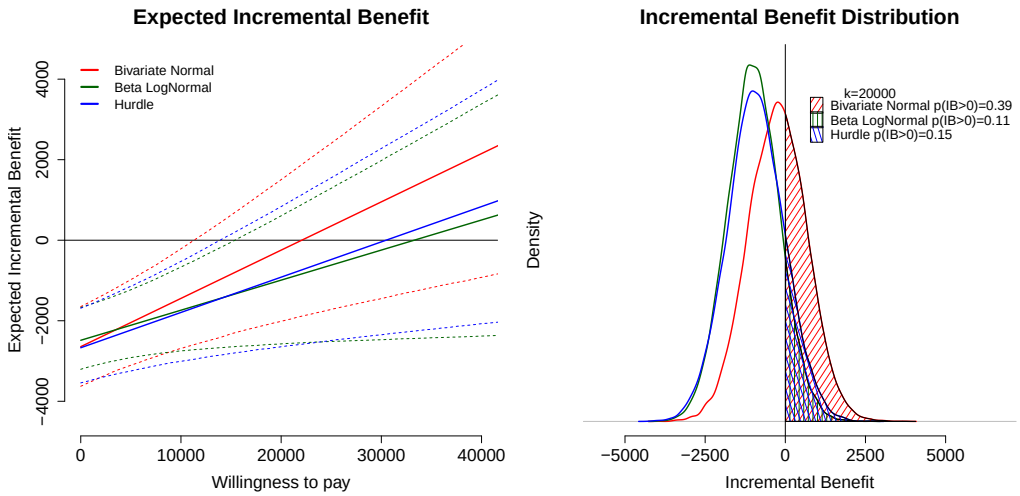
(d) Beta-LogNormal (intervention)

(e) Bivariate Normal (control)

(f) Bivariate Normal (intervention)

Figure 5.11: Imputed QALYs in both groups in the PBS trial based on the Bivariate Normal, Beta-LogNormal and Hurdle Model. Imputations are summarised in terms of means and 90% HPD intervals (coloured dots and lines) while an x symbol denotes the observed cases. Imputed values are distinguished according to whether the corresponding baseline utilities were either observed (blue) or missing (red). The solid black line represents the upper bound for the QALYs (calculated over one year), set at the value of 1.

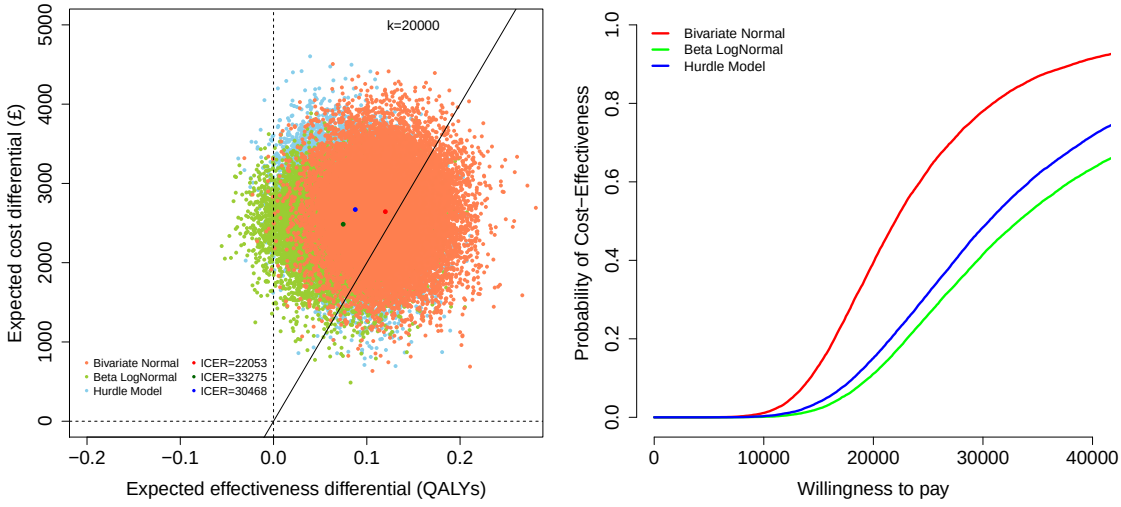
pay $k = £20,000$. Results from the Bivariate Normal model suggest a higher cost-effectiveness of the new intervention and are associated with a lower ICER compared with those of the other two models. This is reflected in the CEACs (Figure 5.13, panel b) which show a probability of cost-effectiveness that is on average 20% higher for the Bivariate Normal with respect to the Beta-LogNormal and Hurdle Model.



(a) Expected Incremental Benefit

(b) Incremental Benefit Distribution

Figure 5.12: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with the Bivariate Normal (red solid and dashed lines), Beta LogNormal (green solid and dashed lines) and Hurdle (blue solid and dashed lines) models fitted to the data from the PBS study under the "all cases" scenario.



(a) Cost-Effectiveness Plane

(b) Cost-Effectiveness Acceptability Curve

Figure 5.13: CEPs (panel a) and CEACs (panel b) associated with the Hurdle (blue dots and line), Bivariate Normal (red dots and line) and Beta-LogNormal (green dots and line) models.

5.10 Discussion

In this chapter we have presented a flexible Bayesian framework that can handle the typical complexities affecting outcome data in CEA, while also being relatively easy to implement using freely available Bayesian software. This is a key advantage that can encourage practitioners to move away from likely biased methods and promote the use of our framework in routine analyses.

The analysis of both case studies shows notable variations in the results, compared with those of the original analyses. In the MenSS trial, accounting for the structural ones and missingness uncertainty has a considerable impact on the cost-effectiveness of the new intervention and future research prioritisation. In the PBS trial, the results of the economic evaluation change substan-

tially when skewness is accounted for. In both cases, the Hurdle Model represents the best model among those assessed as it captures both skewness and structural values, while the other specifications fail to deal with at least one of these features.

Our results are obtained with specific reference to the two case studies. Specifically, the MenSS trial is a pilot RCT in which a large proportion of individuals have missing QALYs and cost values, while in the PBS trial some individuals in both treatment groups have negative QALY values. However, both the MenSS and PBS trials are very much representative of the "typical" dataset used in CEAs alongside RCTs. Thus, 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 many real cases.

We have demonstrated one possible way of assessing the robustness of the results to some "extreme" MNAR scenarios that can be incorporated in the framework at no extra cost in terms of model complexity. This analysis is easy to implement and offers a plausible starting point to assess the impact of missing data uncertainty on the results. In Chapter 6, we will show how the framework can be extended to incorporate a sensitivity analysis to nonignorable missingness using a PMM approach.

In conclusion, our framework can: *a)* jointly model costs and QALYs; *b)* account for skewness and structural values; and *c)* assess the robustness of the results under a set of differing missingness assumptions. The original contribution of this work consists in the joint implementation of methods that account for the complexities of the data within a unique and flexible framework that is relatively easy to apply. In the next chapter we will take a step forward in the analysis and present a longitudinal model that can use all observed utility and cost data in the analysis, explore alternative nonignorable missing data assumptions, while simultaneously handling the complexities that affect the data.

Key points of this chapter:

- We present a flexible Bayesian framework that jointly allows for the typical complexities that affect QALYs and cost data in CEA (e.g. skewness and spikes). The framework relies on the specification of a general modelling structure based on parametric distributions and combined modules in which variables are linked through logical relationships.
- We applied our framework to the MenSS and PBS trials. We compared the performance and inferences from a set of increasingly complex models that accounted for different types of complexities and data characteristics. Posterior results were in general sensitive to model specification and demonstrated the bias associated with routine analyses that ignore at least some of the complexities.
- In the MenSS trial, the model specification was extended to assess the robustness of the results to some "extreme" MNAR departures that can be implemented in the framework at no extra cost in terms of model complexity. As expected from the large proportions of unobserved values, inferences were sensitive to missingness assumptions and demonstrated that results under MAR are likely to be biased for this study.
- Our approach allows to jointly account for different data complexities within a unique and flexible framework, which can be implemented using freely available Bayesian software (e.g. `JAGS`). These are key features that can encourage analysts to abandon likely biased methods and improve the quality of the work in economic evaluations.

Chapter 6

A Bayesian Longitudinal Model for Handling Nonignorable Missingness in Health Economic Evaluation

In the previous chapter we have shown how a flexible Bayesian parametric framework can be constructed to jointly handle some of the complexities that affect trial-based CEA data and to conduct a sensitivity analysis to some "extreme" MNAR scenarios within a cross-sectional setting. In this chapter we extend the missing data analysis to formally account for the longitudinal nature of the data using a proper nonignorabale approach to handle missingness.

We present a Bayesian parametric model for conducting inference on the bivariate health economic response formed by the utility and cost data at each time point in a trial. The model expands the framework in Section 5.1 to a longitudinal setting to more efficiently deal with missingness and accommodates a sensitivity analysis to assess the robustness of the inferences and decision-making to a range of plausible nonignorable assumptions.

The proposed approach is motivated by and applied to the PBS data (Section 3.1.2), which present some features that make the study more suited to the implementation of the proposed framework compared with the MenSS trial. These include: a larger sample size and more moderate missingness rates across the follow-ups, an equal number of utility and cost variables at each time point (the baseline costs were not collected in the MenSS study), and the existence of observed utilities or costs at each time which can be used to inform the imputation of the other outcome at the same time (Section 3.1). Finally, while in the MenSS trial no external information was available to inform the MNAR departures to explore in sensitivity analysis, in the PBS trial it was possible to at least partially inform the direction and magnitude of the departures from MAR based on a discussion with the people involved in the trial. A synthesised version of this chapter in the form of a research article has been submitted for publication in *Journal of the Royal Statistical Society: Series A*.

6.1 Longitudinal Modelling Framework

We first present the longitudinal modelling framework for handling nonignorable missingness in trial-based individual-level CEA. Then, we demonstrate the benefits of adopting this approach using the PBS trial as motivating example.

Our nonignorable approach consists in specifying a full data model $p(\boldsymbol{y}, \boldsymbol{r})$ for the joint distribution of the response $\boldsymbol{y} = (u_{ijt}, c_{ijt})$ and missingness $\boldsymbol{r} = (r^u_{ijt}, r^c_{ijt})$ process based on the utility and cost data collected on the i-th person at time j in treatment group t of the trial. The modelling strategy uses the extrapolation factorisation (Section 1.6.1) to factor $p(\boldsymbol{y}, \boldsymbol{r})$ into the observed data distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r})$ and the extrapolation distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r})$.

The observed data distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r})$ is specified using a working model, which is fitted to $p(\boldsymbol{y}, \boldsymbol{r})$ and from which missingness is integrated out to leave the extrapolation distribution unidentified. A pattern mixture factorisation is used to separately specify a model for the marginal distribution of the missingness patterns $p(\boldsymbol{r} \mid \boldsymbol{\lambda})$ and the distribution of the response conditional on the patterns $p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta})$, respectively indexed by the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$.

Ideally, all pattern-specific response distributions should be modelled separately. However, in the PBS study, missingness is non-monotone and the data in most patterns are sparse (see Section 3.1.2), which makes practically infeasible to fit the response model within each pattern, with the exception of the completers ($\boldsymbol{r} = \mathbf{1}$). Alternative non-ignorable approaches, e.g. Selection Models (Section 1.6) could be considered but these are not well-suited for implementing the extrapolation factorisation and make the identification of sensitivity parameters, and therefore the incorporation of missingness assumptions, not transparent and often difficult to check (Daniels and Hogan, 2008).

Thus, to overcome this problem, we collapse together all the non-completers patterns ($\boldsymbol{r} \neq \mathbf{1}$) and fit the model separately to this aggregated pattern and to the completers. This strategy assumes that the reasons for missingness do not largely differ across the non-completers, which may not be realistic in some cases. However, it allows to identify the distribution of the responses for those who have some unobserved data without relying on the observations from the completers. The model is also well-suited to implement the extrapolation factorisation, which in turn allows to retrieve differences across the non-completers' patterns through the incorporation of sensitivity parameters with distributions that vary by type of outcome and time point (see Section 6.3.1).

Using the pattern mixture model approach (Section 1.6) we can express the full data model as

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta}) p(\boldsymbol{r} \mid \boldsymbol{\lambda}), \tag{6.1}$$

where $\boldsymbol{\omega} = (\boldsymbol{\eta}, \boldsymbol{\lambda})$ are the parameters of the joint distribution. Next, we apply the extrapolation factorisation to the right hand side of Equation 6.1 (Section 1.6.1) and obtain

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{y}^{\boldsymbol{r}}_{obs} \mid \boldsymbol{r}, \boldsymbol{\eta}_{\mathsf{O}}) p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r}, \boldsymbol{\eta}_{\mathsf{E}}) p(\boldsymbol{r} \mid \boldsymbol{\lambda}), \tag{6.2}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_{\mathsf{O}}, \boldsymbol{\eta}_{\mathsf{E}})$ are the parameters which index $p(\boldsymbol{y} \mid \boldsymbol{r})$ and which are associated with the observed data and the extrapolation distributions. Finally, we re-arrange the terms in Equation 6.2 and distinguish the components of $p(\boldsymbol{y} \mid \boldsymbol{r})$ on the right-hand side with respect to the completers ($\boldsymbol{r} = \mathbf{1}$) and non-completers ($\boldsymbol{r} \neq \mathbf{1}$), and the sets of parameters associated with these cases: $\boldsymbol{\eta}_{\mathsf{O}}^{\boldsymbol{r}=\mathbf{1}} = \boldsymbol{\eta}_{\mathsf{C}}$ for the completers, $\boldsymbol{\eta}_{\mathsf{O}}^{\boldsymbol{r}\neq\mathbf{1}} = \boldsymbol{\eta}_{\mathsf{NC}}$ for the observed data among the non-completers, and $\boldsymbol{\eta}_{\mathsf{E}}^{\boldsymbol{r}\neq\mathbf{1}} = \boldsymbol{\eta}_{\mathsf{E}}$ for the missing data among the non-completers. The model can be represented as

$$\left.\begin{aligned} p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega}) = p(\boldsymbol{r} \mid \boldsymbol{\lambda}) \, & [p(\boldsymbol{y}^{\boldsymbol{r}}_{obs} \mid \boldsymbol{\eta}_{\mathsf{C}})]^{\mathbb{I}\{\boldsymbol{r}=\mathbf{1}\}} \\ & [p(\boldsymbol{y}^{\boldsymbol{r}}_{obs} \mid \boldsymbol{\eta}_{\mathsf{NC}})]^{\mathbb{I}\{\boldsymbol{r}\neq\mathbf{1}\}} \end{aligned}\right\} \text{ observed data distribution}$$
$$[p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{\eta}_{\mathsf{E}})]^{\mathbb{I}\{\boldsymbol{r}\neq\mathbf{1}\}}, \quad \text{extrapolation distribution} \tag{6.3}$$

where the upper index $\mathbb{I}$ is used to distinguish the different components of the model according to the sets of cases to which they are related. The parameters $\boldsymbol{\omega} = (\boldsymbol{\eta}, \boldsymbol{\lambda})$ index the full data

model, while $\boldsymbol{\eta} = (\boldsymbol{\eta}_{\mathsf{C}}, \boldsymbol{\eta}_{\mathsf{NC}}, \boldsymbol{\eta}_{\mathsf{E}})$ are the parameters of the pattern-specific response model: $\boldsymbol{\eta}_{\mathsf{C}}$ and $\boldsymbol{\eta}_{\mathsf{NC}}$ are the distinct subsets of $\boldsymbol{\eta}$ that index the response model for the observed data in the completers and non-completers patterns, while $\boldsymbol{\eta}_{\mathsf{E}}$ is the subset of $\boldsymbol{\eta}$ that indexes the extrapolation distribution. Among the non-completers there are both $\boldsymbol{y}^{\boldsymbol{r}}_{mis}$ and $\boldsymbol{y}^{\boldsymbol{r}}_{obs}$ because the model is fitted to the aggregated pattern $\boldsymbol{r} \neq \mathbf{1}$ where intermittent missingness occurs (e.g. some individuals can be observed at $j = 1$ but not at $j = 0$).

The joint distribution $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ has three components. The first is the model for the missingness patterns and the response model fitted to $\boldsymbol{r} = \mathbf{1}$. The second is a model fitted to the joint set of the noncompleters, from which the missing values are integrated out to obtain a model for the observed data in $\boldsymbol{r} \neq \mathbf{1}$. This, together with the first component, forms the observed data distribution. The last is the extrapolation distribution. Example 6.1 shows how each component of the factorisation in Equation 6.3 can be identified within a basic analysis setting.

Table 6.1 displays the utility and cost data from the first few individuals in the PBS study and indicates the components of $\boldsymbol{y}^{\boldsymbol{r}}_{mis}$ and $\boldsymbol{y}^{\boldsymbol{r}}_{obs}$ in $\boldsymbol{r} \neq \mathbf{1}$ with red and blue cells, respectively. Across the non-completers, the missing and observed responses vary between utilities and costs and by time point. For example, at $j = 1$, $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{mis}$ includes the missing utilities for $i = 5, 6, 7, 10$ and $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{obs}$ includes the observed utilities for $i = 4, 8, 9$. Similarly, at time $j = 1$, $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{mis}$ contains the missing costs for $i = 7, 10$ and $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{obs}$ contains the observed costs for $i = 4, 5, 6, 8, 9$.

$\boldsymbol{r}$	i	$j=0$		$j=1$		$j=2$	
pattern	subject	utilities	costs	utilities	costs	utilities	costs
$\boldsymbol{r} = \mathbf{1}$	1	0.173	£9214	0.329	£961	0.436	£1973
$\boldsymbol{r} = \mathbf{1}$	2	0.850	£2492	0.692	£3437	0.336	£4550
$\boldsymbol{r} = \mathbf{1}$	3	-0.166	£15283	0.242	£2535	0.815	£3007
$\boldsymbol{r} \neq \mathbf{1}$	4	–	£1085	0.436	£1595	0.365	£3728
$\boldsymbol{r} \neq \mathbf{1}$	5	0.815	£3799	–	£2176	0.848	£533
$\boldsymbol{r} \neq \mathbf{1}$	6	0.436	£178	–	£474	0.244	£879
$\boldsymbol{r} \neq \mathbf{1}$	7	0.436	£4051	–	–	1	£107
$\boldsymbol{r} \neq \mathbf{1}$	8	0.367	£1197	0.367	£2580	–	–
$\boldsymbol{r} \neq \mathbf{1}$	9	0.708	£3788	0.815	£1597	–	–
$\boldsymbol{r} \neq \mathbf{1}$	10	0.273	£3412	–	–	–	–

$\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{mis}$

$\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{obs}$

Table 6.1: Utility and cost data for the i-th subject at each time j derived from a subset of the first 10 subjects in the PBS study. The response components in $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{mis}$ and $\boldsymbol{y}^{\boldsymbol{r}\neq\mathbf{1}}_{obs}$ are indicated with red and blue coloured cells, respectively. Missing responses are indicated with a $-$ sign.

Example 6.1

Let y_{ij} and r_{ij} denote a partially-observed univariate outcome and the corresponding missing data indicator associated with the $i-$th individual at the $j-$th time point in the study. Consider the simplified framework in which the total number of time points is two ($j = 0, 1$) and three different missingness patterns $\boldsymbol{r} = (\boldsymbol{r}_0, \boldsymbol{r}_1)$ are observed: completers; $\boldsymbol{r} = (1, 1)$, missing only at $j = 0$; $\boldsymbol{r} = (0, 1)$, and missing only at $j = 1$; $\boldsymbol{r} = (1, 0)$.
We specify the full data model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ in terms of the marginal model of the patterns $p(\boldsymbol{r} \mid \boldsymbol{\lambda})$ and the conditional model of the response given the patterns $p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta})$, where $\boldsymbol{\omega} = (\boldsymbol{\eta}, \boldsymbol{\lambda})$. The response model $p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{\eta})$ is specified using a multivariate Normal distribution $\boldsymbol{y}^{\boldsymbol{r}} \sim \text{Normal}\,(\boldsymbol{\mu}^{\boldsymbol{r}}, \boldsymbol{\Sigma}^{\boldsymbol{r}})$, where $\boldsymbol{\mu}^{\boldsymbol{r}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{r}}$ denote the mean and covariance matrix parameters in each pattern. We focus our attention on the mean parameters $\boldsymbol{\mu}^{\boldsymbol{r}}$ and assume that the covariance matrix has a simplified form $\boldsymbol{\Sigma}^{\boldsymbol{r}} = \boldsymbol{\Sigma}$ that can be estimated from the data. Thus, the parameters of interest are given by $\boldsymbol{\mu}^{\boldsymbol{r}} = \left(\mu_j^{\boldsymbol{r}=(1,1)}, \mu_j^{\boldsymbol{r}=(0,1)}, \mu_j^{\boldsymbol{r}=(1,0)}\right)$, for $j = 0, 1$.
Because the responses are missing in certain patterns, not all parameters can be identified from the data. Equation 6.3 identifies $\boldsymbol{\mu}^{\boldsymbol{r}}$ through the extrapolation factorisation, which splits the response distribution into three components.

1 The distribution of the completers $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs} \mid \boldsymbol{\eta}_{\mathsf{C}})$, indexed by the parameters $\boldsymbol{\eta}_{\mathsf{C}} = (\mu_0^{\boldsymbol{r}=(1,1)}, \mu_1^{\boldsymbol{r}=(1,1)})$, includes all the responses $\boldsymbol{y}$ in $\boldsymbol{r} = (1, 1)$.

2 The distribution for the observed data among the non-completers $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs} \mid \boldsymbol{\eta}_{\mathsf{NC}})$, indexed by the parameters $\boldsymbol{\eta}_{\mathsf{NC}} = (\mu_0^{\boldsymbol{r}=(1,0)}, \mu_1^{\boldsymbol{r}=(0,1)})$, only includes the observed responses across $\boldsymbol{r} \neq \mathbf{1}$, i.e. $\boldsymbol{y}_0$ in $\boldsymbol{r} = (1, 0)$ and $\boldsymbol{y}_1$ in $\boldsymbol{r} = (0, 1)$.

3 The extrapolation distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{mis} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{\eta}_{\mathsf{E}})$, indexed by the parameters $\boldsymbol{\eta}_{\mathsf{E}} = (\mu_0^{\boldsymbol{r}=(0,1)}, \mu_1^{\boldsymbol{r}=(1,0)})$, only includes the missing responses across $\boldsymbol{r} \neq \mathbf{1}$, i.e. $\boldsymbol{y}_0$ in $\boldsymbol{r} = (0, 1)$ and $\boldsymbol{y}_1$ in $\boldsymbol{r} = (1, 0)$.

The parameters $\boldsymbol{\eta}_{\mathsf{C}}$ and $\boldsymbol{\eta}_{\mathsf{NC}}$ are estimated from fitting the model to the completers and the joint set of the non-completers patterns under ignorability, respectively. This ensures that inferences are obtained only based on the fit of the model to the observed data (which can be validated) without making any assumptions about the distribution of the missing values (which cannot be checked). Therefore, it is possible to perform model selection and choose the model which has the best fit to the observed cases among those assessed. The identification of the model can be then completed using a combination of identifying restrictions and sensitivity parameters to identify the parameters of the extrapolation distribution $\boldsymbol{\eta}_{\mathsf{E}}$. Different types of identifying restrictions can be used in practice, but they all have the objective to provide a convenient framework for the incorporation of sensitivity parameters into the model on which informative priors are typically specified to define the MNAR departures to explore. For example, in the given example, we can identify $\boldsymbol{\eta}_{\mathsf{E}}$ using the complete case missing value restrictions (Section 1.6.3) and some sensitivity parameters $\boldsymbol{\delta}$,

$$\boldsymbol{\eta}_{\mathsf{E}} = \boldsymbol{\eta}_{\mathsf{C}} + \boldsymbol{\delta}.$$

When $\boldsymbol{\delta} = \mathbf{0}$, the model is fully identified using only information from the observed data and some constraints (which cannot be verified from the data) to identify the parameters that index the missing data distribution (i.e. equal to those of the completers). However, these constraints can then be relaxed through the specification of informative priors on $\boldsymbol{\delta}$ to define a range of nonignorable departures, for example, by eliciting some external source of evidence (e.g. experts' opinion). Compared to alternative nonignorable models, e.g. selection models (Section 1.6), this approach do not rely on assumptions about the joint distribution of observed and missing data, which are often difficult to check and which make the identification of sensitivity parameters and the incorporation of external information into the model more problematic.
Within the given setting, each component in the extrapolation factorisation can be identified in relatively easy way. However, in real applications, analysts are typically faced with a more complex framework (e.g. multivariate outcomes, higher number of time points and missingness patterns) which makes the implementation of the model more challenging. Section 6.4 shows the application of the proposed approach to the PBS study.

Since the mean utilities and costs are the target of the inference, in our analysis we do not require the full identification of the extrapolation distribution. Instead, we only require the identification of the marginal means for the missing responses in each pattern. Thus, we identify the extrapolation distribution using a combination of partial identifying restrictions and sensitivity parameters (Section 1.6.3). We compute the marginal means $\mathsf{E}[\boldsymbol{y}^{\boldsymbol{r}}_{mis}]$ by averaging only with respect to the components of $\boldsymbol{y}^{\boldsymbol{r}}_{obs}$ across $\boldsymbol{r} \neq \mathbf{1}$, and then add some sensitivity parameters $\boldsymbol{\delta}$. We start by setting $\boldsymbol{\delta} = \mathbf{0}$ as a benchmark assumption and then explore the sensitivity of the results to alternative scenarios by using different prior distributions on $\boldsymbol{\delta}$, calibrated on the observed data. For example, we can choose the values of the hyperparameters of these priors in terms of the variability in the observed data (e.g. using the sample standard deviations) to define the extent of the departures from the benchmark. Once the working model has been fitted to the observed data and the extrapolation distribution has been identified, the overall marginal mean for the response model can be computed by marginalising over $\boldsymbol{r}$, i.e. $\mathsf{E}[\boldsymbol{y}] = \sum p(\boldsymbol{r})\mathsf{E}[\boldsymbol{y} \mid \boldsymbol{r}]$.

6.2 Observed Data Distribution

The observed data distribution $p(\boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r})$ is specified in terms of a model for the missingness patterns $\boldsymbol{r}$ and a model for the observed responses $\boldsymbol{y}^{\boldsymbol{r}}_{obs}$ in both the completers ($\boldsymbol{r} = \mathbf{1}$) and non-completers ($\boldsymbol{r} \neq \mathbf{1}$) patterns. We specify these in the following subsections.

6.2.1 Model for the missingness patterns

The model for the missingness patterns is specified as a multinomial distribution

$$\boldsymbol{r} \sim \text{Multinomial}\left(\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}\right), \tag{6.4}$$

which is defined on $\{1, \ldots, R_t\}$, with the total number of patterns R_t and the probabilities $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$ conditional on the treatment assignment t. Since the proportions of the individuals in the completers is considerably higher in both treatment groups of the PBS trial compared with any other pattern (Section 3.1.2), it seems reasonable to account for this characteristic of the data in the construction of the prior for $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$. We specify a prior on $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$ that gives more weight on the completers compared with the other patterns. For all non-completers, we assume equal prior weights since no information was available to justify the choice of pattern-specific weights, which are also expected to have limited impact on the posterior inferences of the model due to the sparsity of the data among the non-completers.

Therefore, we choose a Dirichlet $\left(1 - x, \frac{x}{R^\star}, \ldots, \frac{x}{R^\star}\right)$ prior, where x and $R^\star$ are the expected total dropout rate and the total number of potential patterns, respectively. In the PBS trial, the expected dropout rate is equal to 0.2 and is obtained from the original analysis plan of the study (Hassiotis et al., 2018), while the number of potential patterns is given by $2^{2J} = 64$. This prior choice is consistent with the design of the study, where the experimenter expects at least $(1-x)\%$ of the individuals to provide complete data, i.e. to fall in $\boldsymbol{r} = \mathbf{1}$. In practice, this prior is not likely to affect the results as the amount of observed data is enough to learn the posterior of $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$. For comparison purposes, we also consider another specification based on a noninformative Dirichlet$(1, \ldots, 1)$ prior for $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$ in Equation 6.4. Posterior results for the patterns' probabilities $\boldsymbol{\lambda}^{\boldsymbol{r}}_{t}$ and the marginal mean utilities and costs, which are shown in Figure C.13 and Figure C.14 in Appendix C.4.1, were robust to the alternative prior choices.

6.2.2 Model for the observed responses

The model for the observed responses $\boldsymbol{y}^{\boldsymbol{r}}_{obs}$ is specified separately for each treatment group t and extends the cross-sectional framework described in Section 5.1 to handle the utilities and costs within a longitudinal setting. In particular, the model accounts for three different types of complexities that affect the data.

First, the dependence both between each pair of outcome variables (u_{ij}, c_{ij}) and over time is accounted for using a series of conditional distributions. Second, skewness is handled assuming Beta and Log-Normal distributions for the utilities and costs, respectively. Since in the PBS study negative values for u_{ij} occur, as for the analysis in Section 5.7, utilities are scaled on $[0, 1]$ through the transformation $u^{\star}_{ij} = \frac{u_{ij} - \mathsf{min}(\boldsymbol{u}_j)}{\mathsf{max}(\boldsymbol{u}_j) - \mathsf{min}(\boldsymbol{u}_j)}$. The Beta distributions are then fitted to these transformed variables, which are referred to simply as u_{ij} to ease the notation. Third, to account for the structural values $u_{ij} = 1$ and $c_{ij} = 0$ a hurdle approach is specified by including in the model the corresponding indicator variables $d^u_{ij} := \mathbb{I}(u_{ij} = 1)$ and $d^c_{ij} := \mathbb{I}(c_{ij} = 0)$. These indicators take value 1 if subject i is associated with a structural value at time j and 0 otherwise. The probabilities of observing these values, as well as the mean of each variable, are then modelled conditionally on the utilities and costs at the current and previous times via linear regressions defined on the logit or log scale. Compared to the hurdle approach in Section 5.1, which only accounted for the structural ones in the QALYs, the model includes the indicator variables to handle the structural values in both the utility and cost outcomes. This is due to the fact that, within the cross-sectional framework of Section 5.7, where the modelled variables were c_{it}, no individual was associated with a null total cost. However, within a longitudinal framework, where the costs at each time point c_{ijt} are modelled, some individuals in both treatment groups (3 in the control and 2 in the intervention) are associated with a zero at some time point and a hurdle approach is then used to handle them.

The model can be summarised as follows (for simplicity we omit the treatment index t). At time $j = 0$, we model the nonzero costs and the indicator $d^c_{i0} := \mathbb{I}(c_{i0} = 0)$ as:

$$\begin{aligned} c_{i0} \mid d^c_{i0} = 0 &\sim \mathsf{LogNormal}\left(\phi^c_0, \tau^c_0\right) \\ d^c_{i0} &\sim \mathsf{Bernoulli}(\pi^c_0) \end{aligned} \tag{6.5}$$

where ϕ^c_0 and τ^c_0 are the mean and standard deviation for c_{i0} given $c_{i0} \neq 0$ on the log scale, while π^c_0 is the probability of a zero cost value. We next model the utilities and the indicator $d^u_{i0} := \mathbb{I}(u_{i0} = 1)$ conditionally on the costs at the same time:

$$\begin{aligned} u_{i0} \mid d^u_{i0} = 0, c_{i0} &\sim \mathsf{Beta}\left(\phi^u_{i0}\tau^u_{i0}, (1 - \phi^u_{i0})\tau^u_{i0}\right) \\ \mathsf{logit}(\phi^u_{i0}) &= \alpha_{00} + \alpha_{10} \log c_{i0} \\ d^u_{i0} \mid c_0 &\sim \mathsf{Bernoulli}(\pi^u_{i0}) \\ \mathsf{logit}(\pi^u_{i0}) &= \gamma_{00} + \gamma_{10} \log c_{i0} \end{aligned} \tag{6.6}$$

where ϕ^u_{i0} and τ^u_{i0} are the mean and precision for u_{i0} given $u_{i0} \neq 1$ and c_{i0}, while π^u_{i0} is the probability of having a utility value of one given c_{i0}. We use logistic transformations to define a linear dependence for $p(u_{i0} \mid c_{i0}, u_{i0} \neq 1)$ and include the costs on the log scale to avoid numeric instability in the estimation of the logistic regression parameters. This was due to the sparsity of the data (especially among the non-completers) and the very different range of values between the costs on the natural scale (in the order of thousands) and the probabilities.

At time $j = 1, 2$, we extend the approach illustrated in Equation 6.5 and Equation 6.6 for $j = 0$,

assuming a first-order Markov dependence structure. For the costs we have:

$$\begin{aligned} c_{ij} \mid d^c_{ij} = 0, c_{ij-1}, u_{ij-1} &\sim \text{LogNormal}\left(\phi^c_{ij}, \tau^c_j\right) \\ \phi^c_{ij} &= \beta_{0j} + \beta_{1j} \log c_{ij-1} + \beta_{2j} u_{ij-1} \\ d^c_{ij} \mid c_{ij-1}, u_{ij-1} &\sim \text{Bernoulli}(\pi^c_{ij}) \\ \text{logit}(\pi^c_{ij}) &= \zeta_{0j} + \zeta_{1j} \log c_{ij-1} + \zeta_{2j} u_{ij-1}. \end{aligned} \tag{6.7}$$

Similarly to time $j = 0$, the mean and standard deviation on the log scale and the probability parameters for the costs at time j are indicated with ϕ^c_{ij}, τ^c_j and π^c_{ij}. The regression parameters $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \beta_{2j})$ and $\boldsymbol{\zeta}_j = (\zeta_{0j}, \zeta_{1j}, \zeta_{2j})$ in Equation 6.7 capture the dependence between costs at j and the costs and utilities at $j-1$, for the non-zero and zero components, respectively. The model for the utilities is:

$$\begin{aligned} u_{ij} \mid d^u_{ij} = 0, c_{ij}, u_{ij-1} &\sim \text{Beta}\left(\phi^u_{ij}\tau^u_{ij}, (1-\phi^u_{ij})\tau^u_{ij}\right) \\ \text{logit}(\phi^u_{ij}) &= \alpha_{0j} + \alpha_{1j} \log c_{ij} + \alpha_{2j} u_{ij-1} \\ d^u_{ij} \mid c_{ij}, u_{ij-1} &\sim \text{Bernoulli}(\pi^u_{ij}) \\ \text{logit}(\pi^u_{ij}) &= \gamma_{0j} + \gamma_{1j} \log c_{ij} + \gamma_{2j} u_{ij-1}. \end{aligned} \tag{6.8}$$

We denote with ϕ^u_{ij}, τ^u_{ij} and π^u_{ij} the mean, precision and probability parameters for the utilities at time j. The regression parameters $\boldsymbol{\alpha}_j = (\alpha_{0j}, \alpha_{1j}, \alpha_{2j})$ and $\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \gamma_{2j})$ in Equation 6.8 capture the dependence between utilities at j and costs at j and utilities at $j-1$. Both Equation 6.7 and Equation 6.8 include costs as covariates on the log scale to avoid numeric instability in the estimation of the parameters of the logistic regressions, as in Equation 6.6.

For all relevant parameters in the model vague prior distributions are specified. Specifically we choose a Normal$(0, 1000)$ for the regression coefficients and a Uniform$(0, 10000)$ over a large positive range for the standard deviations (for the Beta distributions, priors on the standard deviations are specified as in Section 5.2.2). Figure 6.1 shows a graphical representation of the structure of the model considering only the baseline response pair (u_{i0}, c_{i0}) for simplicity.

First, the module for d^c_{i0} (brown box) is specified, upon which the two components c^0_{i0} and $c^{>0}_{i0}$ in the baseline costs are distinguished in the mixture module for c_{i0} (red box). Individuals associated with a non-zero cost are then marginally modelled using a LogNormal distribution. Next, the module for d^u_{i0} (green box) is specified conditional on the baseline costs, which allows to separate the two components u^1_{i0} and $u^{<1}_{i0}$ in the mixture module for u_{i0} (blue box). Individuals who are associated with a non-one utility are modelled using a Beta distribution conditionally on the baseline costs.

The modular structure of the framework is then repeated for $j = 1, 2$ and the time dependence between the response is captured through a series of conditional specifications for the utility and cost mixture and hurdle modules. The flexibility of this modelling strategy for the observed data distribution allows to account for all the typical complexities of the data: correlation between utilities and costs, skewness and spikes in the outcome variables.

The marginal cost and utility means at each time j are obtained through Monte Carlo integration. First, we fitted the model separately to the completers ($\boldsymbol{r} = \mathbf{1}$) and the joint set of all other patterns ($\boldsymbol{r} \neq \mathbf{1}$) for $t = 1, 2$. Second, at each iteration of the MCMC output, we draw $l = 1, \ldots, L$ replications $\boldsymbol{y}_{jl} = (c_{jl}, u_{jl})$ from their sampling distributions given the current value of the relevant parameters, with $L = 40,000$. Third, we approximated the posterior distribution of the marginal means for $\boldsymbol{y}^{\boldsymbol{r}}_j$ by taking the expectation over these sampled values at each iteration, i.e. $\text{E}[\boldsymbol{y}^{\boldsymbol{r}}_j] = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{y}_{jl}$. Finally, we derived the overall marginal means in each treatment group $\boldsymbol{\mu}_{jt} = (\mu^c_{jt}, \mu^u_{jt})$ as weighted averages across the marginal means in each pattern, using the posterior $\boldsymbol{\lambda}^{\boldsymbol{r}}_t$ as weights.

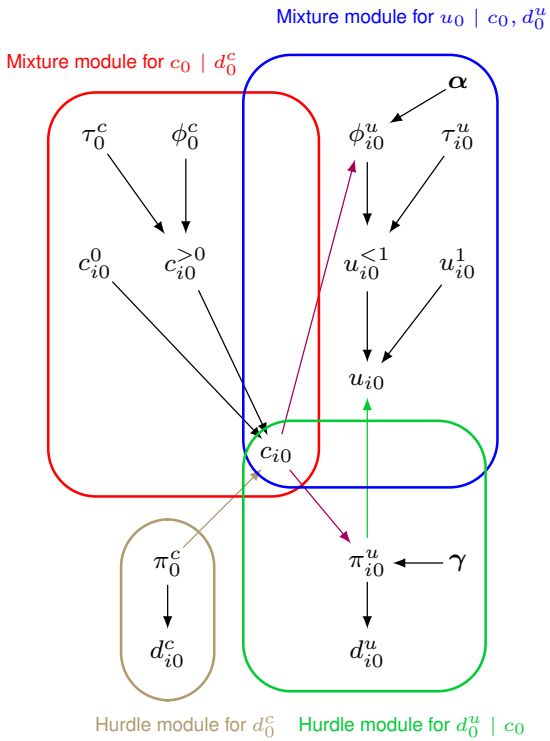


Figure 6.1: Four modules form the framework related to c_{i0} and u_{i0}. The two boxes at the bottom indicate the hurdle modules for d_{i0}^c and d_{i0}^u, which separates the structural (c_{i0}^0, u_{i0}^1) and non-structural $(c_{i0}^{>0}, u_{i0}^{<1})$ values in c_{i0} and u_{i0}, respectively indicated with the brown and green colours. The two boxes at the top are the mixture modules for c_{i0} and u_{i0} in which the non-structural values are modelled using LogNormal and Beta distributions, respectively indicated with the red and blue colours. The solid black arrows show the dependence relationships between the parameters within the modules, while the coloured arrows show the dependence between the parameters of different modules.

Once the marginal utility and cost means are derived from the model using Monte Carlo integration, these are used to compute the population QALYs and total cost mean parameters in each treatment group

$$\mu_{et} = \sum_{j=1}^{J} \left(\mu_{jt}^u + \mu_{j-1t}^u\right) \frac{\delta_j}{2} \quad \text{and} \quad \mu_{ct} = \sum_{j=1}^{J} \mu_{jt}^c, \tag{6.9}$$

where $\delta_j = 0.5$ for each j in the PBS study. The population means in Equation 6.9 are effectively derived using the formulae typically applied to u_{ij} and c_{ij} for computing the QALYs and total cost variables (Section 1.2) to the mean estimates μ_{jt}^u and μ_{jt}^c.

6.3 Extrapolation Distribution

Partial identifying restrictions are used to identify the posterior of the marginal means of the extrapolation distribution. We only require the identification of the marginal means for the missing responses in each pattern because the economic analysis is exclusively based on quantities derived from the cost and utility marginal means.

6.3.1 Partial Identifying Restrictions and Sensitivity Parameters

Sensitivity parameters $\boldsymbol{\delta}$ are embedded within the restrictions to assess the impact of alternative missingness assumptions on the quantities of interest. We choose $\boldsymbol{\delta}_j = (\delta_{c_j}, \delta_{u_j})$ to be time-specific location shifts at the utility and cost means in each pattern. At each time j, the marginal mean of the missing data in each pattern $\boldsymbol{y}_{mis,j}^{\boldsymbol{r}}$ is identified by averaging across the observed components at the same time point $\boldsymbol{y}_{obs,j}^{\boldsymbol{r}}$ for $\boldsymbol{r} \neq \mathbf{1}$ and adding the sensitivity parameters $\boldsymbol{\delta}_j$. More

formally, the mean of the missing responses is identified as

$$\mathsf{E}\left[\boldsymbol{y}^{\boldsymbol{r}}_{mis,j} \mid \boldsymbol{r} \neq \mathbf{1}\right] = \mathsf{E}\left[\boldsymbol{y}^{\boldsymbol{r}}_{obs,j} + \boldsymbol{\delta}_j \mid \boldsymbol{r} \neq \mathbf{1}\right], \tag{6.10}$$

for $j \in \{0, 1, 2\}$. As a reasonable benchmark assumption we set $\boldsymbol{\delta}_j = \mathbf{0}$. Under this scenario, the mean of the missing response in Equation 6.10 is obtained using only the mean parameters estimated from the model, averaged across the non-completers. We then use alternative informative priors on $\boldsymbol{\delta}_j$ to explore plausible departures for missing utilities and costs and assess their impact on cost-effectiveness conclusions.

In the PBS study, no external source of information was available to inform the direction or the magnitude of these departures. We formulate the assumptions about the missing values based on a discussion with the people involved in the original analysis of the trial (direction of departures) and on what we deem to be some plausible scenarios (magnitude of departures). According to these considerations, we assume that subjects with a missing value at time j have a lower utility and a higher cost compared with those who are observed at the same time but were not a completer. We then calibrate the priors on $\boldsymbol{\delta}_j$ using the observed standard deviations for costs and utilities at each time j to define the amplitude of the departures from $\boldsymbol{\delta}_j = \mathbf{0}$.

6.3.2 Priors on the Sensitivity Parameters

We consider three alternative sets of priors on $\boldsymbol{\delta}_j = (\delta_{u_j}, \delta_{c_j})$, calibrated based on the variability in the observed data at each time j. The three types of priors used are the following:

- $\boldsymbol{\delta}^{\text{flat}}$: Flat between 0 and twice the observed standard deviation:

$$\delta_{c_j} \sim \text{Uniform}[0, 2\ \text{sd}(c_j)] \quad \text{and} \quad \delta_{u_j} \sim \text{Uniform}[-2\ \text{sd}(u_j), 0]$$

- $\boldsymbol{\delta}^{\text{skew0}}$: Skewed towards values closer to 0, over the same range as $\boldsymbol{\delta}^{\text{flat}}$:

$$\delta_{c_j} = 2\ \text{sd}(c_j)\left(1 - \sqrt{U}\right) \quad \text{and} \quad \delta_{u_j} = -2\ \text{sd}(u_j)\left(1 - \sqrt{U}\right)$$

- $\boldsymbol{\delta}^{\text{skew1}}$: Skewed towards values far from 0, over the same range as $\boldsymbol{\delta}^{\text{flat}}$:

$$\delta_{c_j} = 2\ \text{sd}(c_j)\left(\sqrt{U}\right) \quad \text{and} \quad \delta_{u_j} = -2\ \text{sd}(u_j)\left(\sqrt{U}\right),$$

where $U \sim \text{Uniform}(0, 1)$ and $\text{sd}(u_j)$ and $\text{sd}(c_j)$ are the standard deviations computed on the observed utilities and costs at time j for $\boldsymbol{r} \neq \mathbf{1}$. Although alternative priors could be considered, since no formal experts' opinion or other information about missingness was collected in the PBS trial, the proposed priors for $\boldsymbol{\delta}_j$ were proposed based on a discussion with the people involved in the trial. In particular, these priors were chosen as departures from $\boldsymbol{\delta}_j = 0$ for both outcomes were believed unlikely to be larger than twice the observed standard deviations at each time j. Figure 6.2 shows a graphical representation of the densities of the three types of priors using the sensitivity parameters $\boldsymbol{\delta}_1 = (\delta_{u_1}, \delta_{c_1})$ as examples.

The three distributions for $\boldsymbol{\delta}_1$ are defined over the same range of values (negative for utilities and positive for costs) but assign different weights to the values within the range. The distribution of $\boldsymbol{\delta}^{\text{flat}}$ (panels a-b) assign the same weight to all values for both outcomes. By contrast, the distributions of $\boldsymbol{\delta}^{\text{skew0}}$ (panels c-d) and $\boldsymbol{\delta}^{\text{skew1}}$ (panels e-f) give more weights to values closer and far from zero and respectively express more and less conservative assumptions about the departures from the benchmark $\boldsymbol{\delta}_1 = 0$. Overall, the prior and posterior distributions of $\boldsymbol{\delta}_j$ were similar under each nonignorable scenario assessed for both treatment groups in the trial. The comparison between the prior and posterior estimates of $\boldsymbol{\delta}_j$ are shown in Figure C.15, Figure C.16

and Figure C.17 (control) and Figure C.18, Figure C.19 and Figure C.20 (intervention), which are provided in Appendix C.4.2.

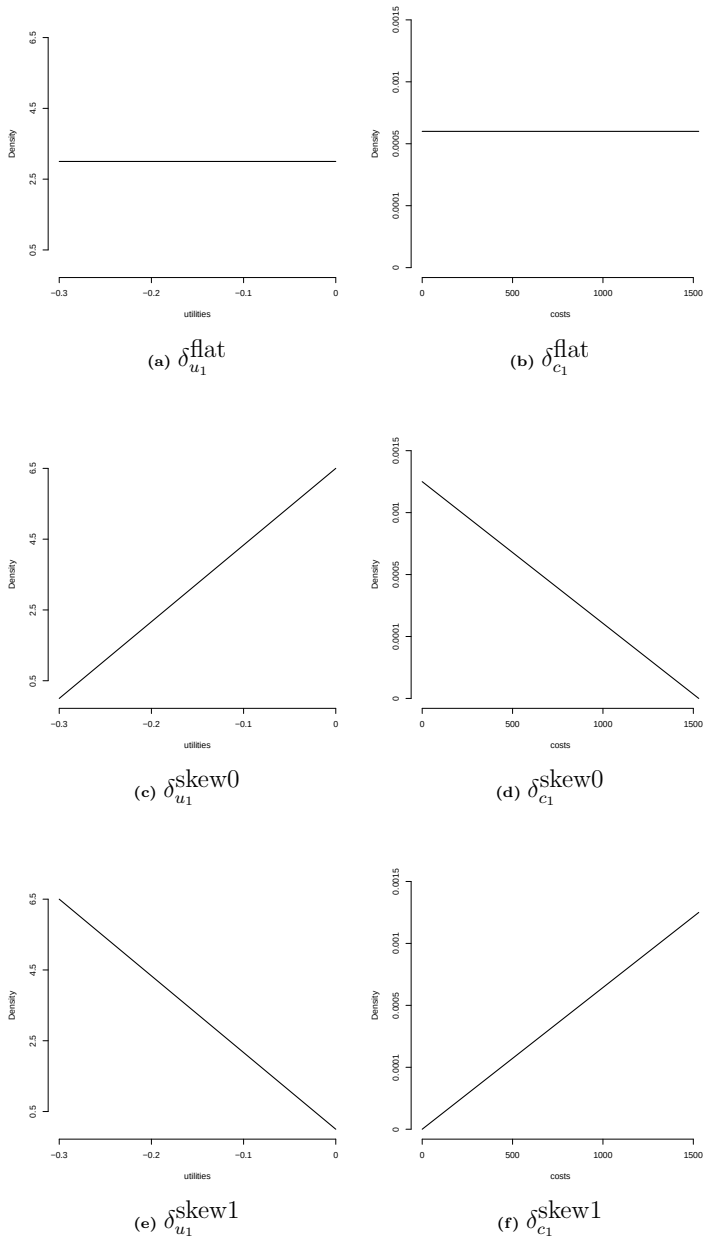


Figure 6.2: Densities of the prior distributions of the sensitivity parameters that identify the means for the missing utility and cost variables at $j = 1$ under three alternative scenarios: $\boldsymbol{\delta}^{\text{flat}}$ (panels a-b), $\boldsymbol{\delta}^{\text{skew0}}$ (panels c-d) and $\boldsymbol{\delta}^{\text{skew1}}$ (panels e-f)

6.4 Application to the PBS study

Among the non-completers ($\boldsymbol{r} \neq \mathbf{1}$), we set to 0 the regression parameters (ζ_{11},ζ_{21}) and (γ_{10},γ_{11},γ_{21}) for the model fitted to the control and intervention group, respectively. This simplification was required because, among the non-completers, there is only one observed $c_j = 0$ at time $j = 1$ in the control group and one observed $u_j = 1$ at time $j = \{0, 1\}$ in the intervention group. We therefore drop from the model the dependence between the probabilities of having a structural value at these times and the variables at the previous or same times to ensure the convergence of the algorithm and avoid identifiability problems. The full `JAGS` code of the model fitted to the

PBS data is provided in Appendix B.3.1. The R code used to derive the mean estimates under each nonignorable scenario is available in Appendix B.3.2.

The model was fitted using JAGS, which is interfaced with R through the package R2jags. We ran two chains with 20,000 iterations per chain, using a burn-in of 5,000, for a total sample of 30,000 iterations for posterior inference (without any thinning). Convergence and autocorrelation of the MCMC simulations were assessed using the same diagnostic measures and criteria described in Section 5.4.1. These include a value for the potential scale reduction factor below 1.05 and an effective sample size of at least 20,000 for the model parameters. Convergence was also assessed with respect to the use of different overdispersed values for the parameters, which were generated from the priors. For the model fitted to the non-completers, some of the model parameters were poorly identified (e.g. the logistic regression parameters for the hurdle models estimated using very few numbers of structural values at some time point). For some of these parameters, although an effective sample size lower then 20,000 were observed, however, the potential scale reduction factor was always below 1.05 and it was therefore assumed that they did not affect the overall convergence the model. The total computational time required for the model based to produce representative samples from the posterior distributions of interest was about 50 minutes.

6.4.1 Model Assessment

We perform model checking to evaluate the adequacy of the fit of the model to the observed data. Specifically, we use the DIC to assess the fit of the model with respect to an alternative parametric specification, where the LogNormal distributions are replaced with Gamma distributions for the cost variables. In our analysis, we consider a DIC on the observed data under MAR as its value does not depend on the values of the sensitivity parameters. Because the sampling distribution of the observed data was not available in closed form, Monte Carlo integration was used to compute it. Results between the two alternative specifications are reported for each modelled variable in Table 6.2. The total values for both DIC and p_D are reported at the bottom of the table.

	Gamma		LogNormal	
variable	DIC	p_D	DIC	p_D
c_0	2147.91	2.05	*2133.39*	1.97
$u_0 \mid c_0$	-377.52	2.87	-377.62	2.82
$c_1 \mid c_0, u_0$	1904.53	4.16	*1827.45*	4.13
$u_1 \mid u_0, c_1$	-468.02	5.37	-468.19	5.32
$c_2 \mid c_1, u_1$	1913.69	4.65	*1856.23*	4.36
$u_2 \mid u_1, c_2$	-454.07	5.87	-453.47	5.99
Total	4667	25	*4518*	25

Table 6.2: DIC and p_D based on the observed data likelihood for each variable in the model. Two models are assessed either assuming LogNormal or Gamma distributions for the cost variables. Total DIC and p_D are also reported at the bottom of the table.

The DIC components for the costs are systematically lower when LogNormal distributions are used compared with Gamma distributions (lower values shown in italics in Table 6.2), and result in an overall better fit to the data for the first model.

Additionally, we assess the absolute fit of the model using posterior predictive checks based on observed data replications. A total of $40,000$ samples for the responses and missingness patterns were drawn from the posterior predictive distribution of the model $p(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{r}} \mid \boldsymbol{y}^{\boldsymbol{r}}_{obs}, \boldsymbol{r}, \boldsymbol{\omega})$. Conditional on the replicated patterns $\tilde{\boldsymbol{r}}$, the replicated observed data in each pattern are defined

as $\tilde{\boldsymbol{y}}^{\tilde{\boldsymbol{r}}}_{obs} = \{\tilde{\boldsymbol{y}}_j : \tilde{\boldsymbol{r}}_j = \mathbf{1}\}$, that is the components of $\tilde{\boldsymbol{y}}$ for which the corresponding missing data indicators at time j in the replicated patterns $\tilde{\boldsymbol{r}}$ are equal to one.

We compute the rank correlations between each pair of variables for each replicated dataset, and compare them with the corresponding values from the real dataset. The results, shown in Figure 6.3, suggest that the proposed parametric model captures most of the correlations well both in the control (panel a) and intervention (panel b) group.

Finally, we assess the fit of the model to the empirical distributions of the observed utility and cost data. Figure 6.4 compares the posterior predictive utility and cost densities at each time j based on the samples generated from the model (light blue lines) with the empirical distributions of the observed cases (dark blue lines) in both the control (panels a and c) and intervention (panels b and d) group of the PBS trial. With the only exception of the utilities at $j = 2$ in the intervention group, the replicated samples closely approximate the empirical distributions of both utilities and costs at each time point.

6.5 Results

6.5.1 Scenarios

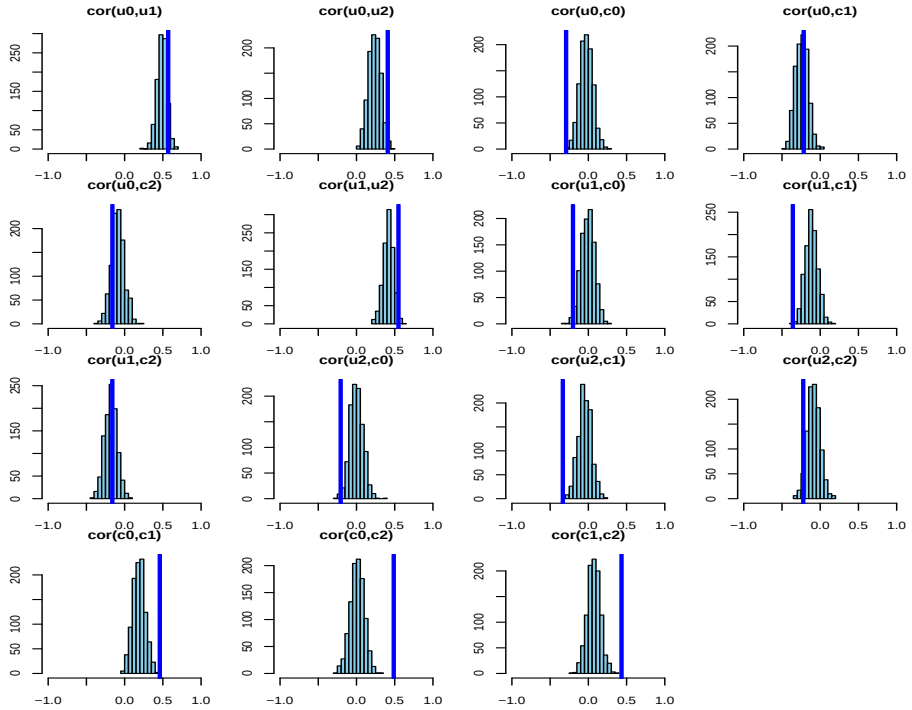
Table 6.3 summarises the different scenarios compared in the economic analysis of the PBS study, derived either from the proposed approach or from alternative models. Within the modelling

model	framework	patterns	missingness
$\boldsymbol{\delta}^{\text{skew1}}$	longitudinal	$\boldsymbol{r} = \mathbf{1}$ & $\boldsymbol{r} \neq \mathbf{1}$	nonignorable
$\boldsymbol{\delta}^{\text{skew0}}$	longitudinal	$\boldsymbol{r} = \mathbf{1}$ & $\boldsymbol{r} \neq \mathbf{1}$	nonignorable
$\boldsymbol{\delta}^{\text{flat}}$	longitudinal	$\boldsymbol{r} = \mathbf{1}$ & $\boldsymbol{r} \neq \mathbf{1}$	nonignorable
$\boldsymbol{\delta} = \mathbf{0}$	longitudinal	$\boldsymbol{r} = \mathbf{1}$ & $\boldsymbol{r} \neq \mathbf{1}$	nonignorable
L-ALL	longitudinal	all	ignorable
L-CC	longitudinal	$\boldsymbol{r} = \mathbf{1}$	ignorable
CS-ALL	cross-sectional	all	ignorable
CS-CC	cross-sectional	$\boldsymbol{r} = \mathbf{1}$	ignorable

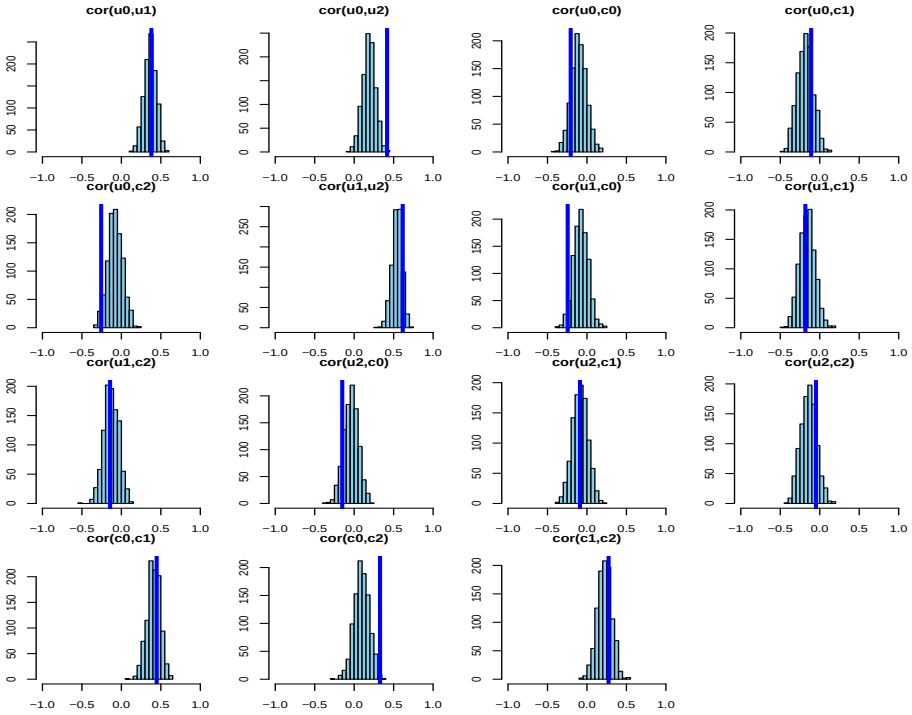
Table 6.3: List of the scenarios compared in the analysis of the PBS study. These are distinguished by modelling framework (either longitudinal or cross-sectional), the patterns to which the models were fitted (either $\boldsymbol{r} = \mathbf{1}$, $\boldsymbol{r} = \mathbf{1}$ & $\boldsymbol{r} \neq \mathbf{1}$ or all patterns jointly), and the assumptions about missingness (either ignorable or nonignorable). All longitudinal models and CS-ALL are specified using Beta-Lognormal distributions with a hurdle approach for both outcomes. CS-CC is specified using the "standard approach" (independent Normal distributions for both outcomes).

framework of Section 6.1, the alternative missingness scenarios explored are indicated with $\boldsymbol{\delta} = \mathbf{0}$ (benchmark) and $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$ (nonignorable departures). In addition, we consider four scenarios. L-ALL and L-CC are two longitudinal models, specified as in Section 6.2.2, but fitted jointly across all patterns under MAR and to the completers, respectively. CS-ALL and CS-CC are two cross-sectional models (i.e. the modelled variables are the QALYs/total costs rather than the utilities/costs): CS-ALL is specified using a Beta-LogNormal distribution with a hurdle approach to handle both unit QALYs and zero costs and is fitted to all cases under MAR, while CS-CC is specified using the "standard approach" (independent Normal distribution for both outcomes and fitted to the completers).

The last four scenarios are included to assess the impact that alternative types of modelling assumptions have on the results. Specifically, variations in the estimates may be due to either the use of a different framework, alternative distributions for the responses, the incorporation of the non-completers in the analysis or a combination thereof. It is important to distinguish how a change in each of these assumptions affects the results of the analysis. For example, we can obtain some information about the expected impact that the inclusion of the partially-observed



(a) control

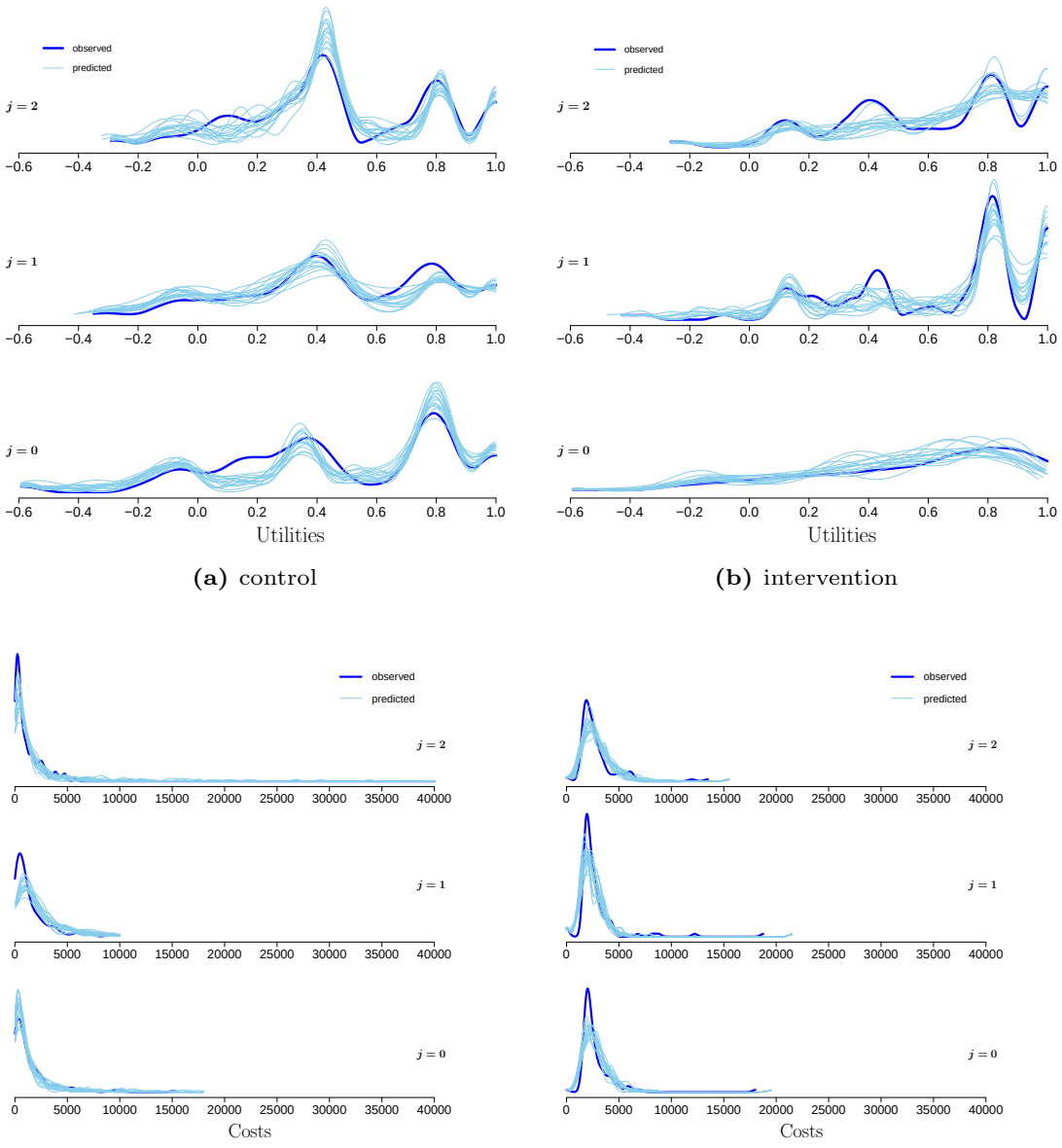


(b) intervention

Figure 6.3: Posterior predictive distributions for the pairwise correlation between utilities and costs variables in the control (panel a) and intervention (panel b) arm across 1000 observed replicated datasets (light blue bars) compared with the observed value in the real dataset (vertical blue lines).

data may have on the inferences by comparing the observed responses between the completers and non-completers in the PBS study.

Figure 6.5 shows the mean utility and cost profiles, calculated as a simple average of the observed data at each time point in the PBS study, for the individuals in the completers (solid



(a) control **(b)** intervention

(c) control **(d)** intervention

Figure 6.4: Posterior predictive utility and cost densities at each time $j = 0, 1, 2$ for the longitudinal model (light blue lines) compared with the empirical distributions of the complete cases (dark blue lines) in the control (panels a and c) and intervention (panel b and d) group in the PBS trial. For each variable, 1000 replications of the data are generated, of which only 15 are displayed in each graph for visualisation purposes

lines) and non-completers (dashed lines), either in the control (red) or intervention group (blue). In the control group (panels a and c), the completers are associated with consistently higher mean profiles with respect to the non-completers for both outcomes; the largest differences occur at the last follow-up for both the utilities (0.036) and the costs (£486). In the intervention group (panels b and d), the mean profiles of the two patterns intersect once throughout the study period; the largest differences are at baseline for both the utilities (0.037) and the costs (£703).

6.5.2 Utility/cost means

Figure 6.6 compares the posterior means and 95% HPD intervals for μ^u_{jt} (panel a) and μ^c_{jt} (panel b) across six alternative scenarios: L-CC, L-ALL, $\boldsymbol{\delta}_j = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$. Since baseline costs are fully observed, only the estimates under L-CC and L-ALL are shown for $\boldsymbol{\mu}^c_0$. Results

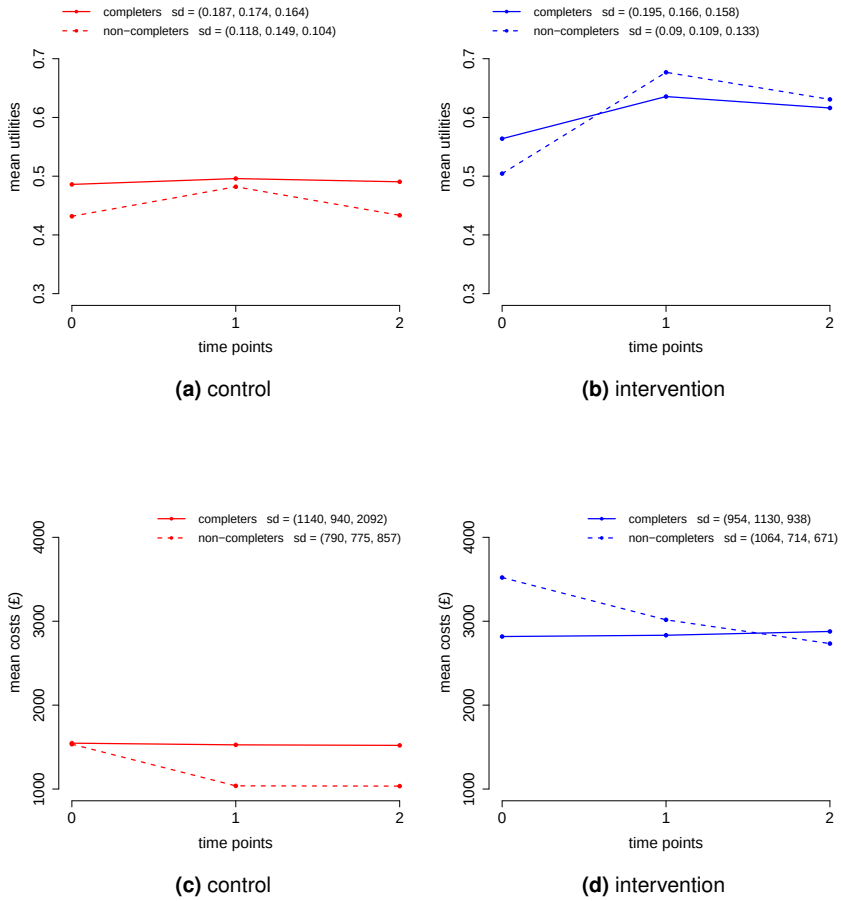


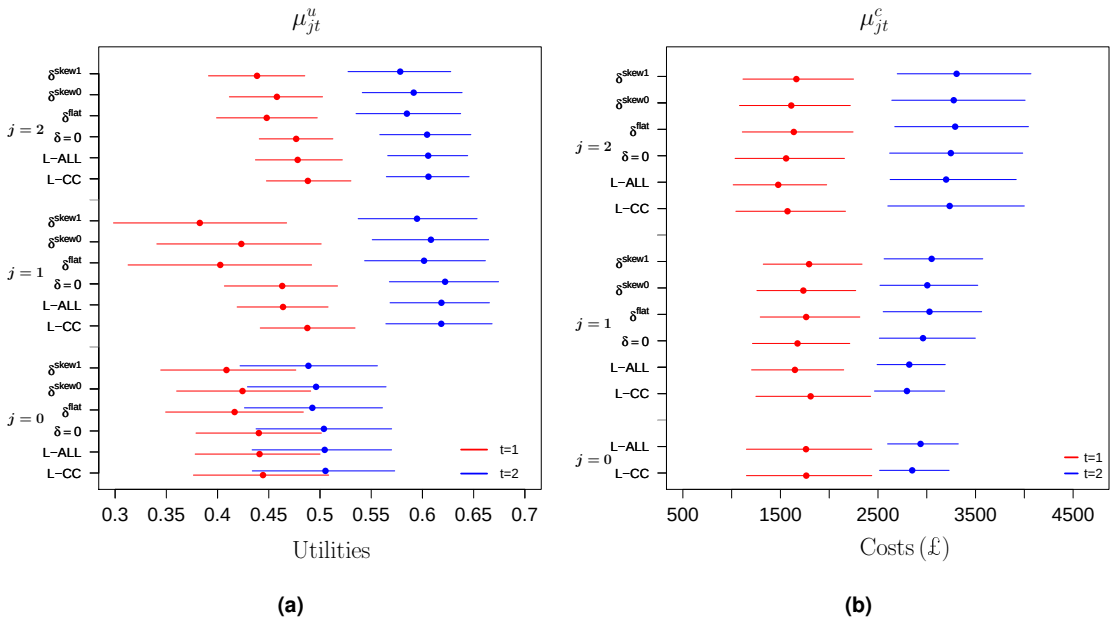
Figure 6.5: Observed mean utility and cost profiles in the PBS trial for the individuals in the completers (solid lines) and non-completers (dashed lines) patterns, either in the control (red) or intervention group (blue).

associated with the control and intervention group are indicated in red and blue, respectively. The distributions of both $\boldsymbol{\mu}_j^u$ and $\boldsymbol{\mu}_j^c$ show values that are higher in the intervention than in the control at each time j. Compared with L-CC, the estimates from L-ALL or $\boldsymbol{\delta} = \mathbf{0}$ show variations that are due to the additional incorporation of the non-completers in the analysis (Figure 6.5), either by fitting the model jointly across the patterns (L-ALL) or using the extrapolation factorisation ($\boldsymbol{\delta} = \mathbf{0}$).

With respect to the benchmark $\boldsymbol{\delta} = \mathbf{0}$, the impact of the alternative priors for the sensitivity parameters (see Section 6.3.2) on the inferences is reflected in the variations of the mean estimates across the nonignroable scenarios. In the control group, compared with $\boldsymbol{\delta} = \mathbf{0}$, mean utilities show an average decrease of 0.023 ($\boldsymbol{\delta}^{\text{flat}}$), 0.016 ($\boldsymbol{\delta}^{\text{skew0}}$) and 0.031 ($\boldsymbol{\delta}^{\text{skew1}}$), while mean costs show an average increase of £83 ($\boldsymbol{\delta}^{\text{flat}}$), £56 ($\boldsymbol{\delta}^{\text{skew0}}$) and £112 ($\boldsymbol{\delta}^{\text{skew1}}$). In the intervention group, mean utilities are on average 0.011 ($\boldsymbol{\delta}^{\text{flat}}$), 0.009 ($\boldsymbol{\delta}^{\text{skew0}}$) and 0.013 ($\boldsymbol{\delta}^{\text{skew1}}$) lower compared with $\boldsymbol{\delta} = \mathbf{0}$, while mean costs are on average £56 ($\boldsymbol{\delta}^{\text{flat}}$), £37 ($\boldsymbol{\delta}^{\text{skew0}}$) and £74 ($\boldsymbol{\delta}^{\text{skew1}}$) higher.

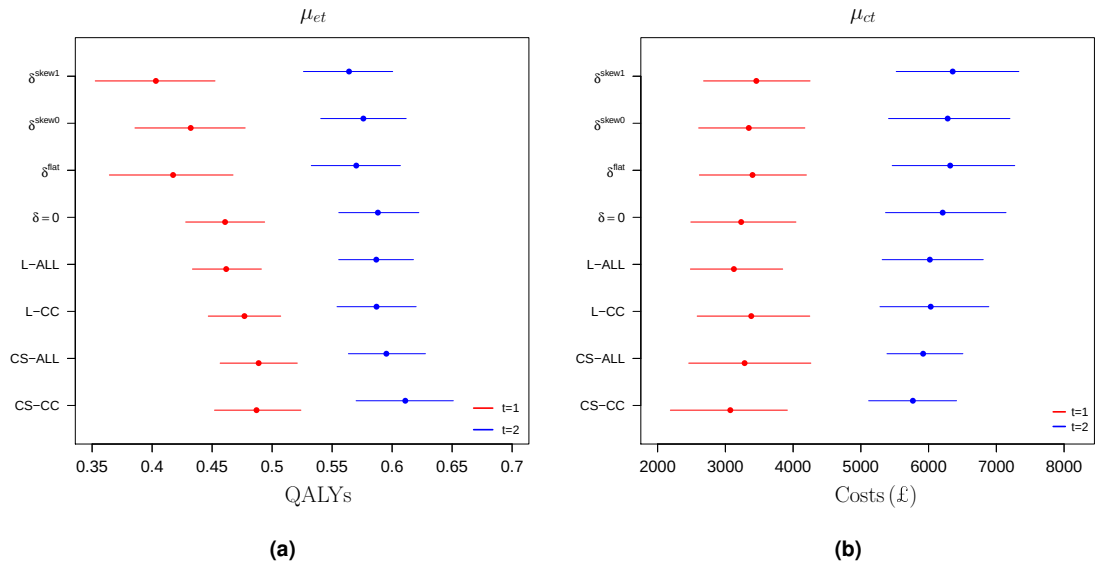
6.5.3 QALYs/total cost means

We additionally compare the estimates of the population QALYs and total cost means μ_{et} and μ_{ct}, derived for each scenario assessed in Section 6.5.2, with those from CS-ALL and CS-CC. Figure 6.7 shows the posterior means and 95% HPD intervals associated with the mean QALYs (panel a) and total cost parameters (panel b) under all scenarios for both the control (denoted with a red colour) and intervention (denoted with a blue colour) groups. Changes in the estimates



(a) **(b)**

Figure 6.6: Posterior means and 95% HPD intervals for the marginal utility and cost means in the control (red dots and lines) and intervention (blue dots and lines) group at each time j in the PBS study across alternative assumptions. Six scenarios are compared: L-CC,L-ALL, $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$, and $\boldsymbol{\delta}^{\text{skew1}}$. Since the baseline costs are fully observed in both groups, only the results under L-CC and L-ALL are displayed for $\boldsymbol{\mu}^c_0$.



(a) **(b)**

Figure 6.7: Posterior means and 95% HPD intervals for the marginal QALYs and total cost means in the control (red dots and lines) and intervention (blue dots and lines) group in the PBS study across alternative assumptions. Eight scenarios are compared: CS-CC, CS-ALL, L-CC, L-ALL, $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$.

are due to different types of modelling assumptions between the scenarios. For example, within a cross-sectional framework, explicitly accounting for the complexities of the data (CS-ALL) leads to almost no impact on the mean QALYs, with an increase of 0.001 for $t = 1$ and a decrease of 0.016 for $t = 2$, and small changes in the mean costs, with an increase of $£210$ for $t = 1$ and $£151$ for $t = 2$, compared with the standard approach (CS-CC). Under the same distributional specification, the change of the modelling framework from cross-sectional (CS-ALL) to longitudinal (L-CC) is associated with limited variations in the mean QALYs and costs in both treatment groups when the models are fitted to the same cases (completers only). However, when the additional

information from the observed data of the non-completers is incorporated into the longitudinal model, more substantial changes are observed in the estimates, especially for the QALYs in the control group. More specifically, the mean estimates derived from the longitudinal model fitted only to the completers (L-CC) are higher with respect to those from the model that additionally incorporates the evidence from the non-completers (L-ALL), both for the QALYs (increases of 0.09 for $t = 1$ and 0.01 for $t = 2$) and costs (increases of $£257$ for $t = 1$ and $£13$ for $t = 2$). These variations are larger in the control, which has higher proportions of missing data compared with the intervention, and where the average observed utilities and costs from the non-completers are systematically lower with respect to the completers (Figure 6.5).

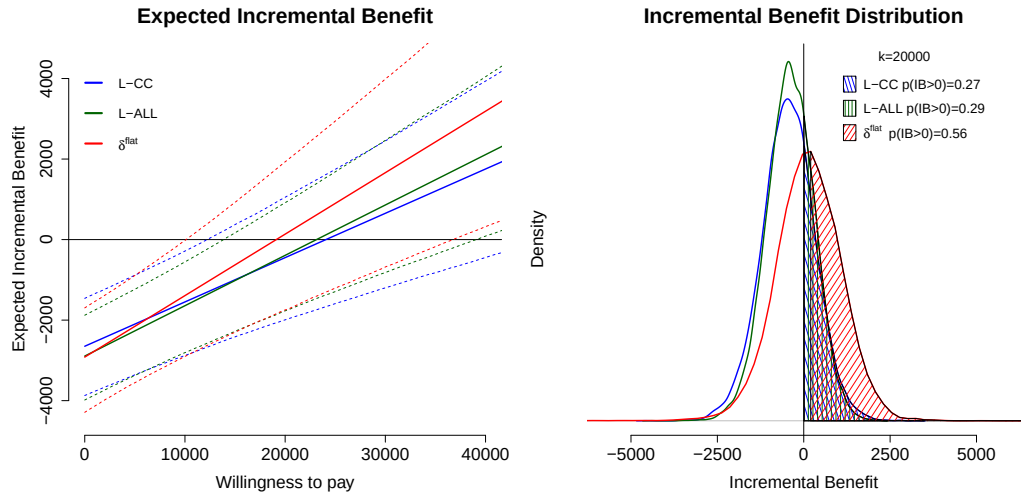
Finally, compared with the benchmark $\boldsymbol{\delta} = \mathbf{0}$, changes of similar amplitude and with the same sign to those observed for $\boldsymbol{\mu}_{jt}$ affect the QALYs and total cost estimates across the nonignorable scenarios $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$. In the control, compared with $\boldsymbol{\delta} = \mathbf{0}$, mean QALYs show decrements of 0.044 ($\boldsymbol{\delta}^{\text{flat}}$), 0.029 ($\boldsymbol{\delta}^{\text{skew0}}$) and 0.058 ($\boldsymbol{\delta}^{\text{skew1}}$), while mean costs show increments of $£167$ ($\boldsymbol{\delta}^{\text{flat}}$), $£112$ ($\boldsymbol{\delta}^{\text{skew0}}$) and £224 ($\boldsymbol{\delta}^{\text{skew1}}$), respectively.

In the intervention, the corresponding decreases in mean QALYs are 0.018 ($\boldsymbol{\delta}^{\text{flat}}$), 0.012 ($\boldsymbol{\delta}^{\text{skew0}}$) and 0.024 ($\boldsymbol{\delta}^{\text{skew1}}$), while the increases in mean total costs are $£110$ ($\boldsymbol{\delta}^{\text{flat}}$), $£73$ ($\boldsymbol{\delta}^{\text{skew0}}$) and $£147$ ($\boldsymbol{\delta}^{\text{skew1}}$). Table C.15, which is provided in Appendix C.4.3, reports the mean and 95% estimates associated with the mean QALYs and costs and other incremental quantities (e.g. net benefit) for each model considered.

6.6 Economic Evaluation

We complete the analysis by assessing the cost-effectiveness of the new intervention with respect to the control, comparing the results across the scenarios in Table 6.5. We specifically rely on the examination of the EIB and IB distribution, as well as the CEP and CEAC to summarise the economic analysis.

Figure 6.8 compares the estimates for the Expected Incremental Benefit and Incremental Benefit distribution obtained under the longitudinal models fitted to only the completers (L-CC), to all cases under MAR (L-ALL) and under the non-ignorable scenario $\boldsymbol{\delta}^{\text{flat}}$. The results for the other nonignorable scenarios are provided in Appendix C.4.4. For almost all the values of the willing-



(a) Expected Incremental Benefit

(b) Incremental Benefit Distribution

Figure 6.8: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with L-CC (blue solid and dashed lines), L-ALL (green solid and dashed lines) and $\boldsymbol{\delta}^{\text{flat}}$ (red solid and dashed lines) models fitted to the data from the PBS study.

ness to pay k, the estimates of the EIB and its lower/upper bounds (panel a) are similar between the L-CC and L-ALL scenarios, with a slightly steeper lines for L-ALL (green colour) compared with L-CC (blue colour), while those under $\boldsymbol{\delta}^{\text{flat}}$ are associated with an upward shift of the lines and the steepest slope (red colour) among the model assessed. This indicates that, under the non-ignorable scenario, the cost-effectiveness assessment is more favourable to the intervention compared with the models fitted under an ignorability assumption. Similar conclusions are obtained by looking at the estimated distribution of the IB (panel b), evaluated at $k = £20,000$, where the probability of cost-effectiveness is higher under $\boldsymbol{\delta}^{\text{flat}}$ (0.56) with respect to both L-ALL (0.29) and L-CC (0.27).

The CEP plot (Figure 6.9, panel a) shows the results only under the same scenarios (light blue for L-CC, light green for L-ALL and light red for $\boldsymbol{\delta}^{\text{flat}}$) for clarity and visualisation purposes. The results for the other nonignorable scenarios are provided in Appendix C.4.4. For all three

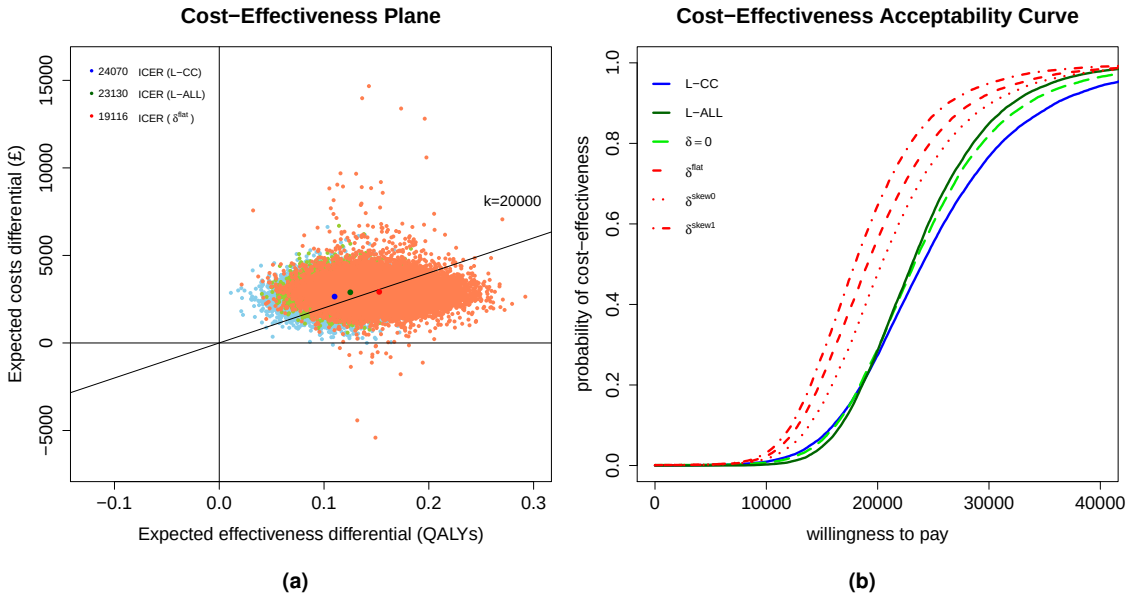


Figure 6.9: CEPs (panel a) and CEACs (panel b) associated with alternative missingness scenarios. In the CEPs, the ICERs based on the results from L-CC, L-ALL and $\boldsymbol{\delta}^{\text{flat}}$ are indicated with corresponding darker coloured dots. For the CEACs, in addition to the results from L-CC and L-ALL (solid lines), the probability values for the alternative scenarios are represented with different coloured dashed lines.

scenarios almost all samples fall in the North-East quadrant and are associated with positive ICERs. The ICER under $\boldsymbol{\delta}^{\text{flat}}$ falls in the sustainability area and indicates a more positive cost-effective assessment for the new intervention compared with L-CC and L-ALL.

The CEAC plot (Figure 6.9, panel b) shows the results under L-CC and L-ALL using blue and green solid lines, respectively. In addition, the results derived under nonignorability are reported using different coloured dashed lines. The CEACs under L-CC, L-ALL and $\boldsymbol{\delta} = \mathbf{0}$ show a similar trend and indicate a probability of cost-effectiveness that is systematically lower compared with the other nonignorable scenarios $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$ for most values of k. The CEAC plot shows that results are sensitive to the assumptions about the missing values, which can lead to a considerable change in the output of the decision process and the cost-effectiveness conclusions.

We finally evaluate the results derived from the models that ignore the information from the non-completers (L-CC, CS-ALL, CS-CC) and compare them with those from the models that incorporate this additional evidence (L-ALL and the nonignorable scenarios). Figure 6.10 shows the CEPs (panel a) associated with L-CC, CS-ALL and CS-CC, respectively indicated with blue, green and red coloured dots. In the CEACs (panel b), in addition to the probability values associated with these scenarios (L-CC – solid blue line, CS-ALL – green dashed line, CS-CC – red dashed line), the results from L-ALL and $\boldsymbol{\delta}^{\text{flat}}$ are indicated with solid green and red lines, respectively.

In the CEP plot (Figure 6.10, panel a), the posterior samples for all three scenarios fall almost

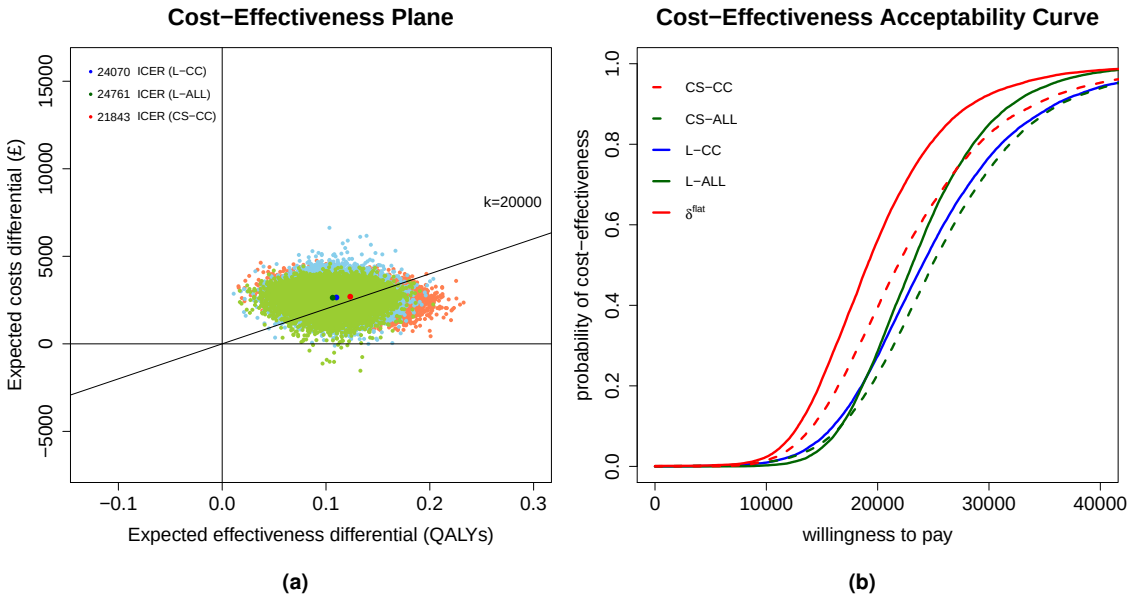


Figure 6.10: CEPs (panel a) and CEACs (panel b) associated with alternative scenarios. In the CEPs, the ICERs based on the results from the L-CC, CS-ALL, CS-CC and L-ALL are indicated with corresponding darker coloured dots. For the CEACs, in addition to the results under CS-ALL and CS-CC (dashed lines), the probability values for L-CC, L-ALL and δ^{flat} are represented with solid lines.

entirely in the North-East quadrant and, at $k =$ £25,000, indicate a mild cost-effectiveness of the new intervention compared with the control. However, ICERs under CS-ALL and L-CC lie at the boundary of the sustainability area and are both higher compared with the ICER under CS-CC.

In the CEAC plot (Figure 6.10, panel b), the acceptability curves under CS-ALL and L-CC show minimal discrepancies. This is due to the fact that, even though the two models are respectively specified within a cross-sectional and longitudinal setting, they are both fitted only to the completers and account for the same number of complexities in the data. Finally, the CEAC for CS-CC is higher compared with L-CC, CS-ALL and L-ALL for most willingness to pay values, but remains systematically lower with respect to δ^{flat}, δ^{skew0} and δ^{skew1} (in Figure 6.10 we only show the results for δ^{flat} for clarity).

6.7 Discussion

In this chapter, we have proposed a new approach for conducting parametric Bayesian inference under nonignorable missingness for a longitudinal bivariate outcome in economic evaluations. The model uses the extrapolation factorisation, within a pattern mixture approach, to specify a model for the distribution of the observed responses at each time point in the study and identify the mean of the missing responses using a combination of partial identifying restrictions and sensitivity parameters.

We motivate and apply our modelling framework to the data from the PBS study. Both inferences and cost-effectiveness conclusions are sensitive to the three alternative nonignorable scenarios explored and suggest a more cost-effective intervention compared with the results obtained under ignorability. This may be due to the fact that the proportions of partially-observed utility and cost data for the individuals in the control group are systematically higher with respect to those in the intervention group (see Table 3.4). This may occur, for example, when the intervention group is more effective in preventing people to drop out from the study, therefore ensuring higher completeness rates in the outcomes than the comparator. Therefore, if the individuals who have

missing values are associated with lower QALYs and higher costs compared to those observed in both groups, it is possible that models based on an ignorable mechanism assumption will tend overestimate the cost-effectiveness of the control with respect to the intervention.

The results from the analysis of the PBS study within a longitudinal framework show some differences compared with those derived from cross-sectional models. These differences become considerable when, as in the original analysis of the trial, the model does not account for all the complexities of the data. This indicates that the statistical methods used by practitioners in routine analyses are likely to be biased and can mislead the cost-effectiveness assessment.

Our approach provides some important advantages when dealing with missing data compared with the standard approach used in trial-based CEAs. First, efficiency is improved since the information from all observed data in each missingness pattern is retained when fitting the model, rather than using only the data from the completers. Second, the model accounts for the time dependence between the responses at different follow-ups, which may provide useful information when imputing the missing values. Third, the complexities affecting the outcome data, such as correlation, skewness and presence of structural values in both utilities and costs, are accounted for through a flexible parametric specification for the observed data distribution. Finally, sensitivity analysis is conducted to assess the robustness of the results to a range of plausible nonignorable missingness assumptions for both utility and cost data. These are specified using informative priors on the sensitivity parameters, calibrated from the observed data.

The model improves and expands the statistical approach used in routine CEAs and can help practitioners to avoid biased results, which may in turn lead to misleading cost-effectiveness conclusions. The availability of methodological and practical tools, such as those presented in this chapter, has the potential to improve the work of modellers and regulators alike, thus advancing the fields of economic evaluation of health care interventions.

Key points of this chapter:

- Our longitudinal model improves current CEA practice by accounting for all observed data in the study, the dependence between outcomes and over time, and the typical complexities that affect the data. In addition, the model assesses the robustness of the results to plausible nonignorable missingness assumptions.
- We demonstrated the benefits of using our approach with the PBS study as a motivating example. We used the extrapolation distribution and a pattern mixture approach to separately fit a flexible parametric model to the observed data, and identify the distribution of the missing data using partial identifying restrictions and sensitivity parameters.
- Marginal mean utility and cost estimates were derived using Monte Carlo integration and then averaging across the pattern-specific estimates. The target population mean QALYs and total cost parameters were then derived by combining the utility and cost means estimated from the model at each time point.
- The fit of the model to the observed data was checked via relative (DIC) and absolute (PCCs) predictive accuracy measures, both suggesting that the model well-captures the characteristics of the observed data. Alternative nonignorable assumptions were incorporated into the model using informative priors on the sensitivity parameters, calibrated on the observed data.
- We compared the results from our model with those from a standard cross-sectional model. Both inferences and cost-effectiveness conclusions were sensitive to the missingness scenarios assessed and indicated a bias associated with the methods used in routine analyses.

Chapter 7

Conclusions and Extensions

In this chapter, we draw together the findings, strengths and limitations of the thesis with respect to the three objectives that constitute the research questions of this work and suggest avenues of future research.

7.1 Summary

The statistical methods implemented in the analyses of the two case studies in this thesis have the objective of providing practical tools to improve the current practice in routine analyses. The methods illustrated follow the general NICE recommendations for conducting within-trial CEAs (NICE, 2013). These include the use of statistical methods that explicitly account for the typical complexities of the data, such as correlation, clustering, skewness, structural values and missingness. In addition, results of the economic evaluations are assessed with respect to both sampling variability and parameter uncertainty (including missing data uncertainty). The first is addressed by exploring and comparing the fit of alternative model specifications to the observed QALYs and costs and how they affect the posterior estimates and cost-effectiveness results. The choice of the parametric distributions for each outcome variable is based on those typically used in the published literature, such as Normal or Beta for QALYs and Normal, Gamma or LogNormal for costs, while model assessment is performed using the DIC or posterior predictive checks as reference approaches to evaluate model fit within the Bayesian framework.

The second is addressed through prior distributions to allow the full propagation of parameter uncertainty in the model, which in turn is directly reflected into variations in the posterior quantities of interest (e.g. mean incremental QALYs and costs) and accounts for the possible dependence between the parameters of the model. The sensitivity of the results to alternative weakly informative prior specifications was assessed to ensure that no unintended informative content in the priors was introduced into the model. Finally, the potential impact of missing data uncertainty on the results is assessed either under MAR and some simple departures from it within a cross-sectional framework or using a proper non-ignorable model via informative priors on the sensitivity parameters to conduct sensitivity analysis to a range of plausible MNAR departures.

In the following sections we discuss how the different objectives of the thesis have been addressed and provide advice to practitioners for addressing more general situations which were not relevant to the specific studies analysed but that may occur in practice.

7.1.1 Objective 1: Literature Review

The main added value provided by the literature review in Chapter 2 is represented by the Quality Evaluation Scheme, a structural tool to qualitatively assess the information reported about missing data in CEAs. The appealing characteristic of the quality evaluation scheme is the possibility to assign scores and to rank the studies based on a list of criteria that summarise the current literature recommendations about the collection and reporting of missing data information. Authors can evaluate the quality evaluation scheme score to assess the quality of their own analysis with respect to the amount and type of information provided on missingness. This would lead to a more comparable formalisation of the missing data uncertainty as well as a better indication of possible issues in assessing the cost-effectiveness of new treatments.

The application of the quality evaluation scheme to the reviewed studies demonstrates the need to improve the quality of CEAs in terms of missing data handling. Specifically, the review highlights a large proportion of articles which are characterised by a lack of transparent assumptions about missingness and a dramatic shortage of sensitivity analyses for most of the studies. The performance and scoring system of the quality evaluation scheme is affected by a subjective construction of the weights assigned to the different types of information provided about the missing values. When applying the quality evaluation scheme to the reviewed articles, we provide a weight allocation criterion that is considered reasonable and proven to be robust to other allocation schemes. However, alternative configurations could be investigated.

7.1.2 Objective 2: Limitations of the Standard Approach and Full Bayesian Framework in CEA

The "standard" statistical approach in CEA is characterised by a series of limitations that undermine the credibility of the results derived from these analyses. These limitations are mostly related to the failure of accounting for at least some complexities that typically affect trial-based outcome data and the often unclear assumptions about the missing values. The analysis in Chapter 4 has shown the potential implications in terms of missingness assumptions that small changes in the implementation of standard methods may have on the final assessment.

The flexible Bayesian framework presented in Chapter 5 overcomes these limitations through jointly tackling the complexities that affect the data, which is achieved through a flexible modular structure that can handle any suitable distributional assumption and covariates in each module of the framework. The Bayesian setting allows for the full characterisation and quantification of the uncertainty for each variable in the model at a relatively small computational cost compared with a frequentist approach. This naturally favours the incorporation of a sensitivity analysis to assess the impact on the inferences and decision-making of alternative missingness assumptions. In addition, the possibility of implementing the framework using freely available software makes it relatively easy to use for practitioners.

We showed the benefits of using our framework with the MenSS and PBS trials and demonstrated its flexibility by comparing the performance of a set of models that account for an increasingly large number of complexities of the data. The results from both studies have substantially different implications in terms of inferences and cost-effectiveness conclusions compared with those of the original analyses, which instead ignored at least some of the complexities.

For the MenSS trial, we also assessed the robustness of the conclusions to some "extreme" missingness assumptions. This type of sensitivity analysis can be incorporated in the framework at no extra cost in terms of model complexity. The analyses from both studies highlight the importance of adopting a comprehensive modelling approach to economic evaluations and the strategic advantages of building these complex models within a Bayesian framework. An `R` pack-

age, called *missingHE*, dedicated to encourage the implementation among practitioners of the proposed Bayesian framework in routine analyses is currently under development. More details on the structure and the types of models handled by the package are available in Appendix D.

A potential drawback of the framework is that, as the sample size and level of complexity increase, these models may not run quickly in standard Bayesian software (e.g. `JAGS`). In addition, when many partially-observed variables are incorporated into the model, the current capability of these software limits the scope for easily implementing complex models and more specialised MCMC sampling algorithms may be required. In these circumstances, a combination of numerical integration and multiple imputation techniques could be used to avoid some of the speed and computational issues associated with full Bayesian models by separating the imputation and analysis tasks. However, the trade-off is the need to ensure that the imputation and analysis models are compatible, which is not trivial.

7.1.3 Objective 3: Longitudinal Missingness Model in CEA

In Chapter 6 we presented a Bayesian parametric model for conducting inference on a partially-observed bivariate health economic longitudinal outcome. Our approach is an alternative, more efficient and likely less biased approach compared to current practice, which includes a longitudinal model that uses the information from all observed data as well as accounts for the time dependence between the responses.

Key advantages of using this modelling strategy are: 1) specification of a flexible parametric model for the observed responses that accounts for the complexities of the data; 2) use of a principled approach to missingness that partially identifies the distribution of the missing data using partial identifying restrictions and sensitivity parameters; 3) possibility to apply the model to any type of missingness patterns and to handle sparse data; 4) exploration of alternative non-ignorable scenarios through different priors for the sensitivity parameters, calibrated on the observed data.

The analysis of the PBS data shows the benefits of using our approach compared with standard cross-sectional models as well as a considerable impact of alternative nonignorable assumptions on the final decision-making conclusions. A major advantage of using a Bayesian modelling framework is to allow for the formal incorporation of external evidence into the analysis through the use of informative prior distributions. This is particularly useful for parameters that are left unidentified by the observed data and that we do not want to be identified by other undesired (and often non-transparent) modelling assumptions. Although the priors for the sensitivity parameters in Section 6.3.2 were not derived from the formal elicitation of experts' opinion, which was not available in the PBS study, they offer a convenient framework to assess the robustness of the conclusions to differing nonignorable departures. This represents a considerable step forward for the handling missingness in economic evaluations compared with the current practice, which typically relies on methods that assume the ignorability of the missingness mechanism and rarely conducts sensitivity analysis to nonignorable departures.

However, further improvements are certainly possible. When experts' knowledge is available, practical tools for eliciting expert opinion (e.g. questionnaires) may be used to extract experts' beliefs and incorporate them into the priors for the sensitivity parameters; for example, questions may be devised to the experts to quantify plausible differences in the mean utility scores between individuals who did and did not complete the health questionnaires. In addition, the modelling framework can be extended to handle partially-observed covariates and use alternative strategies to incorporate the missingness assumptions into the model. These are discussed in Section 7.2.

7.1.4 General advice for trial-based CEAs

The statistical methods and approaches illustrated in this thesis for the analysis two case studies described in Section 3 provide useful tools to improve the current practice in trial-based CEAs. However, in practice, these methods may need to be tailored to address the specific characteristics of data which may vary according to a number of factors (e.g data collection methods, time horizon, missing data information, etc.). We therefore provide some advice to practitioners to extend the methods applied in this thesis to more general situations.

First, the interpretation and range of the QALYs values associated with the individuals in the trial depends on the time horizon of the economic evaluation. For both the case studies in the thesis, the time horizon was one year, which ensures that the range of the observed QALYs coincide with that of the utility scores collected at each time point (see Section 1.2). However, this does not hold for longer time horizon in which the QALYs can assume values outside this range, e.g. larger than 1. In these cases, the use of parametric distributions restricted to the range $(0, 1)$, such as the Beta distribution, is problematic and alternative approaches need to be considered. A possible approach is to re-scale the data using some transformations to make them fall in the desired range before fitting the model (O'Hagan and Stevens, 2001; Basu and Manca, 2012). However, caution should be used when back-transforming the data to ensure that the estimates are derived correctly on the original scale. For example, when modelling the costs on the logarithmic scale, the results have to be back-transformed appropriately, which is not straightforward when there is heteroskedasticity in the data (Barber and Thompson, 2004). Alternatively, other distributional specifications that allow for all the observed values on the original scale of the data can be used (e.g. using the Normal distributions if the degree of skewness is not high or the sample size is large). When structural values occur in either or both outcome variables (e.g. unit utilities and zero costs), hurdle models provide useful methods which allow to improve the fit of the model to the observed data by explicitly handling these values while using appropriate parametric distributions for the other values.

Second, when the randomisation procedures are correctly implemented, the individuals assigned to different groups in RCTs are generally considered to be similar in both observed and unobserved factors, therefore avoiding the need to adjust for potential confounders in the analysis. However, when imbalances between treatment groups occur in some baseline variables that are highly correlated with the outcomes (e.g. the baseline utilities/costs), then these factors should be included in the analysis both to improve efficiency and to adjust the QALYs and total costs mean estimates for the baseline imbalance in these variables (Manca et al., 2005; Hunter et al., 2015).

Third, the specific characteristics of the studies should be taken into account when modelling the data and interpreting the cost-effectiveness results. For example, when the data are collected from different centres, such as in the PBS trial, the multilevel structure of the data must be explicitly recognised in the model to obtain unbiased estimates, e.g. through the inclusion of random effects (Gomes et al., 2012b; Diaz-Ordaz et al., 2014b). In addtion, for pilot trials, e.g. the MenSS trial, the cost-effectiveness results must be interpreted with caution because these studies are typically characterised by a small sample size and the objective of the analysis is only to provide a preliminary economic evaluation, which is then used to inform the decision about the feasibility of assessing the cost-effectiveness of the interventions in a full-scale trial.

Fourth, when information about missing data is collected in the trial, for example in the form of some auxiliary variables which are thought to be predictive of missingness, it should be incorporated in the analysis to make the MAR assumption more plausible. In addition, the plausibility of MAR depends on the characteristics of the data in the spefic context analysed. For example, when missingness follows a monotone pattern, it turns out that MAR has an intuitive representation in terms of the hazard of dropout and valid inferences under MAR can be obtained using

different approaches which respect the MAR condition (Molenberghs and Kenward, 2007; Daniels and Hogan, 2008). However, when missingness is non-monotone, no transparent MAR condition can be defined and the recommended approach is to use all the information in the observed data to improve the estimates of the distribution of the missing data given the observed. In a frequentist framework (e.g. using MI) this can be achieved by including some auxiliary variables and the outcome in the imputation model, provided that this is correctly specified. In a full Bayesian framework the estimation and imputation steps are performed jointly so that only the potential auxiliary variables need to be included in the imputation model to improve the estimation of the quantities of interest. Although auxiliary variables can be used to make MAR more plausible, however, it is never possible to verify this assumption from the data at hand. Thus, alternative MNAR departures should be explored in sensitivity analysis to assess the robustness of the inferences to a range of plausible missingness scenarios.

Nonignorable models provide useful tools to handle MNAR and alternative types of modelling approaches can be used (Daniels and Hogan, 2008; Molenberghs et al., 2015). Among these, pattern mixture models are well-suited for the identification of sensitivity parameters δ, which provide a framework for assessing sensitivity of the inferences to MNAR departures and for incorporating these into the model. The choice of the sensitivity parameters, the direction and the magnitude of the departures is typically informed from sources external to the data, for example by eliciting experts' opinion. In particuar, the sensitivity parameters should be defined such that: 1) there is not information in the data about δ and, 2) upon specification of δ, the effects of interest are identified in the model. In the analysis of the PBS study in Chapter 6, since no formal experts' opinion was available to inform the MNAR departures, specified their prior distributions on the sensitivity parameters based on a discussion with the people who were familiar with the data to indicate the plausible ranges/directions to explore. However, if experts' opinion is available, this should be formaly incorporated into the model using practical elicitation tools, e.g. questionnaires (Mason et al., 2017), to provide more accurate and plausible departures in sensitivity analysis.

Finally, when dealing with partially-observed data, any measure of predictive accuracy of a model, such as the DIC, can be defined in different ways and are therefore not straightforward to interpret. The recommended approach is to compute these measures and perform model selection based only on the fit to the observed data (a sample algorithm that can be used to compute the DIC based on the observed data is provided in Appendix A.2.1). Nevertheless, this can only provide a partial assessment of the model fit since the fit to the unobserved data can never be checked.

7.1.5 Other potential sources of bias

It is important to note that, in trial-based analyses, the inferences and the final cost-effectiveness conclusions could be affected by different sources of bias, such as treatment non-adherence, recruitment methods and measurement error. In the analysis of the MenSS and PBS studies we did not addressed these potential sources of bias as they fall outside the scope of this thesis, but we acknowledge their existence and their potential impact on the results. Some of these sources of bias, together with their implications and possible methods that can be used to address them, are now briefly described.

Treatment non-adherence

Treatment non-adherence in RCTs refers to situations where some participants do not receive their allocated treatment as intended. When non-adherence occurs, randomisation no longer guarantees that the relationship between treatment receipt and the power to detect the treatment

effects may be reduced. Previous systematic reviews investigating the reporting and statistical handling of non-adherence in RCTs have found that adherence to treatment is often under-reported and when reported, sufficient detail on how adherence was defined is often not included (Dodd et al., 2012; Zhang et al., 2014; Agbla and Diaz-Ordaz, 2018). In the presence of non-adherence, estimates under both an intention to treat analysis (inidviduals allocated to the treatments assigned) and an as treated analysis (indiividuals allocated to the treatments received) may be derived to assess the potnetial impact of non-adherence on the results. The latter estimate can be typically obtained using instrumental variable methods, which have been recently extended and applied to trial-based CEAs (Diaz-Ordaz et al., 2018).

Recruitment methods

Individual participant recruitment is a key issue in RCTs, as there are a number of ways in which biased participant recruitment can lead to baseline imbalances in important prognostic factors. Typically, recruitment bias is avoided by concealing the random allocation from the potential participant and researcher until after they have been recruited. However, in cluster RCTs sometimes this is not possible because groups (or clusters) of participants are randomised rather than individuals, yet data are collected on individual participants. Selective recruitment of individual participants can occur in these trials if the people recruiting participants know the participants' allocation, even when allocation of clusters has been adequately concealed (Eldridge et al., 2009). Potential strategies to handle this problem are usually implemented at the design stage of the trial, e.g. no recruiting of individual participants, recruiting of participants before randomisation, recruiting outside the cluster setting or masking of the recruiters.

Measurement error

In trial analyses, when estimating the treatment effect on a continuous response variable observed at baseline and some follow-up points, baseline observations may be prone to measurement error, which is typically a result of within-patient biological variability and technical variability. It is generally known that measurement error at baseline will attenuate the regression slope that summarises the association between the responses observed at baseline and at follow-up points, which can estimated for example using ordinary least squares methods (Chan et al., 2004). In RCTs, if there is no measurement error, the observed baseline differences between treatment groups are entirely due to sampling variation. However, when the baseline data are additionally subject to measurement error, any observed baseline differences arise from the combined effects of sampling variation and measurement error, and there is no easy way of differentiating between these two components. Controversies exist in the literature about whether attenuation of the slope due to measurement error in the baseline data is problematic in RCTs, and whether it should be ignored or adjusted for using alternative estimators of the treatment effect (Chambless and Roeback, 1993; Yanez et al., 1998).

7.2 Extensions

There is scope for extending the work of this thesis in a number of directions, especially with respect to the modelling approach described in Chapter 5 and Chapter 6.

One possibility is to incorporate into the modelling frameworks some baseline covariates that are predictive of missingness, either at the cross-sectional or longitudinal level. There are a number of reasons making the incorporation of covariates in the analysis a potentially difficult task. First, as the number of variables increases, the level of complexity of the model is increased

as well, which may considerably slow down convergence of the MCMC algorithms. Second, when some of these variables are partially-observed, it may be necessary to explicitly model the missing data mechanism to allow for informative missingness in the covariates. The obvious consequence is that substantial complexity is added to the existing full data model through additional sub-models, one for each nonignorable covariate, for which plausible missingness scenarios should be elicited and explored in sensitivity analysis. Implementing models with this level of complexity may not be feasible in standard software such as `JAGS` or `OpenBUGS`, in which case alternative software and/or approximations to the full data model, e.g. tailored MCMC samplers, would need to be sought or developed.

Another possible area of extension is to further increase the flexibility of the proposed approach by relaxing the parametric assumptions of the model and specify semi-parametric or non-parametric distributions for the observed data, which allow a weakening of the model assumptions and likely further improve the fit of the model to the data. Different strategies are available to implement these distributions depending on the type of responses and model specification. Examples are the use of shrinkage priors to share information and improve estimation stability for sparsely observed patterns (Gaskins et al., 2016) or the specification of a Dirichlet process mixture model as a prior on the joint distribution of the working model (Linero and Daniels, 2015).

These approaches have been mostly applied to univariate longitudinal responses, where missingness is typically monotone and the objective of the analysis is the estimation of some treatment effects at some follow-up. It would be interesting to adapt these strategies to improve the performance of the model while simultaneously accounting for the complexities that typically characterise utility and cost data in trial-based CEAs.

Finally, with respect to the nonignorable strategy used in Chapter 6, another possible area of future work is to develop alternative types of identifying restrictions for conducting sensitivity analysis. Many different restrictions have been proposed in the literature, mostly for handling monotone missingness. Conversely, restrictions for nonmonotone missingness, which often occurs in trial-based CEA data, is a topic that has been treated sparingly in the literature (Linero and Daniels, 2018) and further research is needed. For example, the identifying restrictions could be specified such that the sensitivity parameters are introduced at the conditional mean (rather than marginal mean) for each missing response in the model. Once a benchmark scenario is chosen, then plausible departures from it could be explored through suitably-defined informative priors on the sensitivity parameters.

Executive Summary

Three main research objectives were studied in this thesis. These are: **1)** to review the missing data methods used in trial-based economic evaluations and assess the quality of routine analyses with respect to the reporting and handling of missingness; **2)** to identify potential limitations in the standard approach used by practitioners and propose a full Bayesian framework that can improve the current practice and avoid biased results; **3)** to develop a Bayesian principled approach to handle missingness that can combine a model for the observed data with explicit assumptions about the missing values.

In Chapter 1, the theoretical underpinnings for these objectives were summarised with respect to the three main research areas involved, namely the health economics framework, the Bayesian statistical approach and the methods used in missing data analyses. The first objective was addressed in Chapter 2, which showed the results from a literature review on the missing data methods in trial-based economic evaluations. Guidelines were also provided for assessing the quality of missing data analyses in the form of a structural framework, which was applied to the articles studied in the review. An extended version of this chapter is published as a research article in *PharmacoEconomics-Open* (`https://link.springer.com/article/10.1007/s41669-017-0015-6`).

Chapter 3 presented the two case studies analysed in the thesis and described the standard approach used by practitioners in routine analyses. The second objective was addressed in the following two chapters. Chapter 4 illustrated a pitfall related to alternative implementations of the standard approach in terms of missing data assumptions. Both case studies were used to demonstrate the potential bias associated with this approach. The content of this chapter has been submitted in the form of a research article for publication in *Health Economics* and is currently available on *arXiv* (`https://arxiv.org/abs/1805.07149`).

Chapter 5 presented a general Bayesian modelling framework for economic evaluations that improves the standard approach by jointly tackling the typical complexities affecting the data, while also accounting for missing data uncertainty. Both case studies are used to demonstrate the benefits and flexibility of the proposed framework with respect to the standard approach. The content of this chapter has been submitted in the form of a research article for publication in *Statistics in Medicine* and is currently available on *arXiv* (`https://arxiv.org/abs/1801.09541`).

Chapter 6 addressed the third objective through the proposition of a parametric Bayesian longitudinal model that extends the typical modelling framework in economic evaluations to incorporate a principled sensitivity analysis to missingness, while simultaneously accounting for the complexities of the data. This approach was motivated and applied using the data from one of the two case studies analysed in the thesis. The content of this chapter has been submitted in the form of a research article for publication in *Journal of the Royal Statistical Society: Series A* and is currently available on *arXiv* (`https://arxiv.org/abs/1805.07147`).

Finally, Chapter 7 summarised the main conclusions from the thesis and proposed directions for future research. Some of the methods presented in this thesis have been wrapped into an `R` package called *MissingHE*, which is still under development and for which a preliminary version

is available on *GitHub* (`https://github.com/AnGabrio`). The package is specifically dedicated to encourage and facilitate the implementation among practitioners of the Bayesian methods illustrated in this thesis.

Appendix A

Supplementary Information

A.1 Monte Carlo integration

Monte Carlo integration is a simulation method that can be used to derive inferences from a model when the posterior distribution, while possibly known in closed form, is not analytically tractable. In Monte Carlo integration, the integral in the posterior is replaced by the average obtained from simulated values. Let $t(\theta)$ denote a generic summary measure of interest for the parameter θ. Then, the posterior summary $\int t(\theta)p(\theta \mid \boldsymbol{y})d\theta$ is the expected value $\mathrm{E}[t(\theta) \mid \boldsymbol{y}]$ of $t(\theta)$ under the distribution $p(\theta \mid \boldsymbol{y})$. Assume we have S independently sampled values $\{\theta^1, \ldots, \theta^S\}$ for the parameter θ from $p(\theta \mid \boldsymbol{y})$, then for S large

$$\int t(\theta)p(\theta \mid \boldsymbol{y})d\theta \approx \bar{t}_S = \frac{1}{S}\sum_{s=1}^{S} t(\theta^s), \tag{A.1}$$

according to the Strong Law of Large Numbers. The unbiased estimator $\bar{t}_S$ is called the *Monte Carlo estimator* of $\mathrm{E}[t(\theta) \mid \boldsymbol{y}]$. For a discrete θ, the integral in Equation A.1 is replaced by a sum. The consequence of MC integration is that probabilities, summations and integrals can be approximated by the MC method and the empirical distribution of sampled values converges to the true posterior as $S \to \infty$.

According to the Central Limit Theorem, the precision with which the true summary measure is estimated may be quantified by the interval $\left[\bar{t}_S \pm 1.96\frac{s_t}{\sqrt{S}}\right]$. The quantity s_t is the standard deviation of the sampled $t(\theta^s)-$values and $\frac{s_t}{\sqrt{S}}$ is called the *Monte Carlo standard error*. The first is an approximation of the posterior standard deviation while the second is an estimate of the uncertainty of the estimated posterior mean.

A.2 Deviance Information Criterion

As any information criterion, DIC can be used to compare models based on their predictive accuracy and is based on a measure of overall fit of a model, the *deviance*, and a penalty for model complexity, the *effective number of parameters*.

The deviance $D(\boldsymbol{\theta})$ is a general measure that can be used to assess model fit and is defined as

$$D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta} \mid \boldsymbol{y}), \tag{A.2}$$

where $\log L(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the *log-predictive density* or *log-likelihood*, which summarises the information in the data $\boldsymbol{y}$ about the parameters $\boldsymbol{\theta}$, with larger values of the deviance indicating poorer fit. In Bayesian statistics, the quantity in Equation A.2 can be summarised in different ways. For

example, the posterior mean of the deviance $\overline{D(\boldsymbol{\theta})} = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})]$ has been suggested as a sensible Bayesian measure of fit (Dempster, 1973). Alternatively, the deviance can be evaluated at the posterior mean of the parameters, i.e. $D(\bar{\boldsymbol{\theta}}) = D(\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}])$. In general, we use the notation $\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]$ to denote the expectation of $\boldsymbol{\theta}$ with respect to the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Using these two summary measures of the deviance, Spiegelhalter et al. (2002) proposed the penalty term

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \tag{A.3}$$

which is used as an estimate of the "effective number of parameters" in the model. An alternative and slightly different definition of p_D has also been proposed (Gelman et al., 2004):

$$p_D = \mathsf{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})], \tag{A.4}$$

where $\mathsf{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})]$ denotes the variance of the posterior deviance. Compared with the quantity in Equation A.3, the definition of the effective number of parameters in Equation A.4 has the advantage of always being positive and is generally more stable.

Based on $D(\boldsymbol{\theta})$ and p_D, Spiegelhalter et al. (2002) proposed the DIC as a Bayesian model selection criterion, defined as

$$\mathsf{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D = \overline{D(\boldsymbol{\theta})} + p_D, \tag{A.5}$$

with the model taking the smallest value of DIC being preferred. An attractive aspect of using Equation A.5 is that it can readily be calculated from an MCMC run by monitoring $\boldsymbol{\theta}$ and $D(\boldsymbol{\theta})$. Let $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^S\}$ be the posterior samples for the parameters $\boldsymbol{\theta}$ from a converged Markov chain. We can then easily calculate the approximations $D(\bar{\boldsymbol{\theta}}) \approx \frac{1}{S}\sum_{s=1}^{S} D(\boldsymbol{\theta}^s)$ and $\overline{D(\boldsymbol{\theta})} \approx D(\frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\theta}^s)$. DIC can alternatively be written as a function of the log-likelihood

$$\mathsf{DIC} = 2\log L\{\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\boldsymbol{\theta} \mid \boldsymbol{y}\right]\} - 4\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\log L(\boldsymbol{\theta} \mid \boldsymbol{y})\right]. \tag{A.6}$$

DIC can often be calculated automatically by software such as `OpenBUGS`, where Equation A.6 is derived by taking $\overline{D(\boldsymbol{\theta})}$ as the posterior mean of $-2\log L(\boldsymbol{\theta} \mid \boldsymbol{y})$ and evaluating $D(\bar{\boldsymbol{\theta}})$ as -2 times the log-likelihood at the posterior mean of the stochastic nodes.

The DIC implementation and automatic calculation in many MCMC programmes, such as `OpenBUGS` or `JAGS`, has greatly encouraged its use among applied statisticians. However, there are some potential difficulties and pitfalls when using DIC to compare models (Spiegelhalter et al., 2002). First, p_D and therefore DIC are not invariant to reparameterisation of the model. Second, a negative p_D may occur with a non-logconcave likelihood and when the prior is in conflict with the data. Third, the best model as determined by the DIC can change depending on the choice of "likelihood"; for example, for multilevel models the likelihood can take one of two forms: the likelihood obtained by integrating out the random effects or the likelihood without the random effects integrated out. Another limitation, common to all likelihood based criteria, is that for some models, the likelihood is not available in closed form (e.g. the Hurdle models in Chapter 5), for which the likelihood can be typically evaluated using Monte Carlo integration.

Finally, the DIC is not uniquely defined in the presence of missing data and its use and interpretation are not straightforward (Celeux et al., 2006; Mason et al., 2012a). In this situation, a common practice is to compare models using a DIC computed on the observed values alone, using Monte Carlo integration when the sampling distribution of the observed data is not available in closed form (Daniels and Hogan, 2008). The appropriate extension of the observed data DIC for multilevel models with missing data is based on the observed data likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{y}_{obs})$ under ignorability of the missing data but conditional on the random effects (Ntzoufras, 2009). This is

the DIC typically implemented for multilevel models in the BUGS software and, in the context of the missing data, is computed as

$$\begin{aligned}\text{DIC} &= D(\overline{\boldsymbol{\theta}}, \boldsymbol{u} \mid \boldsymbol{y}_{obs}) + 2p_D \\ &= 2\overline{D(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}_{obs})} - D(\overline{\boldsymbol{\theta}, \boldsymbol{u}} \mid \boldsymbol{y}_{obs}) \\ &= 2\log L\{\mathsf{E}_{\boldsymbol{\theta},\boldsymbol{u}|\boldsymbol{y}_{obs}}\left[\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}_{obs}\right]\} - 4\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}_{obs}}\left[\log L(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}_{obs})\right].\end{aligned}$$

where the random effects $\boldsymbol{u}$ are considered as extra parameters to be estimated, i.e. $\overline{D(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}_{obs})} = -2\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}_{obs}}\left[\log L(\boldsymbol{y}_{obs} \mid \boldsymbol{\theta}, \boldsymbol{u})\right]$ and p_D denotes the number of effective parameters, which is defined as $\overline{D(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}_{obs})} - D(\overline{\boldsymbol{\theta}, \boldsymbol{u}} \mid \boldsymbol{y}_{obs})$. When the response distribution does not have a known closed form, Monte Carlo integration is typically used to calculate the corresponding likelihood and the DIC.

A.2.1 Algorithm for the computation of the DIC based on the observed data likelihood

Let $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ be the joint model of interest, formed by the usual bivariate economic outcome $\boldsymbol{y} = (e, c)$ and indexed by the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_c)$. For example, consider the modelling framework described in Chapter 5, which is based on the factorisation of the joint into a marginal distribution of the QALYs $p(e \mid \boldsymbol{\theta}_e)$ and a conditional distribution of the costs $p(c \mid e, \boldsymbol{\theta}_c)$, indexed by the parameters $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_c$, respectively. The algorithm implemented in our applications to calculate the observed DIC proceeds as follows.

1 Carry out a standard MCMC run on the joint model $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(e \mid \boldsymbol{\theta}_e)p(c \mid e, \boldsymbol{\theta}_c)$ and save the posterior samples for the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_e, \boldsymbol{\theta}_c)$, which are denoted as $\boldsymbol{\theta}^{(k)}$, for $k = 1, \ldots, K$.

2 Evaluate the posterior means of $\boldsymbol{\theta}$ across the posterior samples, denoted by $\overline{\boldsymbol{\theta}} = (\overline{\boldsymbol{\theta}}_e, \overline{\boldsymbol{\theta}}_c)$.

3 At each iteration of the posterior distribution, generate a sample $e_{mis}^{(kl)}$, for $l = 1 \ldots, L$, from the appropriate likelihood evaluated at $\boldsymbol{\theta}_e^{(k)}$, e.g. $e_{mis}^{(kl)} \sim \text{Beta}\left(\mu_e^{(k)}\tau_e^{(k)}, (1-\mu_e^{(k)})\tau_e^{(k)}\right)$.

4 Then evaluate

$$h^{(k)} = \mathsf{E}_{e_{mis}|e_{obs},\boldsymbol{\theta}_e^{(k)}}\left[p(c \mid e_{obs}, e_{mis}, \boldsymbol{\theta}_c^{(k)})\right] \approx \frac{1}{L}\sum_{l=1}^{L} p(c \mid e_{obs}, e_{mis}^{(kl)}, \boldsymbol{\theta}_c^{(k)})$$

5 Calculate the posterior expectation of the observed data log likelihood as

$$\frac{1}{K}\sum_{k=1}^{K}\left[\log L(\boldsymbol{\theta}_e^{(k)} \mid e_{obs}) + \log h^{(k)}\right]$$

and multiply this by -2 to get the posterior mean of the deviance, denoted as $\overline{D(\boldsymbol{\theta})}$.

6 Generate a new sample $\boldsymbol{e}_{mis}^{(l)}$, for $l = 1, \ldots, L$, using $\overline{\boldsymbol{\theta}}_e$. Evaluate the plug-in observed data log likelihood using the posterior mean of the parameters as

$$\log L(\overline{\boldsymbol{\theta}}_e \mid e_{obs}) + \log\left(\mathsf{E}_{e_{mis}|e_{obs},\overline{\boldsymbol{\theta}}_e}\left[p(c \mid e_{obs}, e_{mis}, \overline{\boldsymbol{\theta}}_c)\right]\right),$$

where

$$\mathsf{E}_{e_{mis}|e_{obs},\overline{\boldsymbol{\theta}}_e}\left[p(c \mid e_{obs}, e_{mis}, \overline{\boldsymbol{\theta}}_c)\right] \approx \frac{1}{L}\sum_{l=1}^{L} p(c \mid e_{obs}, e_{mis}^{(l)}, \overline{\boldsymbol{\theta}}_c)$$

and multiply this plug-in log likelihood by -2 to get the plug-in deviance, denoted as $D(\overline{\boldsymbol{\theta}})$

7 Finally, calculate DIC $= 2\overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$

A.3 Condition of Validity for Complete Case Analysis

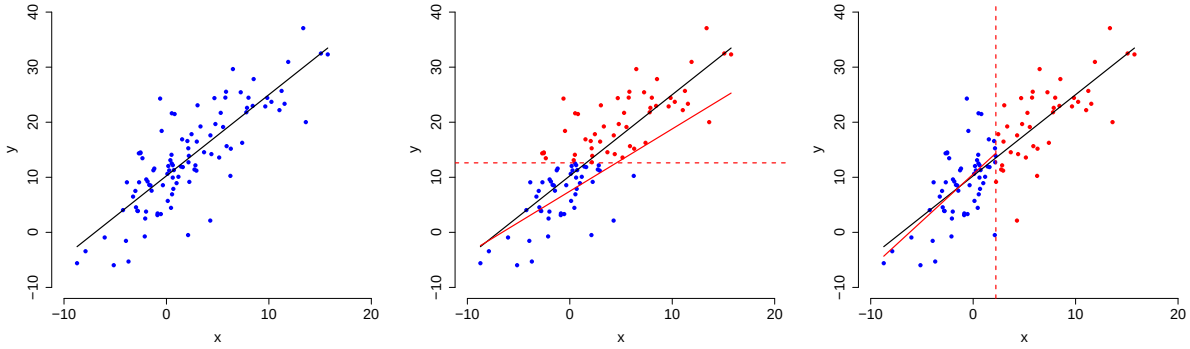
When the focus of the analysis is in some aspect of the conditional distribution of the response $\boldsymbol{y}$ given some covariate $\boldsymbol{x}$ (e.g. regression parameter estimates), then CCA will lead unbiased estimates as long as the distribution of $\boldsymbol{y} \mid \boldsymbol{x}$ in the completers is the same as in the target population. However, this condition does not match exactly the definition of the Rubin's classes since it typically holds when the missingness mechanism is MCAR, but also under certain MAR and MNAR mechanisms (White and Carlin, 2010).

For example, suppose the interest is to perform a linear regression model on some data formed by the outcome $\boldsymbol{y}$ and covariates $\boldsymbol{x}$ variable, for which some values in either $\boldsymbol{y}$ or $\boldsymbol{x}$ could be unobserved. Let $\boldsymbol{r}$ denote whether a subject is a complete case ($\boldsymbol{r} = \mathbf{1}$) or not ($\boldsymbol{r} = \mathbf{0}$). Under a CCA, all missing values are ignored with the analysis being performed on the complete cases alone. Compared with a full data analysis (where all data are observed), the regression parameter estimates will be less precise due to the loss of efficiency caused by the fewer cases on which the model is fitted. However, the condition of validity of CCA (in terms of the bias associated with the regression estimates) changes according to the type of the missing data mechanism and the variable affected by missingness. Specifically, under a MCAR mechanism in either $\boldsymbol{y}$ or $\boldsymbol{x}$, the parameter estimates are unbiased because the complete cases are still a random sample from the population. When missingness is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{x}$ (i.e. $\boldsymbol{r} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{x}$), then the parameter estimates under CCA are still unbiased because the conditional distribution $\boldsymbol{y} \mid \boldsymbol{x}$ in the complete cases is the same as in the target population:

$$
\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{r} = \mathbf{1}) = \frac{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r} = \mathbf{1})}{p(\boldsymbol{x}, \boldsymbol{r} = \mathbf{1})} &= \frac{p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{y}, \boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{x}) p(\boldsymbol{x})} \\
&= \frac{p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{x}) p(\boldsymbol{x})} \\
&= p(\boldsymbol{y} \mid \boldsymbol{x})
\end{aligned}
$$

where, under the assumption that $\boldsymbol{r} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{x}$, then $p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r} = \mathbf{1} \mid \boldsymbol{x})$. When, instead, missingness depends on $\boldsymbol{y}$, even after conditioning on $\boldsymbol{x}$, then the regression estimates are biased. Figure A.1 visually displays the regression line estimated from the linear model of $\boldsymbol{y} \mid \boldsymbol{x}$ using CCA on a simulated dataset under three alternative scenarios: using all cases with no missing values (panel a), when all cases with a response value higher than the mean of $\boldsymbol{y}$ are removed from the analysis (panel b), and when all cases with a covariate value higher than the mean of $\boldsymbol{x}$ are removed from the analysis (panel c). The regression line estimated on all cases (panel a) is the benchmark and most efficient scenario since all data are used for the estimation of the regression parameters. When missingness is dependent only on the outcome $\boldsymbol{y}$ (panel b), the estimated regression parameters are biased, while when missingness is dependent on the covariate $\boldsymbol{x}$ but not on $\boldsymbol{y}$ (panel c), the estimated regression parameters are unbiased.

In general the condition of validity for CCA is that misssingness should be (conditionally) independent of the outcome, which may occur under different types of missing data mechanisms. If missingness only affects $\boldsymbol{y}$, then CCA's validity assumption coincides with the MAR assumption. However, when missingness affects $\boldsymbol{x}$, CCA's validity condition does not fit neatly into the Rubin's categories. For example, suppose we divide the covariates $\boldsymbol{x}$ into $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, where $\boldsymbol{x}_2$ is always observed while $\boldsymbol{x}_1$ can be missing. Then, if missingness depends only on $\boldsymbol{x}_2$, then the data are



(a) Fully observed data

(b) Missingness dependent on $\boldsymbol{y}$

(c) Missingness dependent on $\boldsymbol{x}$

Figure A.1: Estimated regression line from the conditional linear model $\boldsymbol{y} \mid \boldsymbol{x}$ using CCA under three scenarios: with fully observed data (panel a), when missingness depends on $\boldsymbol{y}$ (panel b) and when missingness depends on $\boldsymbol{x}$ (panel c).

MAR but if missingness depends on $\boldsymbol{x}_1$ or jointly on $\boldsymbol{x}$, then the data are MNAR. In both cases, since missingness is conditionally independent of the response $\boldsymbol{y}$, the regression estimates under CCA are unbiased.

Appendix B

Model Code

B.1 Mean Baseline Adjustment

The complete JAGS code for implementing the CC/AC version of the mean baseline adjustment for the model fitted to the MenSS study in Chapter 4 is presented below.

```
model{

# N1 = 75 number of subjects in the control (t=1)
# N2 = 84 number of subjects in the intervention (t=2)
# et = QALYs variables in the two groups
# ct = Total cost variables in the two groups
# u0_t = Baseline utility variables in the two groups
# u0_cc[t] = Mean baseline utilities based on the CC in both groups
# u0_ac[t] = Mean baseline utilities based on the AC in both groups

      for(i in 1 : N1){ #control
          e1[i] ~ dnorm(mu.e[i, 1], tau.e[1])
          mu.e[i, 1] <- beta0[1] + beta1[1] * u0_1[i]
          cost1[i] ~ dnorm(mu.c[1], tau.c[1])
          }

      for(i in 1 : N2){ #intervention
          eff2[i] ~ dnorm(mu.e[i, 2], tau.e[2])
          mu.e[i,2] <- beta0[2] + beta1[2] * u0_2[i]
          cost2[i] ~ dnorm(mu.c[2], tau.c[2])
         }

    # mean baseline utilities provided as data

        mu.e.cc[1] <- beta0[1] + beta1[1] * u0_cc[1]
        mu.e.cc[2] <- beta0[2] + beta1[2] * u0_cc[2]
        mu.e.ac[1] <- beta0[1] + beta1[1] * u0_ac[1]
        mu.e.ac[2] <- beta0[2] + beta1[2] * u0_ac[2]

    # from precision to variance and standard deviation

    for(t in 1 : 2){
         tau.e[t] <- 1 / ss.e[t]
         ss.e[t] <- sd.e[t] * sd.e[t]
         sd.e[t] <- exp(ls.e[t])
         tau.c[t] <- 1 / ss.c[t]
         ss.c[t] <- sd.c[t] * sd.c[t]
         sd.c[t] <- exp(ls.c[t])
```

```
    # priors

        beta0[t] ~ dnorm(0, 0.00001)
        beta1[t] ~ dnorm(0, 0.00001)
        ls.e[t] ~ dunif(-5, 10)
        ls.c[t] ~ dunif(-5, 10)
        mu.c[t] ~ dnorm(0, 0.00001)
     }

    # incremental quantities

        delta_e.cc <- mu.e.cc[2] - mu.e.cc[1] # based on u0_cc
        delta_e.ac <- mu.e.ac[2] - mu.e.ac[1] # based on u0_ac
        delta_c <- mu.c[2] - mu.c[1] #cost increment
}
```

B.2 Hurdle Model

The complete JAGS code for implementing the Hurdle Model in Chapter 5 for the analysis of the MenSS data is given below.

```
model {

# N1 = 75 number of subjects in the control (t=1)
# N2 = 84 number of subjects in the intervention (t=2)
# et = QALYs variables in the two groups
# ct = Total cost variables in the two groups
# u0_t = Baseline utility variables in the two groups
# d.et = Structural one indicators for et in the two groups
# d.u0_1 = Structural one indicators for u0_t in the two groups
# aget, ethnicityt, employmentt = Covariate variables in the two groups

      for(i in 1 : N1) { # control

      # 1. module for the structural ones in the QALYs

        d.e1[i] ~ dbern(pi.e[i, 1])
        logit(pi.e[i, 1]) <- gamma0[1] + gamma1[1] * (u0_1[i] - mean(u0_1[])) +
               gamma2[1] * (age1[i] - mean(age1[])) + gamma3[ethnicity1[i], 1] +
               gamma4[employment1[i], 1]

      #2. module for the structural ones in the baseline utilities

        d.u0_1[i] ~ dbern(pi.u[i, 1])
        logit(pi.u[i, 1]) <- eta0[1] + eta1[1] * (age1[i] - mean(age1[])) +
        eta2[ethnicity1[i], 1] + eta3[employment1[i], 1]

      #3. marginal module for the QALYs

        e1[i] ~ dbeta(phi.e[i, 1] * tau.e[i, 1], (1 - phi.e[i, 1]) * tau.e[i, 1])
        tau.e[i, 1] <- phi.e[i, 1] * (1 - phi.e[i, 1]) /
        pow(sigma.e[d.e1[i] + 1], 2) - 1
        logit(phi.e[i, 1]) <- alpha0[d.e1[i]+1, 1] +
        alpha1[d.e1[i]+1, 1] * (u0_1[i] - mean(u0_1[]))

      #4. marginal module for the baseline utilities
```

```
  u1[i] ~ dbeta(mu.u[d.u0_1[i] + 1, 1] * tau.u[d.u0_1[i] + 1, 1],
  (1 - mu.u[d.u0_1[i] + 1, 1]) * tau.u[d.u0_1[i] + 1, 1])

 #5. conditional module for the costs

  c1[i] ~ dgamma(phi.c[i, 1] * tau.c[i, 1], tau.c[i, 1])
  tau.c[i, 1] <- phi.c[i, 1] / pow(sigma.c[1], 2)
  log( phi.c[i, 1]) <- beta0[1] + beta1[1] * (e1[i] - mu.e[1])
 }

 for(i in 1 : N2) { #intervention

#1. module for the structural ones in the QALYs

  d.e2[i] ~ dbern(pi.e[i, 2])
  logit(pi.e[i, 2]) <- gamma0[2] + gamma1[2] * (u0_2[i] - mean(u0_2[])) +
         gamma2[2] * (age2[i] - mean(age2[])) + gamma3[ethnicity2[i], 2] +
         gamma4[employment2[i], 2]

#2. module for the structural ones in the baseline utilities

  d.u0_2[i] ~ dbern(pi.u[i, 2])
  logit(pi.u[i, 2]) <- eta0[2] + eta1[2] * (age2[i] - mean(age2[])) +
  eta2[ethnicity2[i], 2] + eta3[employment2[i], 2]

#3. marginal module for the QALYs

  e2[i] ~ dbeta(phi.e[i, 2] * tau.e[i, 2], (1 - phi.e[i, 2]) * tau.e[i, 2])
  tau.e[i,2] <- phi.e[i, 2] * (1 - phi.e[i, 2]) /
  pow(sigma.e[d.e2[i] + 1], 2) - 1
  logit(phi.e[i, 2]) <- alpha0[d.e2[i] + 1, 2] +
  alpha1[d.e2[i] + 1, 2] * (u0_2[i] - mean(u0_2[]))

#4. marginal module for the baseline utilities

  u0_2[i] ~ dbeta(mu.u[d.u0_2[i] + 1, 2] * tau.u[d.u0_2[i] + 1, 2],
  (1 - mu.u[d.u0_2[i] + 1, 2]) * tau.u[d.u0_2[i] + 1, 2])

#5. conditional module for the costs

  c2[i] ~ dgamma(phi.c[i, 2] * tau.c[i, 2], tau.c[i, 2])
  tau.c[i, 1] <- phi.c[i, 1] / pow(sigma.c[2], 2)
  log( phi.c[i, 2]) <- beta0[2] + beta1[2] * (e2[i] - mu.e[2])
 }

# priors for module 1 and 2

 for(t in 1 : 2) {
  gamma0[t] ~ dlogis(0, 1)
  gamma1[t] ~ dnorm(0, 0.00001)
  gamma2[t] ~ dnorm(0, 0.00001)
  eta0[t] ~ dlogis(0, 1)
  eta2[t] ~ dnorm(0, 0.00001)

# priors on coefficients for categorical covariates
#(set reference category as 0)

  gamma3[1, t] <- 0
  gamma4[1, t] <- 0
  eta2[1, t] <- 0
```

```
    eta3[1, t] <- 0
  }

# priors for all other categories
# mu and tau values provided as data variables
#with zero means and small precisions (0.00001)
# ethnicity has different numbers of categories between arms

    gamma3[2:14, 1] ~ dmnorm(mu1.gamma3[], tau1.gamma3[, ])
    gamma3[2:12, 2] ~ dmnorm(mu2.gamma3[], tau2.gamma3[, ])
    gamma4[2:6, 1] ~ dmnorm(mu1.gamma4[], tau1.gamma4[, ])
    gamma4[2:6, 2] ~ dmnorm(mu2.gamma4[], tau2.gamma4[, ])
    eta2[2:14, 1] ~ dmnorm(mu1.eta2[], tau1.eta2[, ])
    eta2[2:12, 2] ~ dmnorm(mu2.eta2[], tau2.eta2[, ])
    eta3[2:6, 1] ~ dmnorm(mu1.eta3[], tau1.eta3[, ])
    eta3[2:6, 2] ~ dmnorm(mu2.eta3[], tau2.eta3[, ])

 for(t in 1 : 2) {

# priors for model 3
# priors for the ones group in the QALYs

    alpha0[2, t] <- logit(0.999999)
    alpha1[2, t] <- 0
    sigma.e[2, t] <- 0.00001

# priors for the non-ones group in the QALYs

    alpha0[1, t] ~ dnorm(0, 0.000001)
    alpha1[1, t] ~ dnorm(0, 0.000001)
    sigma.e[1, t] ~ dunif(0, sd.limit.e[t])
    sd.limit.e[t] <- pow(mu.e[1, t] * (1 - mu.e[1, t]), 0.5)

# priors for model 4
# priors for the ones group in the baseline utilities

    tau.u[2, t] <- mu.u[2, t] * (1 - mu.u[2, t]) /
    pow(sigma.u[2, t], 2) - 1
    logit(mu.u[2, t]) <- delta0[2, t]
    delta0[2, t] <- logit(0.999999)
    sigma.u[2, t] <- 0.00001

# priors for the non-ones group in the baseline utilities

    tau.u[1, t] <- mu.u[1,t] * (1 - mu.u[1, t]) /
    pow(sigma.u[1, t], 2) - 1
    logit(mu.u[1, t]) <- delta0[1,t]
    delta0[1, t] ~ dnorm(0, 0.00001)
    sigma.u[1, t] ~ dunif(0, sd.limit.u[t])
    sd.limit.u[t] <- pow(mu.u[1, t] * (1 - mu.u[1, t]), 0.5)

# priors for module 5

    beta0[t] ~ dnorm(0, 0.00001)
    sigma.c[t] ~ dunif(0, 1000)
    beta1[t] ~ dnorm(0, 0.00001)

# obtain marginal probabilities for weighting

    p[t] <- ilogit(gamma0[t])
```

```
    # obtain the weighted marginal mean QALYs

       mu.e[t] <- p[t] + (1-p[t]) * ilogit(alpha0[t])
     }

    # compute incremental QALYs and costs

       delta_e <- mu.e[2] - mu.e[1]
       delta_c <- mu.c[2] - mu.c[1]
}
```

B.3 Longitudinal Model

B.3.1 Model Code

The JAGS code for fitting the longitudinal model in Chapter 6 to the completers and non-completers patterns in the PBS study is given below.

```
model{

# r = pattern indicator
# uj, cj = utilities and costs at time j = 0,1,2
# dj.u, dj.c = indicators for the structural values at time j
# trt = treatment indicator (1 = control, 2 = intervention)
# N = 244 total number of individuals
# N.r1 = 204 number of individuals in the completers pattern
# N.r2 = N - N.r1 = 40 number of individuals in the non-completers pattern
# R.star = number of theoretical patterns
# prior.psi = vector containing the hyperprior values for psi

    # model for patterns r

     for(i in 1:N){
        r[i] ~ dcat(psi[1:R.star,trt[i]])
     }

    # priors on model for r

    for(t in 1:2){
       psi[1:R.star,t] ~ ddirch(prior.psi[])
    }

   # model for the responses in the completers (r1) stratified by treatment

   for(i in 1:N.r1){
     c0[i] ~ dlnorm(nu0.c.r1[d0.c[i]+1,trt[i]],tau0.c.r1[d0.c[i]+1,trt[i]])
     d0.c[i] ~ dbern(pi0.c.r1[trt[i]])
     u0[i] ~ dbeta(a0.r1[i,trt[i]],b0.r1[i,trt[i]])
     a0.r1[i,trt[i]] <- nu0.u.r1[i,trt[i]]*(nu0.u.r1[i,trt[i]]*
     (1-nu0.u.r1[i,trt[i]])/pow(sigma0.u.r1[d0.u[i]+1,trt[i]],2)-1)
     b0.r1[i,trt[i]] <- (1-nu0.u.r1[i,trt[i]])*(nu0.u.r1[i,trt[i]]*
     (1-nu0.u.r1[i,trt[i]])/pow(sigma0.u.r1[d0.u[i]+1,trt[i]],2)-1)
     logit(nu0.u.r1[i,trt[i]]) <- alpha00.r1[d0.u[i]+1,trt[i]]+
     alpha10.r1[d0.u[i]+1,trt[i]]*log(c0[i])
     d0.u[i] ~ dbern(pi0.u.r1[i,trt[i]])
     logit(pi0.u.r1[i,trt[i]]) <- gamma00.r1[trt[i]]+
     gamma10.r1[trt[i]]*log(c0[i])
```

```
  c1[i] ~ dlnorm(nu1.c.r1[i,trt[i]],tau1.c.r1[d1.c[i]+1,trt[i]])
  nu1.c.r1[i,trt[i]] <- beta01.r1[d1.c[i]+1,trt[i]]+
  beta11.r1[d1.c[i]+1,trt[i]]*log(c0[i])+beta21.r1[d1.c[i]+1,trt[i]]*u0[i]
  d1.c[i] ~ dbern(pi1.c.r1[i,trt[i]])
  logit(pi1.c.r1[i,trt[i]]) <- zeta01.r1[trt[i]]+
  zeta11.r1[trt[i]]*log(c0[i])+zeta21.r1[trt[i]]*u0[i]
  u1[i] ~ dbeta(a1.r1[i,trt[i]],b1.r1[i,trt[i]])
  a1.r1[i,trt[i]] <- nu1.u.r1[i,trt[i]]*(nu1.u.r1[i,trt[i]]*
  (1-nu1.u.r1[i,trt[i]])/pow(sigma1.u.r1[d1.u[i]+1,trt[i]],2)-1)
  b1.r1[i,trt[i]] <- (1-nu1.u.r1[i,trt[i]])*(nu1.u.r1[i,trt[i]]*
  (1-nu1.u.r1[i,trt[i]])/pow(sigma1.u.r1[d1.u[i]+1,trt[i]],2)-1)
  logit(nu1.u.r1[i,trt[i]]) <- alpha01.r1[d1.u[i]+1,trt[i]]+
  alpha11.r1[d1.u[i]+1,trt[i]]*log(c1[i])+alpha21.r1[d1.u[i]+1,trt[i]]*u0[i]
  d1.u[i] ~ dbern(pi1.u.r1[i,trt[i]])
  logit(pi1.u.r1[i,trt[i]]) <- gamma01.r1[trt[i]]+
  gamma11.r1[trt[i]]*log(c0[i])+gamma21.r1[trt[i]]*u0[i]

  c2[i] ~ dlnorm(nu2.c.r1[i,trt[i]],tau2.c.r1[d2.c[i]+1,trt[i]])
  nu2.c.r1[i,trt[i]]<- beta02.r1[d2.c[i]+1,trt[i]]+
  beta12.r1[d2.c[i]+1,trt[i]]*log(c1[i])+beta22.r1[d2.c[i]+1,trt[i]]*u1[i]
  d2.c[i] ~ dbern(pi2.c.r1[i,trt[i]])
  logit(pi2.c.r1[i,trt[i]]) <- zeta02.r1[trt[i]]+
  zeta12.r1[trt[i]]*log(c1[i])+zeta22.r1[trt[i]]*u1[i]
  u2[i] ~ dbeta(a2.r1[i,trt[i]],b2.r1[i,trt[i]])
  a2.r1[i,trt[i]] <- nu2.u.r1[i,trt[i]]*(nu2.u.r1[i,trt[i]]*
  (1-nu2.u.r1[i,trt[i]])/pow(sigma2.u.r1[d2.u[i]+1,trt[i]],2)-1)
  b2.r1[i,trt[i]] <- (1-nu2.u.r1[i,trt[i]])*(nu2.u.r1[i,trt[i]]*
  (1-nu2.u.r1[i,trt[i]])/ pow(sigma2.u.r1[d2.u[i]+1,trt[i]],2)-1)
  logit(nu2.u.r1[i,trt[i]]) <- alpha02.r1[d2.u[i]+1,trt[i]]+
  alpha12.r1[d2.u[i]+1,trt[i]]*log(c2[i])+alpha22.r1[d2.u[i]+1,trt[i]]*u0[i]
  d2.u[i] ~ dbern(pi2.u.r1[i,trt[i]])
  logit(pi2.u.r1[i,trt[i]]) <- gamma02.r1[trt[i]]+
  gamma12.r1[trt[i]]*log(c1[i])+gamma22.r1[trt[i]]*u1[i]
 }

 # obtain standard deviations from precisions for cj

 for(d in 1:2){
   for(t in 1:2){
       tau0.c.r1[d,t] <- 1/pow(sigma0.c.r1[d,t],2)
       tau1.c.r1[d,t] <- 1/pow(sigma1.c.r1[d,t],2)
       tau2.c.r1[d,t] <- 1/pow(sigma2.c.r1[d,t],2)
   }
 }

# priors on model for uj and cj
# priors on cj>0 and uj<1

 for(t in 1:2){
    alpha00.r1[1,t] ~ dnorm(0,0.0001)
    alpha01.r1[1,t] ~ dnorm(0,0.0001)
    alpha02.r1[1,t] ~ dnorm(0,0.0001)
    alpha10.r1[1,t] ~ dnorm(0,0.0001)
    alpha11.r1[1,t] ~ dnorm(0,0.0001)
    alpha12.r1[1,t] ~ dnorm(0,0.0001)
    alpha21.r1[1,t] ~ dnorm(0,0.0001)
    alpha22.r1[1,t] ~ dnorm(0,0.0001)
    sigma0.u.limit.r1[t] <- sqrt(ilogit(alpha00.r1[1,t])*
    (1-ilogit(alpha00.r1[1,t])))
    sigma0.u.r1[1,t] ~ dunif(0,sigma0.u.limit.r1[t])
```

```
    sigma1.u.limit.r1[t] <- sqrt(ilogit(alpha01.r1[1,t])*
    (1-ilogit(alpha01.r1[1,t])))
    sigma1.u.r1[1,t] ~ dunif(0,sigma1.u.limit.r1[t])
    sigma2.u.limit.r1[t] <- sqrt(ilogit(alpha02.r1[1,t])*
    (1-ilogit(alpha02.r1[1,t])))
    sigma2.u.r1[1,t] ~ dunif(0,sigma2.u.limit.r1[t])
    nu0.c.r1[1,t] ~ dnorm(0,0.0001)
    beta01.r1[1,t] ~ dnorm(0,0.0001)
    beta02.r1[1,t] ~ dnorm(0,0.0001)
    beta11.r1[1,t] ~ dnorm(0,0.0001)
    beta12.r1[1,t] ~ dnorm(0,0.0001)
    beta21.r1[1,t] ~ dnorm(0,0.0001)
    beta22.r1[1,t] ~ dnorm(0,0.0001)
    sigma0.c.r1[1,t] ~ dunif(0,10000)
    sigma1.c.r1[1,t] ~ dunif(0,10000)
    sigma2.c.r1[1,t] ~ dunif(0,10000)

# priors on cj=0 and uj=1

    alpha00.r1[2,t] <- logit(0.999999)
    alpha01.r1[2,t] <- logit(0.999999)
    alpha02.r1[2,t] <- logit(0.999999)
    alpha10.r1[2,t] <- 0
    alpha11.r1[2,t] <- 0
    alpha12.r1[2,t] <- 0
    alpha21.r1[2,t] <- 0
    alpha22.r1[2,t] <- 0
    sigma0.u.r1[2,t] <- 0.000001
    sigma1.u.r1[2,t] <- 0.000001
    sigma2.u.r1[2,t] <- 0.000001
    nu0.c.r1[2,t] <- log(pow(0.000001,2)/
    (sqrt(pow(0.000001,2)+pow(0.000001,2))))
    beta01.r1[2,t] <- log(pow(0.000001,2)/
    (sqrt(pow(0.000001,2)+pow(0.000001,2))))
    beta02.r1[2,t] <- log(pow(0.000001,2)/
    (sqrt(pow(0.000001,2)+pow(0.000001,2))))
    beta11.r1[2,t] <- 0
    beta12.r1[2,t] <- 0
    beta21.r1[2,t] <- 0
    beta22.r1[2,t] <- 0
    sigma0.c.r1[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))
    sigma1.c.r1[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))
    sigma2.c.r1[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))

 # priors on model for d.uj and d.cj

    gamma00.r1[t] ~ dnorm(0,0.0001)
    gamma01.r1[t] ~ dnorm(0,0.0001)
    gamma02.r1[t] ~ dnorm(0,0.0001)
    gamma10.r1[t] ~ dnorm(0,0.0001)
    gamma11.r1[t] ~ dnorm(0,0.0001)
    gamma12.r1[t] ~ dnorm(0,0.0001)
    gamma21.r1[t] ~ dnorm(0,0.0001)
    gamma22.r1[t] ~ dnorm(0,0.0001)
    pi0.c.r1[t] ~ dunif(0,1)
    zeta01.r1[t] ~ dnorm(0,0.0001)
    zeta02.r1[t] ~ dnorm(0,0.0001)
    zeta11.r1[t] ~ dnorm(0,0.0001)
    zeta12.r1[t] ~ dnorm(0,0.0001)
    zeta21.r1[t] ~ dnorm(0,0.0001)
    zeta22.r1[t] ~ dnorm(0,0.0001)
```

```
}


# model for the responses in the non-completers (r2) stratified by treatment

for(i in N.r1+1:N){
   c0.r2[i] ~ dlnorm(nu0.c.r2[d0.c[i]+1,trt[i]],tau0.c.r2[d0.c[i]+1,trt[i]])
   d0.c[i] ~ dbern(pi0.c.r2[trt[i]])
   u0[i] ~ dbeta(a0.r2[i,trt[i]],b0.r2[i,trt[i]])
   a0.r2[i,trt[i]] <- nu0.u.r2[i,trt[i]]*(nu0.u.r2[i,trt[i]]*
   (1-nu0.u.r2[i,trt[i]])/pow(sigma0.u.r2[d0.u[i]+1,trt[i]],2)-1)
   b0.r2[i,trt[i]] <- (1-nu0.u.r2[i,trt[i]])*(nu0.u.r2[i,trt[i]]*
   (1-nu0.u.r2[i,trt[i]])/pow(sigma0.u.r2[d0.u[i]+1,trt[i]],2)-1)
   logit(nu0.u.r2[i,trt[i]]) <- alpha00.r2[d0.u[i]+1,trt[i]]+
   alpha10.r2[d0.u[i]+1,trt[i]]*log(c0[i])
   d0.u[i] ~ dbern(pi0.u.r2[i,trt[i]])
   logit(pi0.u.r2[i,trt[i]]) <- gamma00.r2[trt[i]]+
   gamma10.r2[trt[i]]*log(c0[i])

  c1[i] ~ dlnorm(nu1.c.r2[i,trt[i]],tau1.c.r2[d1.c[i]+1,trt[i]])
  nu1.c.r2[i,trt[i]] <- beta01.r2[d1.c[i]+1,trt[i]]+
  beta11.r2[d1.c[i]+1,trt[i]]*log(c0[i])+beta21.r2[d1.c[i]+1,trt[i]]*u0[i]
  d1.c[i] ~ dbern(pi1.c.r2[i,trt[i]])
  logit(pi1.c.r2[i,trt[i]]) <- zeta01.r2[trt[i]]+
  zeta11.r2[trt[i]]*log(c0[i])+zeta21.r2[trt[i]]*u0[i]
  u1[i] ~ dbeta(a1.r2[i,trt[i]],b1.r2[i,trt[i]])
  a1.r2[i,trt[i]] <- nu1.u.r2[i,trt[i]]*(nu1.u.r2[i,trt[i]]*
  (1-nu1.u.r2[i,trt[i]])/pow(sigma1.u.r2[d1.u[i]+1,trt[i]],2)-1)
  b1.r2[i,trt[i]] <- (1-nu1.u.r2[i,trt[i]])*(nu1.u.r2[i,trt[i]]*
  (1-nu1.u.r2[i,trt[i]])/pow(sigma1.u.r2[d1.u[i]+1,trt[i]],2)-1)
  logit(nu1.u.r2[i,trt[i]]) <- alpha01.r2[d1.u[i]+1,trt[i]]+
  alpha11.r2[d1.u[i]+1,trt[i]]*log(c1[i])+alpha21.r2[d1.u[i]+1,trt[i]]*u0[i]
  d1.u[i] ~ dbern(pi1.u.r2[i,trt[i]])
  logit(pi1.u.r2[i,trt[i]]) <- gamma01.r2[trt[i]]+
  gamma11.r2[trt[i]]*log(c0[i])+gamma21.r2[trt[i]]*u0[i]

  c2[i] ~ dlnorm(nu2.c.r2[i,trt[i]],tau2.c.r2[d2.c[i]+1,trt[i]])
  nu2.c.r2[i,trt[i]] <- beta02.r2[d2.c[i]+1,trt[i]]+
  beta12.r2[d2.c[i]+1,trt[i]]*log(c1[i])+beta22.r2[d2.c[i]+1,trt[i]]*u1[i]
  d2.c[i] ~ dbern(pi2.c.r2[i,trt[i]])
  logit(pi2.c.r2[i,trt[i]]) <- zeta02.r2[trt[i]]+
  zeta12.r2[trt[i]]*log(c1[i])+zeta22.r2[trt[i]]*u1[i]
  u2[i] ~ dbeta(a2.r2[i,trt[i]],b2.r2[i,trt[i]])
  a2.r2[i,trt[i]] <- nu2.u.r2[i,trt[i]]*(nu2.u.r2[i,trt[i]]*
  (1-nu2.u.r2[i,trt[i]])/pow(sigma2.u.r2[d2.u[i]+1,trt[i]],2)-1)
  b2.r2[i,trt[i]] <- (1-nu2.u.r2[i,trt[i]])*(nu2.u.r2[i,trt[i]]*
  (1-nu2.u.r2[i,trt[i]])/pow(sigma2.u.r2[d2.u[i]+1,trt[i]],2)-1)
  logit(nu2.u.r2[i,trt[i]]) <- alpha02.r2[d2.u[i]+1,trt[i]]+
  alpha12.r2[d2.u[i]+1,trt[i]]*log(c2[i])+alpha22.r2[d2.u[i]+1,trt[i]]*u0[i]
  d2.u[i] ~ dbern(pi2.u.r2[i,trt[i]])
  logit(pi2.u.r2[i,trt[i]]) <- gamma02.r2[trt[i]]+
  gamma12.r2[trt[i]]*log(c1[i])+gamma22.r2[trt[i]]*u1[i]
}

 # obtain standard deviations from precisions for cj

 for(d in 1:2){
    for(t in 1:2){
        tau0.c.r2[d,t] <- 1/pow(sigma0.c.r2[d,t],2)
        tau1.c.r2[d,t] <- 1/pow(sigma1.c.r2[d,t],2)
        tau2.c.r2[d,t] <- 1/pow(sigma2.c.r2[d,t],2)
```

```
   }
}

# priors on model for uj and cj
# priors on cj>0 and uj<1

for(t in 1:2){
  alpha00.r2[1,t] ~ dnorm(0,0.0001)
  alpha01.r2[1,t] ~ dnorm(0,0.0001)
  alpha02.r2[1,t] ~ dnorm(0,0.0001)
  alpha10.r2[1,t] ~ dnorm(0,0.0001)
  alpha11.r2[1,t] ~ dnorm(0,0.0001)
  alpha12.r2[1,t] ~ dnorm(0,0.0001)
  alpha21.r2[1,t] ~ dnorm(0,0.0001)
  alpha22.r2[1,t] ~ dnorm(0,0.0001)
  sigma0.u.limit.r2[t] <- sqrt(ilogit(alpha00.r2[1,t])*
  (1-ilogit(alpha00.r2[1,t])))
  sigma0.u.r2[1,t] ~ dunif(0,sigma0.u.limit.r2[t])
  sigma1.u.limit.r2[t] <- sqrt(ilogit(alpha01.r2[1,t])*
  (1-ilogit(alpha01.r2[1,t])))
  sigma1.u.r2[1,t] ~ dunif(0,sigma1.u.limit.r2[t])
  sigma2.u.limit.r2[t] <- sqrt(ilogit(alpha02.r2[1,t])*
  (1-ilogit(alpha02.r2[1,t])))
  sigma2.u.r2[1,t] ~ dunif(0,sigma2.u.limit.r2[t])
  nu0.c.r2[1,t] ~ dnorm(0,0.0001)
  beta01.r2[1,t] ~ dnorm(0,0.0001)
  beta02.r2[1,t] ~ dnorm(0,0.0001)
  beta11.r2[1,t] ~ dnorm(0,0.0001)
  beta12.r2[1,t] ~ dnorm(0,0.0001)
  beta21.r2[1,t] ~ dnorm(0,0.0001)
  beta22.r2[1,t] ~ dnorm(0,0.0001)
  sigma0.c.r2[1,t] ~ dunif(0,10000)
  sigma1.c.r2[1,t] ~ dunif(0,10000)
  sigma2.c.r2[1,t] ~ dunif(0,10000)

# priors on for cj=0 and uj=1

  alpha00.r2[2,t] <- logit(0.999999)
  alpha01.r2[2,t] <- logit(0.999999)
  alpha02.r2[2,t] <- logit(0.999999)
  alpha10.r2[2,t] <- 0
  alpha11.r2[2,t] <- 0
  alpha12.r2[2,t] <- 0
  alpha21.r2[2,t] <- 0
  alpha22.r2[2,t] <- 0
  sigma0.u.r2[2,t] <- 0.000001
  sigma1.u.r2[2,t] <- 0.000001
  sigma2.u.r2[2,t] <- 0.000001
  nu0.c.r2[2,t] <- log(pow(0.000001,2)/
  (sqrt(pow(0.000001,2)+pow(0.000001,2))))
  beta01.r2[2,t] <- log(pow(0.000001,2)/
  (sqrt(pow(0.000001,2)+pow(0.000001,2))))
  beta02.r2[2,t] <- log(pow(0.000001,2)/
  (sqrt(pow(0.000001,2)+pow(0.000001,2))))
  beta11.r2[2,t] <- 0
  beta12.r2[2,t] <- 0
  beta21.r2[2,t] <- 0
  beta22.r2[2,t] <- 0
  sigma0.c.r2[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))
  sigma1.c.r2[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))
  sigma2.c.r2[2,t] <- sqrt(log(pow(0.000001,2)/pow(0.000001,2)+1))
```

```
   # priors on model for d.uj and d.cj

    gamma00.r2[t] ~ dnorm(0,0.0001)
    gamma01.r2[t] ~ dnorm(0,0.0001)
    gamma02.r2[t] ~ dnorm(0,0.0001)
    gamma12.r2[t] ~ dnorm(0,0.0001)
    gamma22.r2[t] ~ dnorm(0,0.0001)
    pi0.c.r2[t] ~ dunif(0,1)
    zeta01.r2[t] ~ dnorm(0,0.0001)
    zeta02.r2[t] ~ dnorm(0,0.0001)
    zeta12.r2[t] ~ dnorm(0,0.0001)
    zeta22.r2[t] ~ dnorm(0,0.0001)
}

    gamma10.r2[1] ~ dnorm(0,0.0001)
    gamma11.r2[1] ~ dnorm(0,0.0001)
    gamma21.r2[1] ~ dnorm(0,0.0001)
    gamma10.r2[2] <- 0
    gamma11.r2[2] <- 0
    gamma21.r2[2] <- 0
    zeta11.r2[1] <- 0
    zeta21.r2[1] <- 0
    zeta11.r2[2] ~ dnorm(0,0.0001)
    zeta21.r2[2] ~ dnorm(0,0.0001)
}
```

B.3.2 Monte Carlo Integration and Marginal Means Computation

R code for the MCI used to approximate the posterior distribution of the marginal utility and cost means for the model in Chapter 6 under the nonignorable scenarios $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$.

```
# N.sim = 20,000 number of the iterations
# N.rep = 40,000 replicated samples at each iteration
# r1 = completers pattern
# r2 = non-completers pattern
# r.con = number of observed patterns in the control
# r.int = number of observed patterns in the intervention

# all relevant parameters saved from JAGS into R objects
# e.g. nu0.c.r1[1:N.sim, 1:2] = c0 mean for r1
# psi.obs.con[1:N.sim,1:r.con] = patterns' probabilities (control)
# psi.obs.int[1:N.sim, 1:r.int] = patterns' probabilities (intervention)

# compute Monte Carlo estimates for mean uj and cj at each time j=0,1,2

mu.c0.r1 <- mu.c1.r1 <- mu.c2.r1 <- matrix(NA,N.sim,2)
mu.u0.r1 <- mu.u1.r1 <- mu.u2.r1 <- matrix(NA,N.sim,2)
mu.c0.r2 <- mu.c1.r2 <- mu.c2.r2 <- matrix(NA,N.sim,2)
mu.u0.r2 <- mu.u1.r2 <- mu.u2.r2 <- matrix(NA,N.sim,2)

# control

for(k in 1:N.sim){
   mu.c0.r1[k,1] <- mean(rlnorm(N.rep,nu0.c.r1[k,1],sigma0.c.r1[k,1]))
   mu.c1.r1[k,1] <- mean(rlnorm(N.rep,nu1.c.r1[k,1],sigma1.c.r1[k,1]))
   mu.c2.r1[k,1] <- mean(rlnorm(N.rep,nu2.c.r1[k,1],sigma2.c.r1[k,1]))
   mu.u0.r1[k,1] <- mean(rbeta(N.rep,a0.r1[k,1],b0.r1[k,1]))
```

```
   mu.u1.r1[k,1] <- mean(rbeta(N.rep,a1.r1[k,1],b1.r1[k,1]))
   mu.u2.r1[k,1] <- mean(rbeta(N.rep,a2.r1[k,1],b2.r1[k,1]))
   mu.c0.r2[k,1] <- mean(rlnorm(N.rep,nu0.c.r2[k,1],sigma0.c.r2[k,1]))
   mu.c1.r2[k,1] <- mean(rlnorm(N.rep,nu1.c.r2[k,1],sigma1.c.r2[k,1]))
   mu.c2.r2[k,1] <- mean(rlnorm(N.rep,nu2.c.r2[k,1],sigma2.c.r2[k,1]))
   mu.u0.r2[k,1] <- mean(rbeta(N.rep,a0.r2[k,1],b0.r2[k,1]))
   mu.u1.r2[k,1] <- mean(rbeta(N.rep,a1.r2[k,1],b1.r2[k,1]))
   mu.u2.r2[k,1] <- mean(rbeta(N.rep,a2.r2[k,1],b2.r2[k,1]))

# intervention

   mu.c0.r1[k,2] <- mean(rlnorm(N.rep,nu0.c.r1[k,2],sigma0.c.r1[k,2]))
   mu.c1.r1[k,2] <- mean(rlnorm(N.rep,nu1.c.r1[k,2],sigma1.c.r1[k,2]))
   mu.c2.r1[k,2] <- mean(rlnorm(N.rep,nu2.c.r1[k,2],sigma2.c.r1[k,2]))
   mu.u0.r1[k,2] <- mean(rbeta(N.rep,a0.r1[k,2],b0.r1[k,2]))
   mu.u1.r1[k,2] <- mean(rbeta(N.rep,a1.r1[k,2],b1.r1[k,2]))
   mu.u2.r1[k,2] <- mean(rbeta(N.rep,a2.r1[k,2],b2.r1[k,2]))
   mu.c0.r2[k,2] <- mean(rlnorm(N.rep,nu0.c.r2[k,2],sigma0.c.r2[k,2]))
   mu.c1.r2[k,2] <- mean(rlnorm(N.rep,nu1.c.r2[k,2],sigma1.c.r2[k,2]))
   mu.c2.r2[k,2] <- mean(rlnorm(N.rep,nu2.c.r2[k,2],sigma2.c.r2[k,2]))
   mu.u0.r2[k,2] <- mean(rbeta(N.rep,a0.r2[k,2],b0.r2[k,2]))
   mu.u1.r2[k,2] <- mean(rbeta(N.rep,a1.r2[k,2],b1.r2[k,2]))
   mu.u2.r2[k,2] <- mean(rbeta(N.rep,a2.r2[k,2],b2.r2[k,2]))
}


# priors on the sensitivity parameters
# flat prior

delta.c0.flat <- runif(N.sim,0,2*sd(c0))
delta.u0.flat <- runif(N.sim,-2*sd(u0),0)
delta.c1.flat <- runif(N.sim,0,2*sd(c1))
delta.u1.flat <- runif(N.sim,-2*sd(u1),0)
delta.c2.flat <- runif(N.sim,0,2*sd(c2))
delta.u2.flat <- runif(N.sim,-2*sd(u2),0)

# skew0 prior

delta.c0.skew0<- 2*sd(c0)*(1-sqrt(runif(N.sim,0,1))
delta.u0.skew0<- -2*sd(u0)*(1-sqrt(runif(N.sim,0,1))
delta.c1.skew0<- 2*sd(c1)*(1-sqrt(runif(N.sim,0,1))
delta.u1.skew0<- -2*sd(u1)*(1-sqrt(runif(N.sim,0,1))
delta.c2.skew0<- 2*sd(c2)*(1-sqrt(runif(N.sim,0,1))
delta.u2.skew0<- -2*sd(u2)*(1-sqrt(runif(N.sim,0,1))

#skew1 prior

delta.c0.skew1 <- 2*sd(c0)*sqrt(runif(N.sim,0,1)
delta.u0.skew1 <- -2*sd(u0)*sqrt(runif(N.sim,0,1)
delta.c1.skew1 <- 2*sd(c1)*sqrt(runif(N.sim,0,1)
delta.u1.skew1 <- -2*sd(u1)*sqrt(runif(N.sim,0,1)
delta.c2.skew1 <- 2*sd(c2)*sqrt(runif(N.sim,0,1)
delta.u2.skew1 <- -2*sd(u2)*sqrt(runif(N.sim,0,1)

# stratify the marginal means by pattern and arm under delta=0
# control

mu.c0.r.con <- mu.c1.r.con <- mu.c2.r.con <- matrix(NA,N.sim,9)
mu.c0.r.con[,1] <- mu.c0.r1[,1]
mu.c0.r.con[,2:9] <- mu.c0.r2[,1]
mu.c1.r.con[,1] <- mu.c1.r1[,1]
```

```
mu.c1.r.con[,2:9] <- mu.c1.r2[,1]
mu.c2.r.con[,1] <- mu.c2.r1[,1]
mu.c2.r.con[,2:9] <- mu.c2.r2[,1]
mu.u0.r.con <- mu.u1.r.con <- mu.u2.r.con <- matrix(NA,N.sim,9)
mu.u0.r.con[,1] <- mu.u0.r1[,1]
mu.u0.r.con[,2:9] <- mu.u0.r2[,1]
mu.u1.r.con[,1] <- mu.u1.r1[,1]
mu.u1.r.con[,2:9] <- mu.u1.r2[,1]
mu.u2.r.con[,1] <- mu.u2.r1[,1]
mu.u2.r.con[,2:9] <- mu.u2.r2[,1]

# intervention

mu.c0.r.int <- mu.c1.r.int <- mu.c2.r.int <- matrix(NA,N.sim,6)
mu.c0.r.int[,1] <- mu.c0.r1[,2]
mu.c0.r.int[,2:6] <- mu.c0.r2[,2]
mu.c1.r.int[,1] <- mu.c1.r1[,2]
mu.c1.r.int[,2:6] <- mu.c1.r2[,2]
mu.c2.r.int[,1] <- mu.c2.r1[,2]
mu.c2.r.int[,2:6] <- mu.c2.r2[,2]
mu.u0.r.int <- mu.u1.r.int <- mu.u2.r.int <- matrix(NA,N.sim,6)
mu.u0.r.int[,1] <- mu.u0.r1[,2]
mu.u0.r.int[,2:6] <- mu.u0.r2[,2]
mu.u1.r.int[,1] <- mu.u1.r1[,2]
mu.u1.r.int[,2:6] <- mu.u1.r2[,2]
mu.u2.r.int[,1] <- mu.u2.r1[,2]
mu.u2.r.int[,2:6] <- mu.u2.r2[,2]

# marginal means for missing data in each pattern and arm under each prior

mu.c1.r.con.flat <- mu.c1.r.con.skew0 <- mu.c1.r.con.skew1 <- mu.c1.r.con
mu.c2.r.con.flat <- mu.c2.r.con.skew0 <- mu.c2.r.con.skew1 <- mu.c2.r.con
mu.u0.r.con.flat <- mu.u0.r.con.skew0 <- mu.u0.r.con.skew1 <- mu.u0.r.con
mu.u1.r.con.flat <- mu.u1.r.con.skew0 <- mu.u1.r.con.skew1 <- mu.u1.r.con
mu.u2.r.con.flat <- mu.u2.r.con.skew0 <- mu.u2.r.con.skew1 <- mu.u2.r.con
mu.c1.r.int.flat <- mu.c1.r.int.skew0 <- mu.c1.r.int.skew1 <- mu.c1.r.int
mu.c2.r.int.flat <- mu.c2.r.int.skew0 <- mu.c2.r.int.skew1 <- mu.c2.r.int
mu.u0.r.int.flat <- mu.u0.r.int.skew0 <- mu.u0.r.int.skew1 <- mu.u0.r.int
mu.u1.r.int.flat <- mu.u1.r.int.skew0 <- mu.u1.r.int.skew1 <- mu.u1.r.int
mu.u2.r.int.flat <- mu.u2.r.int.skew0 <- mu.u2.r.int.skew1 <- mu.u2.r.int

# control
# c, j=1 (missing for r.con=8,9)

mu.c1.r.con.flat[,c(8,9)] <- mu.c1.r2[,1]+delta.c1.flat
mu.c1.r.con.skew0[,c(8,9)] <- mu.c1.r2[,1]+delta.c1.skew0
mu.c1.r.con.skew1[,c(8,9)] <- mu.c1.r2[,1]+delta.c1.skew1

# c, j=2 (missing for r.con=6,9)

mu.c2.r.con.flat[,c(6,9)] <- mu.c2.r2[,1]+delta.c2.flat
mu.c2.r.con.skew0[,c(6,9)] <- mu.c2.r2[,1]+delta.c2.skew0
mu.c2.r.con.skew1[,c(6,9)] <- mu.c2.r2[,1]+delta.c2.skew1

# u, j=0 (missing for r.con=2,5)

mu.u0.r.con.flat[,c(2,5)] <- mu.u0.r2[,1]+delta.u0.flat
mu.u0.r.con.skew0[,c(2,5)] <- mu.u0.r2[,1]+delta.u0.skew0
mu.u0.r.con.skew1[,c(2,5)] <- mu.u0.r2[,1]+delta.u0.skew1

# u, j=1 (missing for r.con=3,5,7,8,9)
```

```
mu.u1.r.con.flat[,c(3,5,7,8,9)] <- mu.u1.r2[,1]+delta.u1.flat
mu.u1.r.con.skew0[,c(3,5,7,8,9)] <- mu.u1.r2[,1]+delta.u1.skew0
mu.u1.r.con.skew1[,c(3,5,7,8,9)] <- mu.u1.r2[,1]+delta.u1.skew1

# u, j=2 (missing for r.con=4,6,7,9)

mu.u2.r.con.flat[,c(4,6,7,9)] <- mu.u2.r2[,1]+delta.u2.flat
mu.u2.r.con.skew0[,c(4,6,7,9)] <- mu.u2.r2[,1]+delta.u2.skew0
mu.u2.r.con.skew1[,c(4,6,7,9)] <- mu.u2.r2[,1]+delta.u2.skew1

# intervention
# c, j=1 (missing for r.int=5,6)

mu.c1.r.int.flat[,c(5,6)] <- mu.c1.r2[,2]+delta.c1.flat
mu.c1.r.int.skew0[,c(5,6)] <- mu.c1.r2[,2]+delta.c1.skew0
mu.c1.r.int.skew1[,c(5,6)] <- mu.c1.r2[,2]+delta.c1.skew1

# c, j=2 (missing for r.int=6)

mu.c2.r.int.flat[,6] <- mu.c2.r2[,2]+delta.c2.flat
mu.c2.r.int.skew0[,6] <- mu.c2.r2[,2]+delta.c2.skew0
mu.c2.r.int.skew1[,6] <- mu.c2.r2[,2]+delta.c2.skew1

# u, j=0 (missing for r.int=2)

mu.u0.r.int.flat[,2] <- mu.u0.r2[,2]+delta.u0.flat
mu.u0.r.int.skew0[,2] <- mu.u0.r2[,2]+delta.u0.skew0
mu.u0.r.int.skew1[,2] <- mu.u0.r2[,2]+delta.u0.skew1

# u, j=1 (missing for r.int=3,5,6)

mu.u1.r.int.flat[,c(3,5,6)] <- mu.u1.r2[,2]+delta.u1.flat
mu.u1.r.int.skew0[,c(3,5,6)] <- mu.u1.r2[,2]+delta.u1.skew0
mu.u1.r.int.skew1[,c(3,5,6)] <- mu.u1.r2[,2]+delta.u1.skew1

# u, j=2 (missing for r.int=4,6,7,9)

mu.u2.r.int.flat[,c(4,6)] <- mu.u2.r2[,2]+delta.u2.flat
mu.u2.r.int.skew0[,c(4,6)] <- mu.u2.r2[,2]+delta.u2.skew0
mu.u2.r.int.skew1[,c(4,6)] <- mu.u2.r2[,2]+delta.u2.skew1

# marginal means across observed patterns in each group under each prior

mu.c0.bench <- mu.c1.bench <- mu.c2.bench <- matrix(NA,N.sim,2)
mu.u0.bench <- mu.u1.bench <- mu.u2.bench <- matrix(NA,N.sim,2)
mu.c0.flat <- mu.c1.flat <- mu.c2.flat <- matrix(NA,N.sim,2)
mu.u0.flat <- mu.u1.flat <- mu.u2.flat <- matrix(NA,N.sim,2)
mu.c0.skew0 <- mu.c1.skew0 <- mu.c2.skew0 <- matrix(NA,N.sim,2)
mu.u0.skew0 <- mu.u1.skew0 <- mu.u2.skew0 <- matrix(NA,N.sim,2)
mu.c0.skew1 <- mu.c1.skew1 <- mu.c2.skew1 <- matrix(NA,N.sim,2)
mu.u0.skew1 <- mu.u1.skew1 <- mu.u2.skew1 <- matrix(NA,N.sim,2)

# control

mu.c0.bench[,1] <- rowSums(mu.c0.r.con*psi.obs.con)
mu.c1.bench[,1] <- rowSums(mu.c1.r.con*psi.obs.con)
mu.c2.bench[,1] <- rowSums(mu.c2.r.con*psi.obs.con)
mu.u0.bench[,1] <- rowSums(mu.u0.r.con*psi.obs.con)
mu.u1.bench[,1] <- rowSums(mu.u1.r.con*psi.obs.con)
mu.u2.bench[,1] <- rowSums(mu.u2.r.con*psi.obs.con)
```

```
mu.c1.flat[,1] <- rowSums(mu.c1.r.con.flat*psi.obs.con)
mu.c2.flat[,1] <- rowSums(mu.c2.r.con.flat*psi.obs.con)
mu.u0.flat[,1] <- rowSums(mu.u0.r.con.flat*psi.obs.con)
mu.u1.flat[,1] <- rowSums(mu.u1.r.con.flat*psi.obs.con)
mu.u2.flat[,1] <- rowSums(mu.u2.r.con.flat*psi.obs.con)
mu.c1.skew0[,1] <- rowSums(mu.c1.r.con.skew0*psi.obs.con)
mu.c2.skew0[,1] <- rowSums(mu.c2.r.con.skew0*psi.obs.con)
mu.u0.skew0[,1] <- rowSums(mu.u0.r.con.skew0*psi.obs.con)
mu.u1.skew0[,1] <- rowSums(mu.u1.r.con.skew0*psi.obs.con)
mu.u2.skew0[,1] <- rowSums(mu.u2.r.con.skew0*psi.obs.con)
mu.c1.skew1[,1] <- rowSums(mu.c1.r.con.skew1*psi.obs.con)
mu.c2.skew1[,1] <- rowSums(mu.c2.r.con.skew1*psi.obs.con)
mu.u0.skew1[,1] <- rowSums(mu.u0.r.con.skew1*psi.obs.con)
mu.u1.skew1[,1] <- rowSums(mu.u1.r.con.skew1*psi.obs.con)
mu.u2.skew1[,1] <- rowSums(mu.u2.r.con.skew1*psi.obs.con)

# intervention

mu.c0.bench[,2] <- rowSums(mu.c0.r.int*psi.obs.int)
mu.c1.bench[,2] <- rowSums(mu.c1.r.int*psi.obs.int)
mu.c2.bench[,2] <- rowSums(mu.c2.r.int*psi.obs.int)
mu.u0.bench[,2] <- rowSums(mu.u0.r.int*psi.obs.int)
mu.u1.bench[,2] <- rowSums(mu.u1.r.int*psi.obs.int)
mu.u2.bench[,2] <- rowSums(mu.u2.r.int*psi.obs.int)
mu.c1.flat[,2] <- rowSums(mu.c1.r.int.flat*psi.obs.int)
mu.c2.flat[,2] <- rowSums(mu.c2.r.int.flat*psi.obs.int)
mu.u0.flat[,2] <- rowSums(mu.u0.r.int.flat*psi.obs.int)
mu.u1.flat[,2] <- rowSums(mu.u1.r.int.flat*psi.obs.int)
mu.u2.flat[,2] <- rowSums(mu.u2.r.int.flat*psi.obs.int)
mu.c1.skew0[,2] <- rowSums(mu.c1.r.int.skew0*psi.obs.int)
mu.c2.skew0[,2] <- rowSums(mu.c2.r.int.skew0*psi.obs.int)
mu.u0.skew0[,2] <- rowSums(mu.u0.r.int.skew0*psi.obs.int)
mu.u1.skew0[,2] <- rowSums(mu.u1.r.int.skew0*psi.obs.int)
mu.u2.skew0[,2] <- rowSums(mu.u2.r.int.skew0*psi.obs.int)
mu.c1.skew1[,2] <- rowSums(mu.c1.r.int.skew1*psi.obs.int)
mu.c2.skew1[,2] <- rowSums(mu.c2.r.int.skew1*psi.obs.int)
mu.u0.skew1[,2] <- rowSums(mu.u0.r.int.skew1*psi.obs.int)
mu.u1.skew1[,2] <- rowSums(mu.u1.r.int.skew1*psi.obs.int)
mu.u2.skew1[,2] <- rowSums(mu.u2.r.int.skew1*psi.obs.int)
```

Appendix C

Supplementary Analyses

In this Appendix we present information and results from supplementary analyses based on the models presented in Chapter 5 and Chapter 6.

C.1 Supplementary Analyses: Chapter 2

C.1.1 Alternative versions of the quality evaluation scheme

As a way of supporting the recommended scheme which weights the description (D), Method (M) and Limitations (L) components using an allocation of 3:2:1 (denoted as version I), we have also considered two alternative strategies. The first, assigns equal weight to each component with an allocation of 1:1:1 (denoted as version II), while the second assigns the weights assuming a greater discrepancy between the importance associated with each component using an allocation of 6:3:1 (denoted as version III).

The three schemes are compared in Table C.1, which shows the score assigned to each of the three components in the analysis (Description, Method and Limitations) by each weight allocation version.

	Version I (3:2:1)			Version II (1:1:1)			Version III (6:3:1)		
Content	Score			Score			Score		
	D	M	L	D	M	L	D	M	L
Full	6	4	2	2	2	2	12	6	3
Partial	3	2	1	1	1	1	6	3	1
Limitations	0	0	0	0	0	0	0	0	0

Table C.1: Comparison of three weighting schemes, based on the information provided on missingness in each component of the analysis: Description (D), Method (M) and Limitations (L)

Table C.2 shows the different scoring system and the classification into ordered categories based on the total scores associated with the three alternative versions of the scheme.

	Version I (3:2:1)	Version II (1:1:1)	Version III (6:3:1)
Category	Total score	Total score	Total score
A	12	6	20
B	9-11	5	15-19
C	6-8	4	10-14
D	3-5	2-3	5-9
E	0-2	0-1	0-4

Table C.2: Scoring system associated with each group category in three weight allocation versions of the Quality Evaluation Scheme.

Table C.3 (costs) and Table C.4 (effects) show the number of articles falling within each category defined by the three different schemes. The diagram representation visually compares the original scheme (version I) against the two alternative schemes (version II and III) used. The categories indicated on the right-hand side of the graph, allow comparison between the three scoring schemes and grouping methods. Since no major differences can to be detected, we believe that our original weighting scheme can be considered as robust to (reasonable) changes. Other extreme weight combinations could be tested (such as reversing the initial weight assignment from a ratio 3:2:1 to a ratio 1:2:3) but the meaningfulness of such schemes is difficult to justify

Table C.3 shows limited differences in the classification of the articles for missing costs in 2009-2015 across the categories of the scheme between the versions compared. Specifically, compared with version I, there is a decrease in the number of articles for category B (-1 in version II and -2 in version III), C (-5 in version II and -4 in version III) and D (-1 only in version III), and an increase for category D (+6 in version II and +7 in version III).

	Version I (3:2:1)	Version II (1:1:1)	Version III (6:3:1)
Category	# Articles	# Articles	# Articles
A	5	5	5
B	6	5	4
C	14	9	10
D	17	23	24
E	39	39	38

Table C.3: Missing cost articles distribution (total number of articles $= 81$) across the categories of the Quality Evaluation Scheme for the three weight allocation versions compared.

Table C.4 shows limited differences in the classification of the articles for missing effects in 2009-2015 across the categories of the scheme between the versions compared. Specifically, compared with version I, there is a decrease in the number of articles for category B (-3 in version II and -1 in version III) and C (-10 in version II and -2 in version III), with an increase for category D (+13 only in version II) or category E (+3 only in version III).

	Version I (3:2:1)	Version II (1:1:1)	Version III (6:3:1)
Category	# Articles	# Articles	# Articles
A	5	5	5
B	5	2	4
C	18	8	16
D	16	29	16
E	37	37	40

Table C.4: Missing effect articles distribution (total number of articles $= 81$) across the categories of the Quality Evaluation Scheme for the three weight allocation versions compared.

As can be seen from Figure C.1, based on the information provided in the previous tables, the overall distribution of the articles across the different graded categories indicated by letters is not greatly affected. Indeed, no substantial difference is detected for the classification of the articles between version I and III of the scheme. With respect to version II, the largest change is related to an increase in the number of articles falling within category D in version II compared with version I, due to the more limited structure and fewer scores for the former. Such a difference is mostly due to the larger weight assigned to the Description component in the original version, which allows it to more flexibly capture the difference across the articles' quality performance compared to the equal weight version.

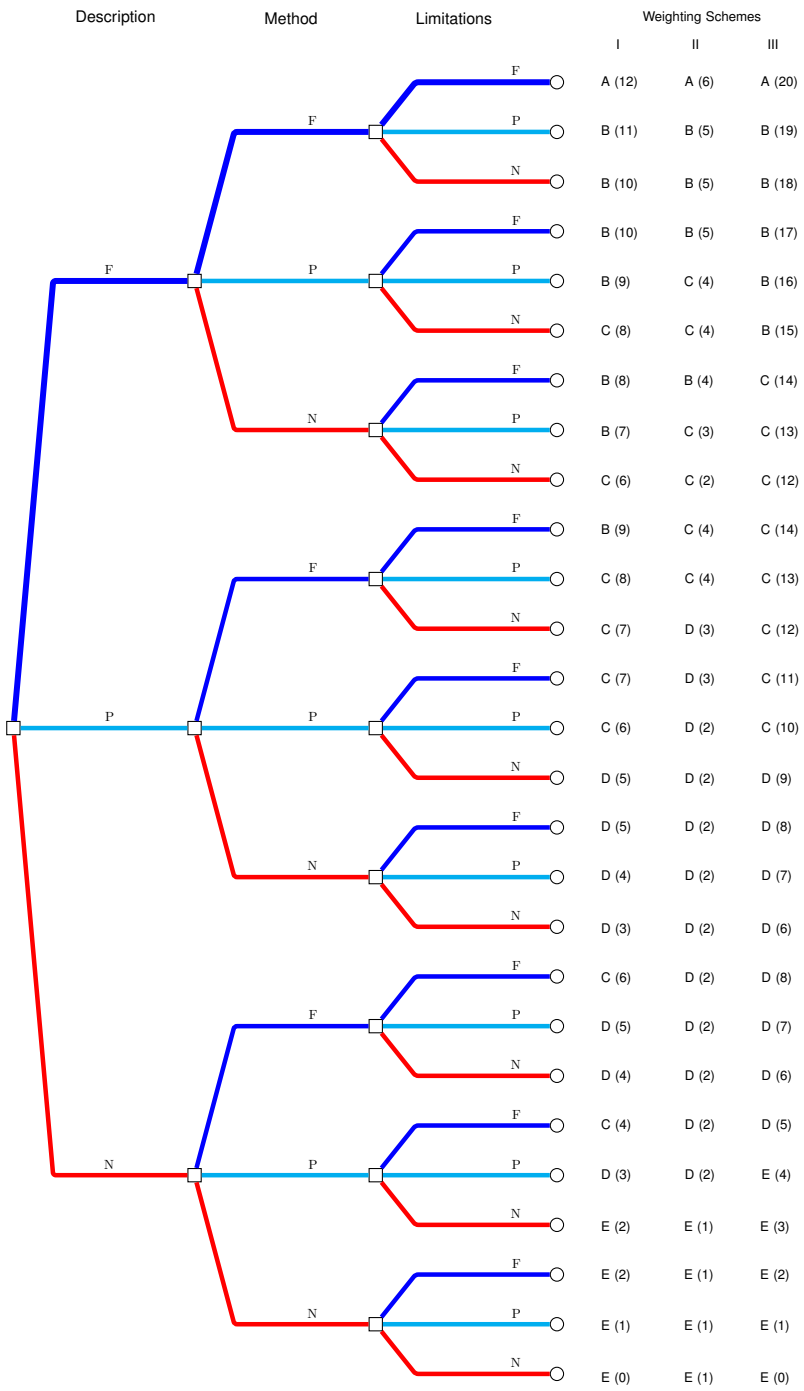


Figure C.1: Diagram representation for the quality score grades (E-A) based on final scores (0 - 12) in three alternative weight allocation schemes for the Description, Method and Limitations components. The weight allocations are 3:2:1 (version I), 1:1:1 (version II) and 6:3:1 (version III). Branches colour represents the different way information can be provided: Red=Null information (N), Light Blue=Partial information (P), Blue=Full information (F).

C.1.2 Robustness Method Analysis

Table C.5 and Table C.6 show the number of the articles in the literature review in the period (2003-2009) associated with each combination of base-case and robustness methods for missing costs and effects, respectively.

Table C.7 and Table C.8 show the number of the articles in the literature review in the period (2009-2015) associated with each combination of base-case and robustness methods for missing costs and effects, respectively.

Costs		Robustness method									
		None	Others	Lin Ext	LVCF	Mean	Cond	CCA	MI	More than one	
Base-case method	Unclear	20	0	0	0	0	0	0	0	0	
	Others	9	0	0	0	0	0	1	1	1	20
	Lin Ext	2	0	0	0	0	0	0	0	0	11
	LVCF	1	0	0	0	0	0	0	0	1	2
	Mean	8	0	0	0	0	0	1	0	0	9
	Cond	4	0	0	0	0	0	1	2	1	8
	CCA	19	2	0	1	0	0	0	4	1	27
	MI	6	0	0	0	0	0	1	0	1	8
		69	2	0	1	0	0	4	7	5	

Table C.5: Comparison of methods used in the base-case analysis (rows) and those used as alternatives in a robustness analysis (columns) for the articles between 2003-2009 for missing costs. The total number of articles using each base-case and robustness method is reported in the last column at the right of the table and in the last row at the bottom of the table. Legend: unspecified methods (Unclear), other methods (Others), Linear Extrapolation (Lin Ext), Last Value Carried Forward (LVCF), Mean Imputation (Mean), Conditional Imputation (Cond), Complete Case Analysis (CCA), Multiple Imputation (MI).

Effects		Robustness method									
		None	Others	Lin Ext	LVCF	Mean	Cond	CCA	MI	More than one	
Base-case method	Unclear	21	0	0	0	0	0	0	0	0	21
	Others	7	0	0	0	0	0	0	1	0	8
	Lin Ext	3	0	0	0	0	0	0	1	0	4
	LVCF	5	1	0	0	0	0	1	0	2	9
	Mean	6	1	0	0	0	0	0	0	0	7
	Cond	1	0	0	0	0	0	0	0	1	2
	CCA	16	0	0	1	2	1	0	3	1	24
	MI	8	0	0	0	0	0	3	1	1	13
		67	2	0	1	2	1	4	6	5	

Table C.6: Comparison of methods used in the base-case analysis (rows) and those used as alternatives in a robustness analysis (columns) for the articles between 2003-2009 for missing effects. The total number of articles using each base-case and robustness method is reported in the last column at the right of the table and in the last row at the bottom of the table. Legend: unspecified methods (Unclear), other methods (Others), Linear Extrapolation (Lin Ext), Last Value Carried Forward (LVCF), Mean Imputation (Mean), Conditional Imputation (Cond), Complete Case Analysis (CCA), Multiple Imputation (MI).

Costs		Robustness method									
		None	Others	Lin Ext	LVCF	Mean	Cond	CCA	MI	More than one	
Base-case method	Unclear	22	0	0	0	0	0	0	0	0	22
	Others	7	0	0	0	0	0	0	0	2	9
	Lin Ext	1	0	0	0	0	0	0	0	0	1
	LVCF	4	0	0	0	0	0	0	0	0	4
	Mean	1	0	0	0	0	0	1	0	0	2
	Cond	2	0	0	0	0	0	0	0	0	2
	CCA	10	2	0	1	0	0	0	1	0	14
	MI	16	0	0	1	0	0	9	0	1	27
		63	2	0	2	0	0	10	1	3	

Table C.7: Comparison of methods used in the base-case analysis (rows) and those used as alternatives in a robustness analysis (columns) for the articles between 2009-2015 for missing costs. The total number of articles using each base-case and robustness method is reported in the last column at the right of the table and in the last row at the bottom of the table. Legend: unspecified methods (Unclear), other methods (Others), Linear Extrapolation (Lin Ext), Last Value Carried Forward (LVCF), Mean Imputation (Mean), Conditional Imputation (Cond), Complete Case Analysis (CCA), Multiple Imputation (MI).

Effects		Robustness method									
		None	Others	Lin Ext	LVCF	Mean	Cond	CCA	MI	More than one	
Base-case method	Unclear	8	0	0	0	0	0	0	0	0	8
	Others	10	0	0	0	0	0	0	0	1	11
	Lin Ext	1	0	0	0	0	0	0	0	0	1
	LVCF	5	0	0	0	1	0	0	0	0	6
	Mean	6	1	0	0	0	0	0	0	0	7
	Cond	3	0	0	0	0	0	0	0	0	3
	CCA	8	5	0	2	0	0	0	1	1	17
	MI	16	1	0	1	0	0	7	1	2	28
		57	7	0	3	1	0	7	2	4	

Table C.8: Comparison of methods used in the base-case analysis (rows) and those used as alternatives in a robustness analysis (columns) for the articles between 2009-2015 for missing effects. The total number of articles using each base-case and robustness method is reported in the last column at the right of the table and in the last row at the bottom of the table. Legend: unspecified methods (Unclear), other methods (Others), Linear Extrapolation (Lin Ext), Last Value Carried Forward (LVCF), Mean Imputation (Mean), Conditional Imputation (Cond), Complete Case Analysis (CCA), Multiple Imputation (MI).

Table C.9 shows the number of the articles in the literature review in the period (2009-2015) associated with each combination of quality score derived from the Quality Evaluation Scheme (see Section 2.1) and strength of missing data assumptions (classified in terms of the type of missingness method used – see Section 2.3). Results are separately reported for missing costs and effects.

Costs		Strength of the assumptions					
		UNK	SI	CCA	MI	SA	
Quality score	A	0	1	0	4	0	5
	B	0	3	0	3	0	6
	C	0	2	2	10	0	14
	D	7	6	2	2	0	17
	E	14	8	8	9	0	39
		21	19	12	24	0	

Effects		Strength of the assumptions					
		UNK	SI	CCA	MI	SA	
Quality score	A	1	0	0	3	1	5
	B	0	2	0	3	0	5
	C	0	3	7	8	0	18
	D	2	8	1	5	0	16
	E	5	15	8	9	0	37
		8	28	16	28	1	

Table C.9: Comparison of the quality scores (rows) and strength of missing data assumptions (columns) associated with the studies between 2009-2015 for the missing costs and effects.The total number of articles for each category of the quality scores and strength of assumptions is reported in the last column at the right of the table and in the last row at the bottom of the table. The strength of assumptions for each study are broadly summarised according to the type of missing data method used: UNK (Unknown), SI (Single Imputation), CCA (Complete Case Analysis), MI (Multiple Imputation), SA (Sensitivity Analysis).

C.1.3 Statistical methods used in the reviewed studies

Table C.10 shows the numbers and proportions of articles in the literature review associated with different types of statistical methods used to perform the cost-effectiveness analysis.

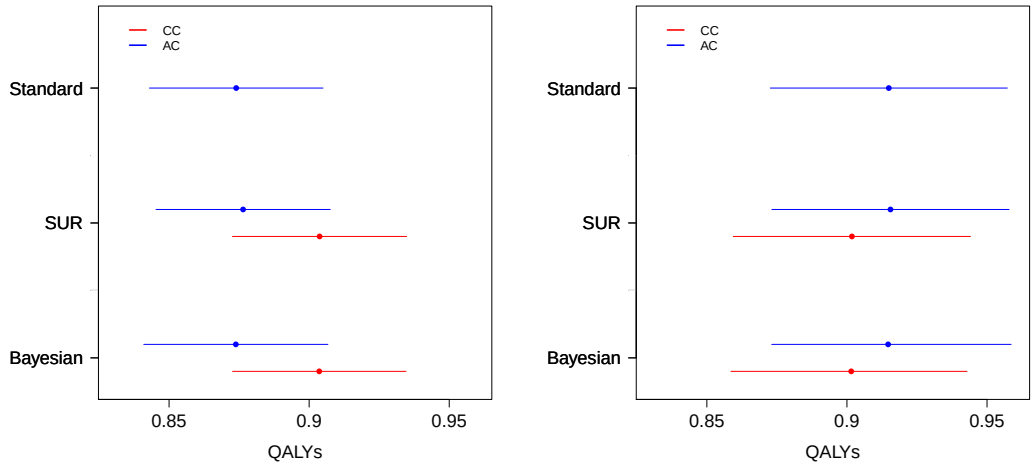
Method	n of studies	% of studies
Independent regressions with bootstrapping	30	38%
Independent regressions without bootstrapping	23	29%
Markov models	3	3%
Generalised estimating equations	2	2%
Seemingly unrelated regressions	2	2%
Generalised linear models	2	2%
MCMC methods	2	2%
Two-stage bootstrap methods	2	2%
Unclear	15	20%
Total	81	100%

Table C.10: Number (and frequency) of the studies by type of statistical methods used in the primary CEAs across the articles included in the literature review in Chapter 2 between 2009-2015.

C.2 Supplementary Analyses: Chapter 4

C.2.1 MenSS study

Figure C.2 compares the point and 95% CI estimates for the mean QALYs and costs parameters in the two treatment groups of the MenSS trial, which are obtained under three modelling approaches fitted to the completers using either the CC (red) or AC (blue) mean baseline utility adjustment. The three approaches are: ordinary least square regressions independently for QALYs and costs (Standard), Seemingly Unrelated Regressions (SUR) and Full Bayesian Model (Bayesian). In particular, SURs have been implememted using multi-equation estimation methods in the function `systemfit`, available from the `R` package `systemfit` (Henningsen and Hamann, 2018)



(a) control **(b)** Intervention

Figure C.2: Mean QALYs and costs estimates in the control (panel a) and intervention (panel b) group of the MenSS trial derived from three models fitted to the completers using either the CC (red) or AC (blue) mean baseline utility adjustment. The models are: Standard Approach (Standard), which uses only the AC mean baseline adjustment, Seemingly Unrelated Regression (SUR) and Full Bayesian Model (Bayesian).

Table C.11 compares the point and 95% CI estimates for the mean QALYs and costs and other incremental parameters (e.g. net benefit), for the two treatment groups of the MenSS trial, which are obtained under three modelling approaches: ordinary least square regressions independently for QALYs and costs (Standard), Seemingly Unrelated Regressions using maximum likelihood (SUR) and Full Bayesian Model (Bayesian).

C.2.2 PBS study

Table C.12 compares the point and 95% CI estimates for the mean QALYs and costs and other incremental parameters (e.g. net benefit), for the two treatment groups of the PBS trial, which are obtained under three modelling approaches: ordinary least square regressions independently for QALYs and costs (Standard), Multivariate Linear Mixed effects Models using maximum likelihood (MLME) and Full Bayesian Model (Bayesian). In particular, MLMEs have been implememted using mixed-effects maximum likelihood estimation methods in the function `lmer`, available from the `R` package `lme4` (Bates et al., 2019).

Parameter	Standard		SUR CC		SUR AC		Bayesian CC		Bayesian AC	
	Mean	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$										
mean QALY (μ_{e1})	0.874	(0.843;0.905)	0.904	(0.873;0.935)	0.876	(0.845;0.908)	0.904	(0.873;0.935)	0.874	(0.841;0.907)
mean cost (μ_{c1})	208	(106;310)	208	(106;310)	208	(106;310)	207	(104;307)	207	(104;307)
Intervention $(t = 2)$										
mean QALY (μ_{e2})	0.915	(0.873;0.957)	0.902	(0.859;0.944)	0.916	(0.873;0.958)	0.902	(0.859;0.943)	0.915	(0.873;0.959)
mean cost (μ_{c2})	189	(112;266)	189	(131;285)	189	(131;285)	189	(111;266)	189	(111;266)
Incremental										
QALY differential (Δ_e)	0.04	(-0.02;0.09)	-0.002	(-0.05;0.05)	0.04	(-0.02;0.1)	-0.002	(-0.054;0.05)	0.041	(-0.013;0.094)
Cost differential (Δ_c)	-18	(-145;97)	-18	(-160;107)	-18	(-160;107)	-18	(-146;110)	-18	(-146;110)
IB (at $k = 20000$)			-20	(-1385;1016)	801	(-484;2064)	-23	(-1063;1042)	835	(-241;1928)
ICER			9625		-482		8822		-449	

Table C.11: Means and 95% credible/confidence interval estimates of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the MenSS trial obtained from different models. The models are: the original model used by Bailey et al. (2016) (Standard), Seemingly Unrelated Regression (SUR) and Full Bayesian model (Bayesian). The results of the last two models are distinguished according to whether they were fitted using the CC or AC mean baseline utilities. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Cost values are expressed in £.

Parameter	Standard		MLME CC		MLME AC		Bayesian CC		Bayesian AC	
	Mean	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$										
mean QALY (μ_{e1})	0.529	(0.493;0.565)	0.493	(0.449;0.536)	0.523	(0.479;0.566)	0.493	(0.441;0.546)	0.521	(0.469;0.577)
mean cost (μ_{c1})	2974	(2104;3843)	2984	(2050;4068)	2984	(2054;4070)	3054	(2122;3997)	3055	(2124;4000)
Intervention $(t = 2)$										
mean QALY (μ_{e2})	0.611	(0.575;0.647)	0.593	(0.531;0.676)	0.592	(0.531;0.676)	0.585	(0.531;0.637)	0.583	(0.529;0.635)
mean cost (μ_{c2})	4568	(3915;5221)	5346	(4540;6287)	5349	(4543;6291)	5449	(4673;6217)	5456	(4693;6221)
Incremental										
QALY differential (Δ_e)	0.08	(0.012;0.14)	0.08	(-0.02;0.18)	0.05	(-0.05;0.15)	0.09	(0.03;0.15)	0.06	(0.01;0.12)
Cost differential (Δ_c)	1608	(212;2434)	2362	(1125;3590)	2365	(1223;3562)	2395	(1172;3592)	2401	(1155;3567)
IB (at $k = 20000$)			-353	(-3240;1493)	-996	(-3841;893)	-552	(-2220;1108)	-1166	(-2876;452)
ICER			23971		32610		25985		38871	

Table C.12: Means and 95% credible/confidence interval estimates of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the PBS trial obtained from different models. The models are: the original model used by Hassiotis et al. (2018) (Standard), Multivariate Linear Mixed effects Model (MLME) and Full Bayesian model (Bayesian). The results of the last two models are distinguished according to whether they were fitted using the CC or AC mean baseline utilities. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Cost values are expressed in £.

Figure C.3 compares the point and 95% CI estimates for the mean QALYs and costs parameters in the two treatment groups of the PBS trial, which are obtained under three modelling approaches fitted to the completers using either the CC (red) or AC (blue) mean baseline utility adjustment. The three approaches are: ordinary least square regressions independently for QALYs and costs (Standard), Multivariate Linear Mixed effects Model using maximum likelihood (MLME) and Full Bayesian Model (Bayesian).

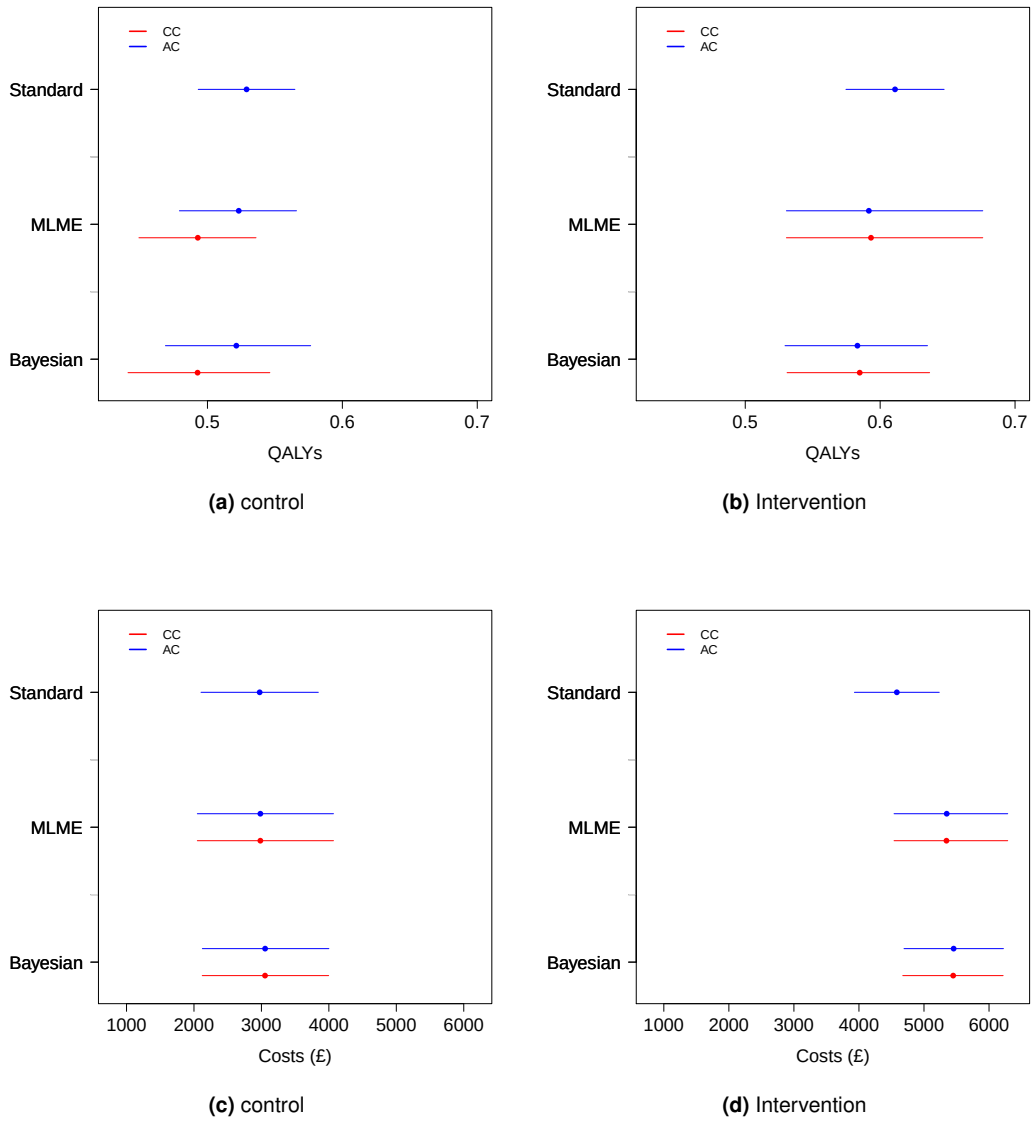
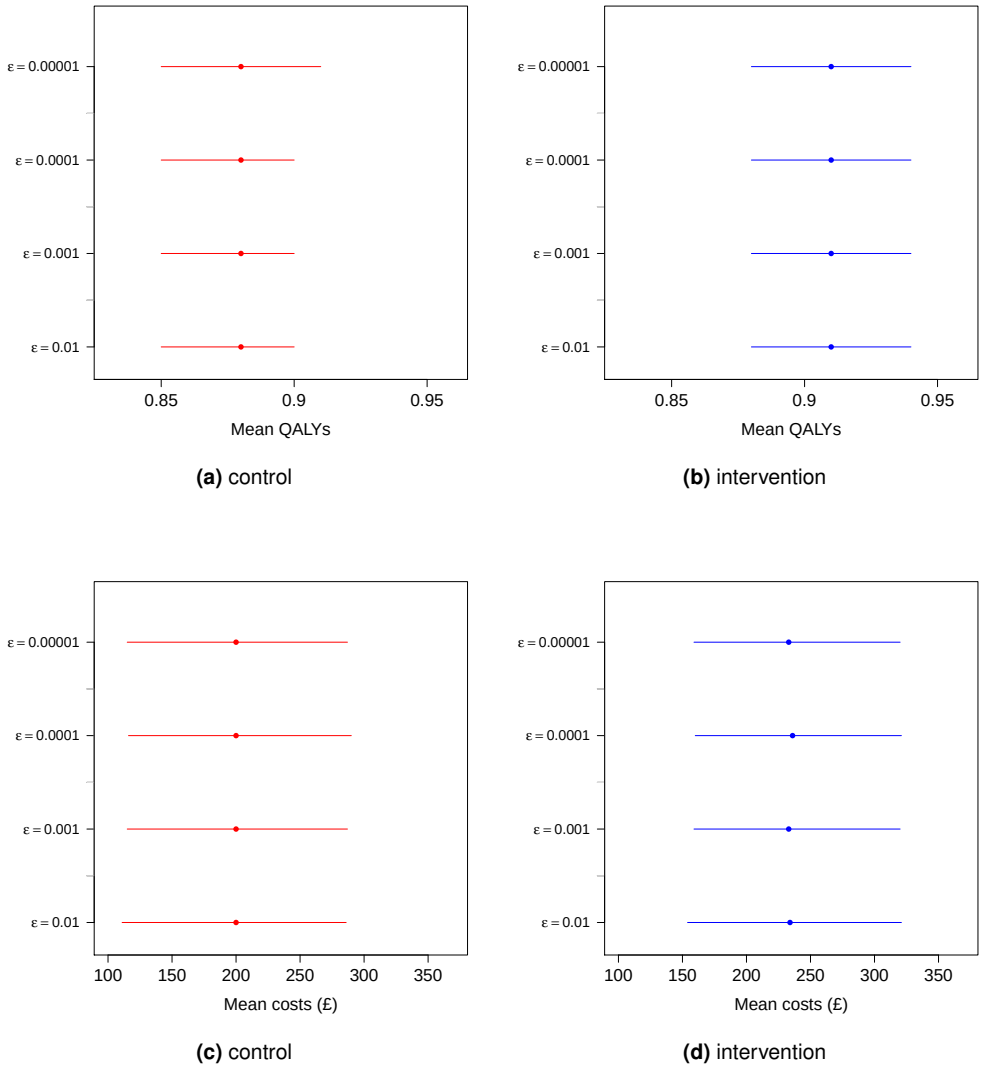


Figure C.3: Mean QALYs and cost estimates in the control (panels a and c) and intervention (panels b and d) group of the PBS trial derived from three models fitted to the completers using either the CC (red) or AC (blue) mean baseline utility adjustment. The models are: Standard Approach (Standard), which uses only the AC mean baseline adjustment, Multivariate Linear Mixed effects Model (MLME) and Full Bayesian Model (Bayesian).

C.3 Supplementary Analyses: Chapter 5

C.3.1 Sensitivity to the choice of the scaling parameter for the costs

Figure C.4 shows the sensitivity of the inferences across the alternative specifications for the scaling factor ϵ, which is added to the data of the MenSS trial to avoid zero values when using a Gamma distributions for modelling the costs. Results in terms of mean posterior estimates and 90% HPD intervals were almost unchanged in all the cases. Thus, we can assert that model performance was unaffected by the choice of the value for ϵ.



(a) control **(b)** intervention

(c) control **(d)** intervention

Figure C.4: Sensitivity analysis for the choice of the scaling parameter ϵ when fitting the Beta-Gamma model to the QALYs and cost data under the "all cases" scenario for the MenSS trial. The different ϵ values are subtracted and added to the set of observed QALYs and cost data to avoid the boundary values that fall outside the support of the Beta and Gamma distributions (1 for the QALYs and 0 for the costs), respectively. For each value of ϵ tested, posterior means and 90% HPD intervals for the mean QALYs and cost parameters are respectively represented with dots and lines: red for the control (panels a and c) and blue for the intervention (panels b and d).

C.3.2 Implementation "trick" for the Hurdle Model

The Hurdle Model described in Chapter 5 can be represented using a different sampling distribution for the QALYs, depending on the observed value of the indicator d_{ie}

$$e_i \mid d_{ie} \sim \begin{cases} p(e_i \mid d_{ie}=0) = p(e_i \mid \boldsymbol{\theta}^{<1}), & \text{if } e_i < 1 \\ p(e_i \mid d_{ie}=1) = p(e_i \mid \boldsymbol{\theta}^{1}), & \text{if } e_i = 1, \end{cases} \tag{C.1}$$

where the model for $e_i = 1$ is degenerate at a point mass at 1, while that for $e_i < 1$ is defined in terms of a Beta distribution. We can conveniently re-write Equation C.1 more succinctly and with specific reference to our case as

$$e_i \sim \text{Beta}\left(\phi_{ie}^{d_{ie}}\tau_{ie}^{d_{ie}}, \left(1-\phi_{ie}^{d_{ie}}\right)\tau_{ie}^{d_{ie}}\right). \tag{C.2}$$

If we set $\phi^1_{ie} = 1$ and select τ^1_{ie} in order to induce a variance as close to 0 as possible, the two specifications are identical. Unfortunately, it is not possible to directly fit the model in Equation C.2 into BUGS/JAGS, because the Beta distribution is specified in the open interval $(0, 1)$ and thus setting $\phi^1_{ie} = 1$ implies that $\tau^1_{ie} = 0$, which is not allowed.

However, the required behaviour is very closely mimicked if we define our model with logit$(\phi^1_{ie}) = \alpha^1_0 [+ \ldots]$ and set $\alpha^1_0 =$ logit(0.999999) and $\sigma_e \approx 0$, which implies $\mu_e \approx 1$ with virtually no uncertainty. In other words, we can specify extremely informative priors on the parameters $\boldsymbol{\theta}^1$ so that the implied distribution for the structural ones components of the mixture is concentrated around 1 with essentially no uncertainty. More importantly, with such a prior no amount of data can modify the posterior. The critical aspect of this strategy, however, is that inferences may be potentially sensitive to the way such priors are specified, that is whether a small variation in the hyperprior values can affect the posterior estimates.

In fact, the estimation of the other parameters is not really affected by this choice, provided that the encoded prior really induces the variance towards zero. It is also plausible that different values for σ^1_e have an impact on measures of model fit, such as the DIC. This is essentially due to the fact that the population is really comprised of two groups, one of which shows QALYs that are identically one. Thus, the closer the approximation to zero for the variance the better the fit to the observed data and therefore the smaller the resulting DIC. With this in mind, we have used different values for σ^1_e to assess the impact on the mean QALYs estimates. Fixing the value of the mean for the ones group to $\mu^1_e = 0.999999$ corresponds to an upper bound for the standard deviation of 0.0001 (see Section 5.2.2). We have explored a range of possibilities by progressively decreasing this value and assessed their impact on posterior results.

Figure C.5 shows the sensitivity of the inferences across the alternative specifications for σ^1_e. Results in terms of mean posterior estimates and 90% HPD intervals were almost unchanged in all the cases. Thus, we can assert that model performance was unaffected by the choice of the value for σ^1_e. We also observe that the DIC becomes smaller when the standard deviation parameter decreases and the best-fitting model is the one associated with the smallest values, although the results are hardly different from both an estimation and convergence perspective for all the parameters.

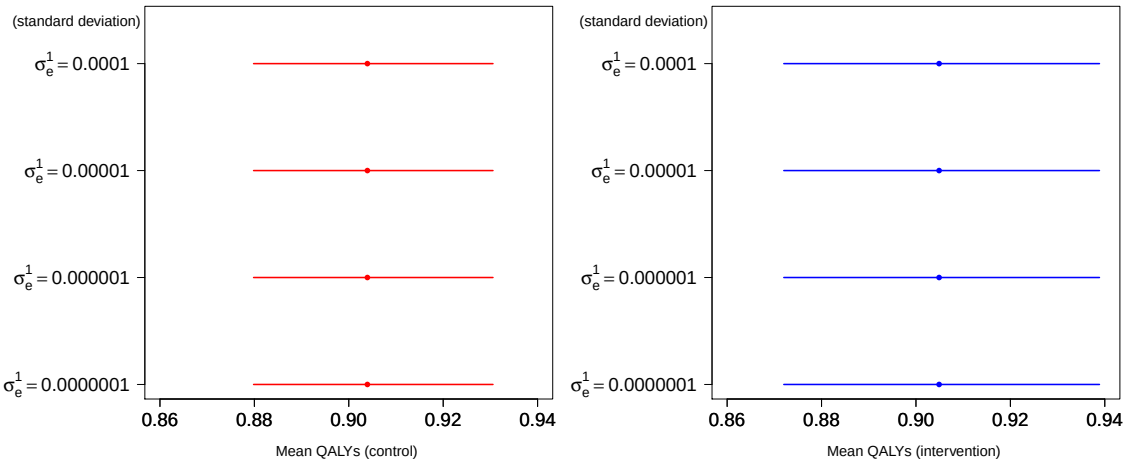


Figure C.5: Sensitivity analysis for the choice of the standard deviation for the distribution of the structural ones in the QALYs. For each value of σ^1_e tested, posterior means and 90% HPD intervals for the mean QALYs parameters are respectively represented with dots and lines (red for the control and blue for the intervention group).

C.3.3 Prior sensitivity

Figure C.6 and Figure C.7 compare the point and 95% CI estimates for the expected effectiveness and cost differentials derived from the models fitted to the MenSS and PBS data in Chapter 5 with those obtained under three alternative prior specifications. The original prior specifications (Original) assume Normal priors for the regression parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} \sim$ Normal$(0, 1000)$ and uniform priors for the standard deviations $\boldsymbol{\sigma} \sim$ Uniform$(0, 1000)$ of the costs and the QALYs (for the latter, whenever a Beta distribution is assumed, truncated uniform priors as described in Section 5.2.2 are used). The three alternative versions considered vary the priors on the standard deviations, using either Uniform$(0, 10000)$ (Uniform) or truncated Normal$(0, 1000)$T$[0,]$ (Half-Normal) specifications, and on the regression parameters, using Normal$(0, 100000)$ (Normal) priors.

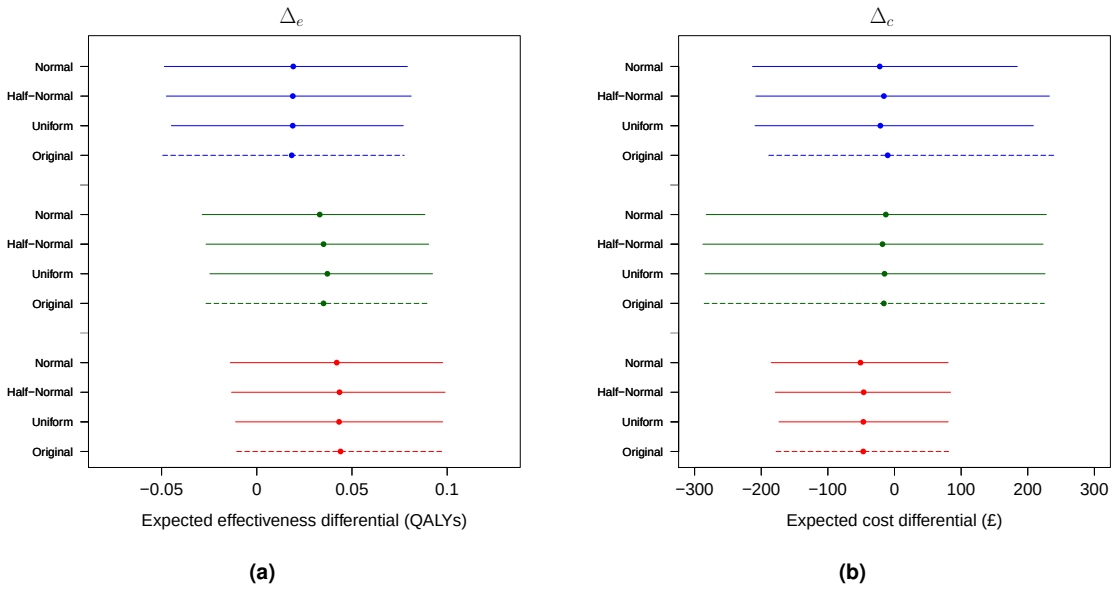


Figure C.6: Mean and 95% credible interval estimates of the expected effectiveness (panel a) and cost (panel b) differentials for the Bivariate Normal, Beta-Gamma and Hurdle Model (respectively indicated by red, green and blue colours in the graphs) in the MenSS trial under different priors. The prior specification described in Section 5.2 (Original), denoted with dashed lines in the graphs, assumes Normal priors for the regression parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} \sim$ Normal$(0, 1000)$ and uniform priors for the standard deviations $\boldsymbol{\sigma} \sim$ Uniform$(0, 1000)$ of the costs and the QALYs (for the latter, the Beta-Gamma and Hurdle Model assume a truncated uniform prior as described in Section 5.2.2). Three alternative versions are considered which vary the priors on the standard deviations, using either Uniform$(0, 10000)$ (Uniform) or truncated Normal$(0, 1000)$T$[0,]$ (Half-Normal) specifications, and on the regression parameters, using Normal$(0, 100000)$ (Normal) priors.

C.3.4 Posterior estimates

Table C.13 compares the point and 95% CI estimates for the mean QALYs and costs and other incremental parameters (e.g. net benefit), in the two treatment groups of the MenSS trial between three alternative specifications of the modelling framework described in Section 5.1: Normal for both QALYs and costs (Bivariate Normal), Beta for the QALYs and Gamma for the costs (Beta-Gamma) and Hurdle for the QALYs and Gamma for the costs (Hurdle Model).

Table C.14 compares the point and 95% CI estimates for the mean QALYs and costs and other incremental parameters (e.g. net benefit), in the two treatment groups of the PBS trial between three alternative specifications of the modelling framework described in Section 5.1: Normal for both QALYs and costs (Bivariate Normal), Beta for the QALYs and LogNormal for the costs (Beta-Gamma) and Hurdle for the QALYs and LogNormal for the costs (Hurdle Model).

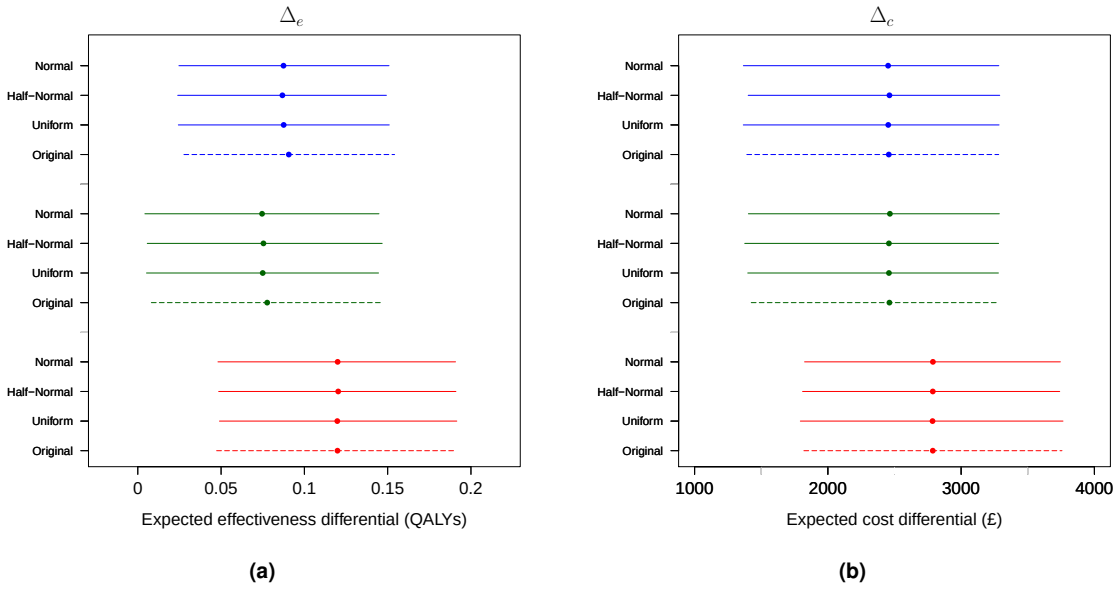


Figure C.7: Mean and 95% credible interval estimates of the expected effectiveness (panel a) and cost (panel b) differentials for the Bivariate Normal, Beta-LogNormal and Hurdle Model (respectively indicated by red, green and blue colours in the graphs) in the PBS trial under different priors. The prior specification described in Section 5.2 (Original), denoted with dashed lines in the graphs, assumes Normal priors for the regression parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} \sim \text{Normal}(0, 1000)$ and uniform priors for the standard deviations $\boldsymbol{\sigma} \sim \text{Uniform}(0, 1000)$ of the costs and the QALYs (for the latter, the Beta-LogNormal and Hurdle Model assume a truncated uniform prior as described in Section 5.2.2). Three alternative versions are considered which vary the priors on the standard deviations, using either $\text{Uniform}(0, 10000)$ (Uniform) or truncated $\text{Normal}(0, 1000)\text{T}[0,]$ (Half-Normal) specifications, and on the regression parameters, using $\text{Normal}(0, 100000)$ (Normal) priors.

Parameter	Bivariate Normal		Beta-Gamma		Hurdle Model	
	Mean	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$						
mean QALY (μ_{e1})	0.873	(0.84;0.906)	0.876	(0.836;0.907)	0.88	(0.839;0.912)
mean cost (μ_{c1})	234	(135;336)	200	(116;336)	198	(116;343)
Intervention $(t = 2)$						
mean QALY (μ_{e2})	0.92	(0.874;0.961)	0.907	(0.859;0.94)	0.895	(0.836;0.938)
mean cost (μ_{c2})	187	(106;267)	189	(89;401)	193	(91;419)
Incremental						
QALY differential (Δ_e)	0.04	(-0.01;0.1)	0.03	(-0.03;0.08)	0.02	(-0.05;0.07)
Cost differential (Δ_c)	47	(-178;81)	-11	(-184;217)	-5	(-184;244)
IB (at $k = 20000$)	927	(-188;2055)	631	(-559;1730)	312	(-1072;1545)
ICER	1067		-355		-339	

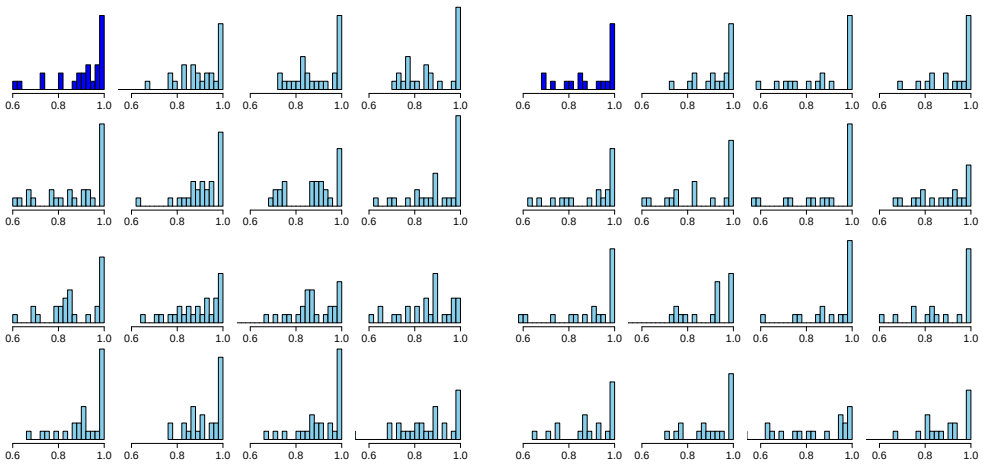
Table C.13: Means and 95% credible interval estimates of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the MenSS trial obtained from the Bivariate Normal, Beta-Gamma and Hurdle model under the "all cases" scenario. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Cost values are expressed in £.

C.3.5 Posterior Predictive Checks

Figure C.8 compares the histograms of the empirical distributions of the QALYs in both treatment groups of the MenSS study (dark blue bars) with respect to 15 replications of the dataset based on the samples drawn from the posterior predictive distribution of the Hurdle Model (light blue bars) in Chapter 5 Most of the replicated QALYs seem to well-capture the empirical distributions of the observed values, including the spikes at 1 for the data in both the control (panel a) and intervention (panel b) groups.

Parameter	Bivariate Normal		Beta LogNormal		Hurdle Model	
	Mean	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$						
mean QALY (μ_{e1})	0.491	(0.442;0.553)	0.501	(0.452;0.551)	0.491	(0.442;0.544)
mean cost (μ_{c1})	2927	(2165;3692)	2605	(2067;3414)	2639	(2046;3546)
Intervention $(t = 2)$						
mean QALY (μ_{e2})	0.613	(0.56;0.662)	0.571	(0.523;0.623)	0.58	(0.531;0.632)
mean cost (μ_{c2})	5573	(4953;6196)	5111	(4586;5731)	5334	(4665;6090)
Incremental						
QALY differential (Δ_e)	0.12	(0.05;0.19)	0.07	(0.01;0.14)	0.09	(0.02;0.15)
Cost differential (Δ_c)	2643	(1648;3626)	2484	(1672;3205)	2670	(1688;3543)
IB (at $k = 20000$)	-246	(-2014;1505)	-991	(-2575;605)	-917	(-2645;835)
ICER	22053		33274		30467	

Table C.14: Means and 95% credible interval estimates of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the PBS trial obtained from the Bivariate Normal, Beta LogNormal and Hurdle model under the "all cases" scenario. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Cost values are expressed in £.

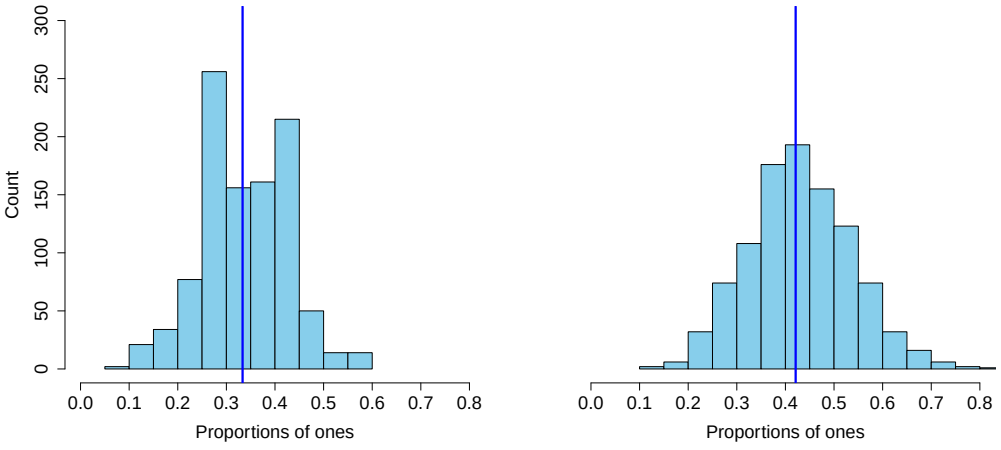


(a) control **(b)** intervention

Figure C.8: Histograms of the empirical distributions of the observed QALYs (dark blue bars) in the control (panel a) and intervention (panel b) in the MenSS trial, which are compared with the replications of the data based on the samples drawn from the posterior predictive distribution of the Hurdle Model. For visualisation purposes, only 15 replications (out of the 1000 generated) are displayed in the graphs.

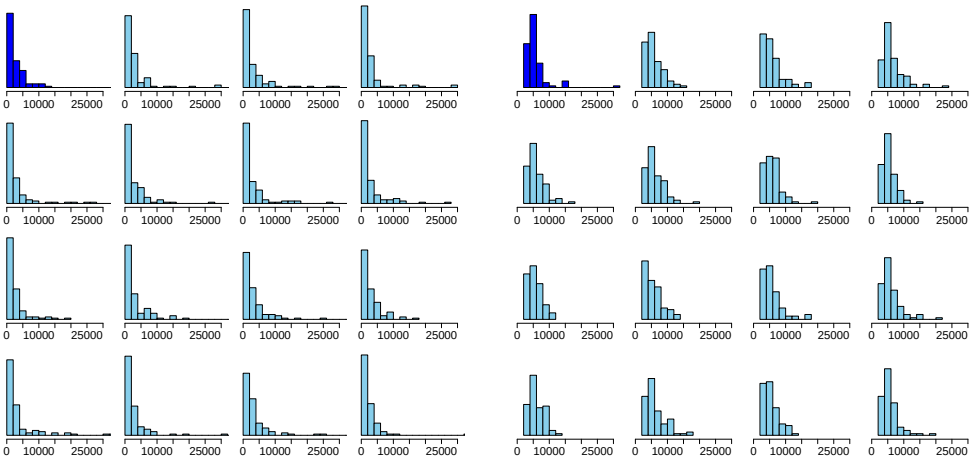
Figure C.9 shows the histograms of the proportions of structural ones derived from 1000 replicates of the QALYs in the MenSS study drawn from the posterior predictive distribution of the Hurdle Model, which are compared with the proportions of ones in the observed QALYs. The distribution of the replicated proportions of ones is on average close to the corresponding values observed in the data for both the control (panel a) and intervention (panel b) groups.

Figure C.10 compares the histograms of the empirical distributions of the costs in both treatment groups of the PBS study (dark blue bars) with respect to 15 replications of the dataset based on the samples drawn from the posterior predictive distribution of the Hurdle Model (light blue bars)



(a) control **(b)** intervention

Figure C.9: Histograms of the proportions of ones across 1000 replications of the QALYs computed from the samples of the posterior predictive distribution of the Hurdle Model (light blue bars) compared with the proportions observed (dark blue lines) in the control (panel a) and intervention (panel b) group in the MenSS trial.

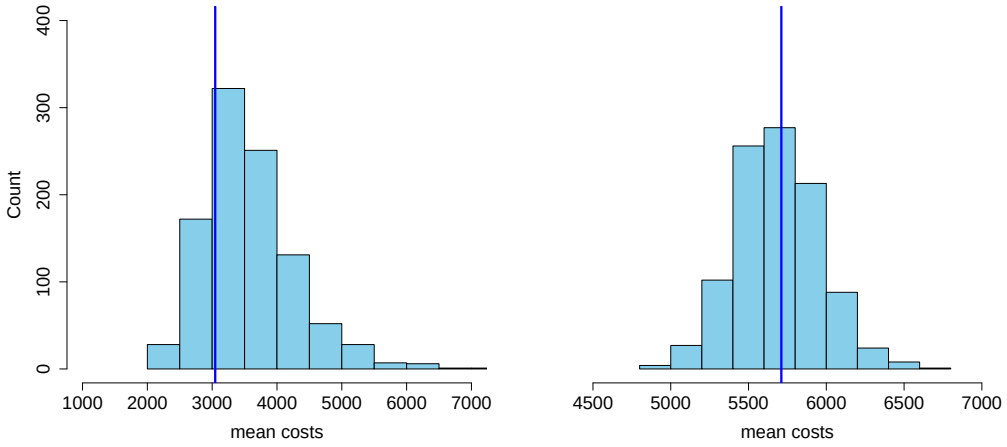


(a) control **(b)** intervention

Figure C.10: Histograms of the empirical distributions of the observed costs (dark blue bars) in the control (panel a) and intervention (panel b) in the PBS trial, which are compared with the replications of the data based on the samples drawn from the posterior predictive distribution of the Hurdle Model. For visualisation purposes, only 15 replications (out of the 1000 generated) are displayed in the graphs.

The replicated datasets seem to closely approximate the empirical cost distributions in both the control (panel a) and intervention (panel b) groups.

Figure C.11 shows the histograms of the proportions of sample means derived from 1000 replicates of the costs in the PBS study drawn from the posterior predictive distribution of the Hurdle Model, which are compared with the sample means in the observed costs.



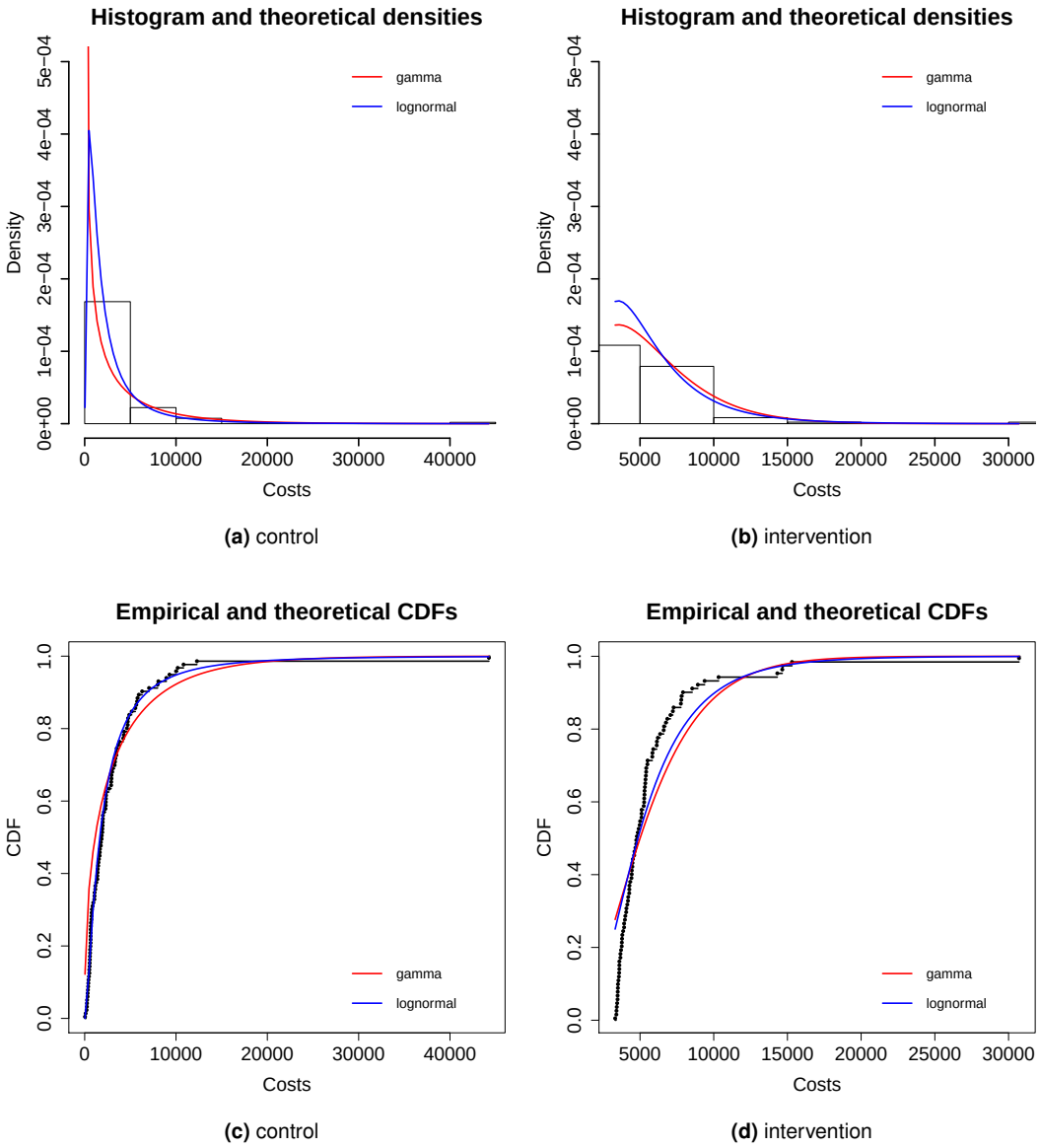
(a) control **(b)** intervention

Figure C.11: Histograms of sample means across 1000 replications of the costs computed from the samples of the posterior predictive distribution of the Hurdle Model (light blue bars) compared with the observed means (dark blue lines) in the control (panel a) and intervention (panel b) group in the PBS trial.

The distribution of the replicated cost means is centered around the sample mean values computed on the observed data for both the control (panel a) and intervention (panel b) groups.

C.3.6 Gamma vs LogNormal

Figure C.12 compares the density and cumulative density functions (black lines) estimated from the empirical costs in the two treatment groups of the PBS trial with respect to the theoretical values obtained from fitting Gamma (red lines) or LogNormal (blue lines) distributions to the data using maximum likelihood methods.



(a) control **(b)** intervention

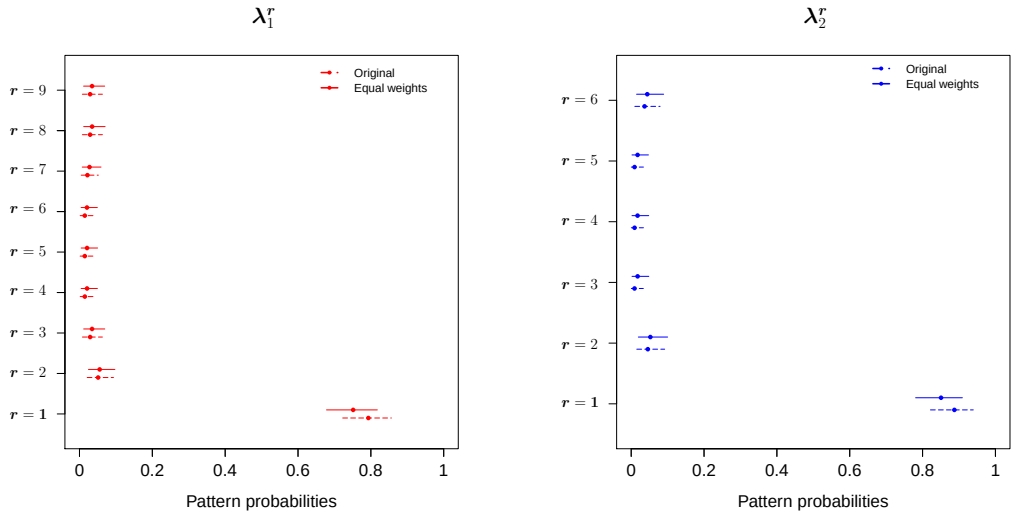
(c) control **(d)** intervention

Figure C.12: Density and cumulative density plots of the empirical cost data against the theoretical values obtained from fitting a Gamma or LogNormal distribution via maximum likelihood methods for both the control (panels a and c) and intervention (panels b and d) group in the PBS trial.

C.4 Supplementary Analyses: Chapter 6

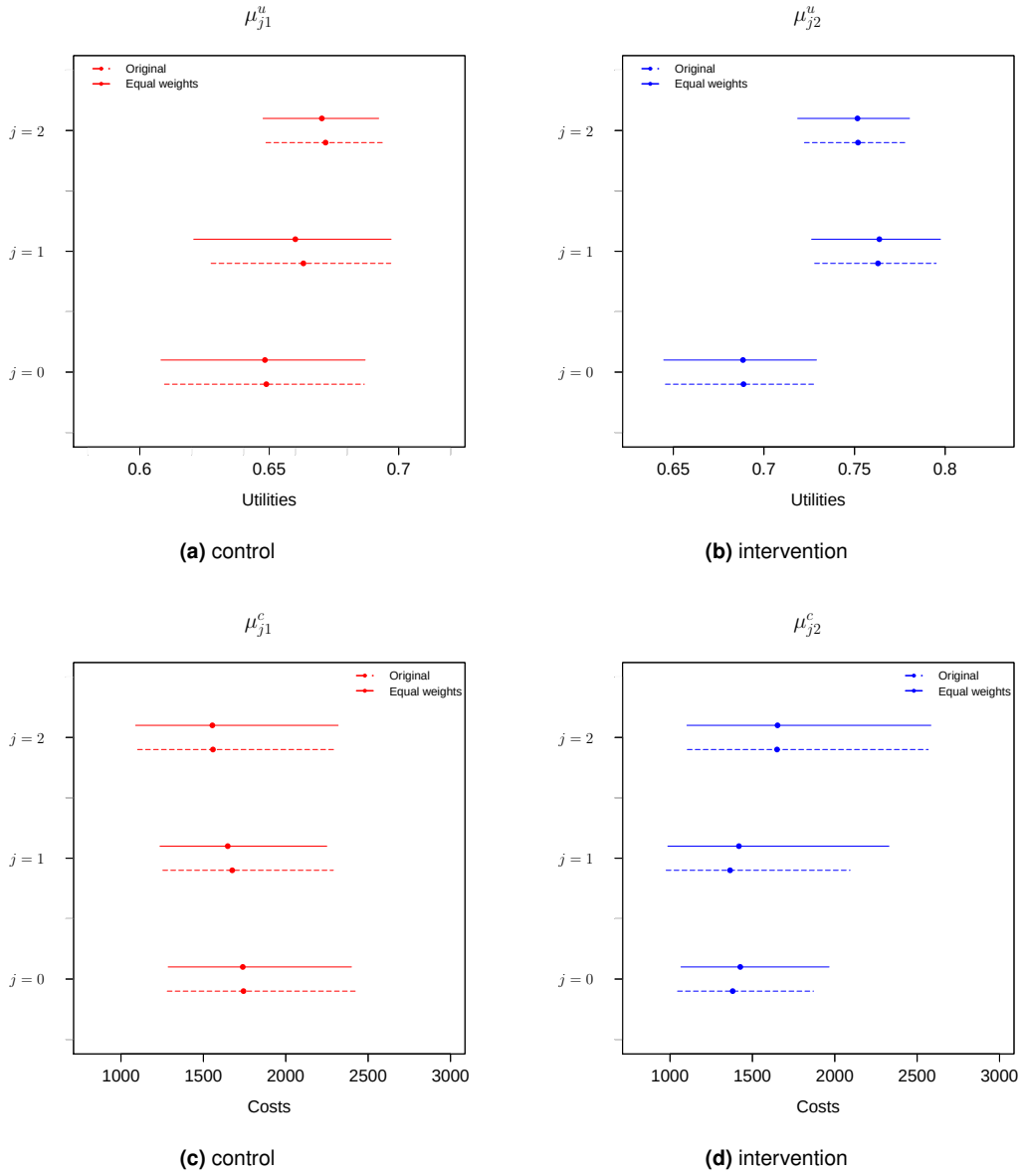
C.4.1 Prior sensitivity

Figure C.13 and Figure C.14 compare the point and 95% CI estimates for the missingness patterns probabilities ($\boldsymbol{\lambda}_t^r$) and marginal mean utilities and costs (μ_{jt}^u, μ_{jt}^c) derived from the models fitted to the two treatment groups of the PBS trial in Chapter 6 with those obtained under an alternative prior specification for $\boldsymbol{\lambda}_t^r$. The original Dirichlet prior (Original), gives more weight to the completers pattern and the same weights across the non-completers pattern (see Section 6.2.1), while the alternative Dirichlet$(1, \ldots, 1)$ prior (Equal weights) assigns the same weight to all the patterns.



(a) control **(b)** intervention

Figure C.13: Mean and 95% credible interval estimates of the missingness patterns' probabilities $\boldsymbol{r}$ in the control (red dots and lines, panel a) and intervention (blue dots and lines, panel b) group of the PBS study. The estimates are shown under either the Dirichlet prior for $\boldsymbol{\lambda}_t^r$ described in Section 6.2.1 (Original) or the alternative specification Dirichlet$(1, \ldots, 1)$, which assigns the same weight to each pattern (Equal weights).



(a) control **(b)** intervention

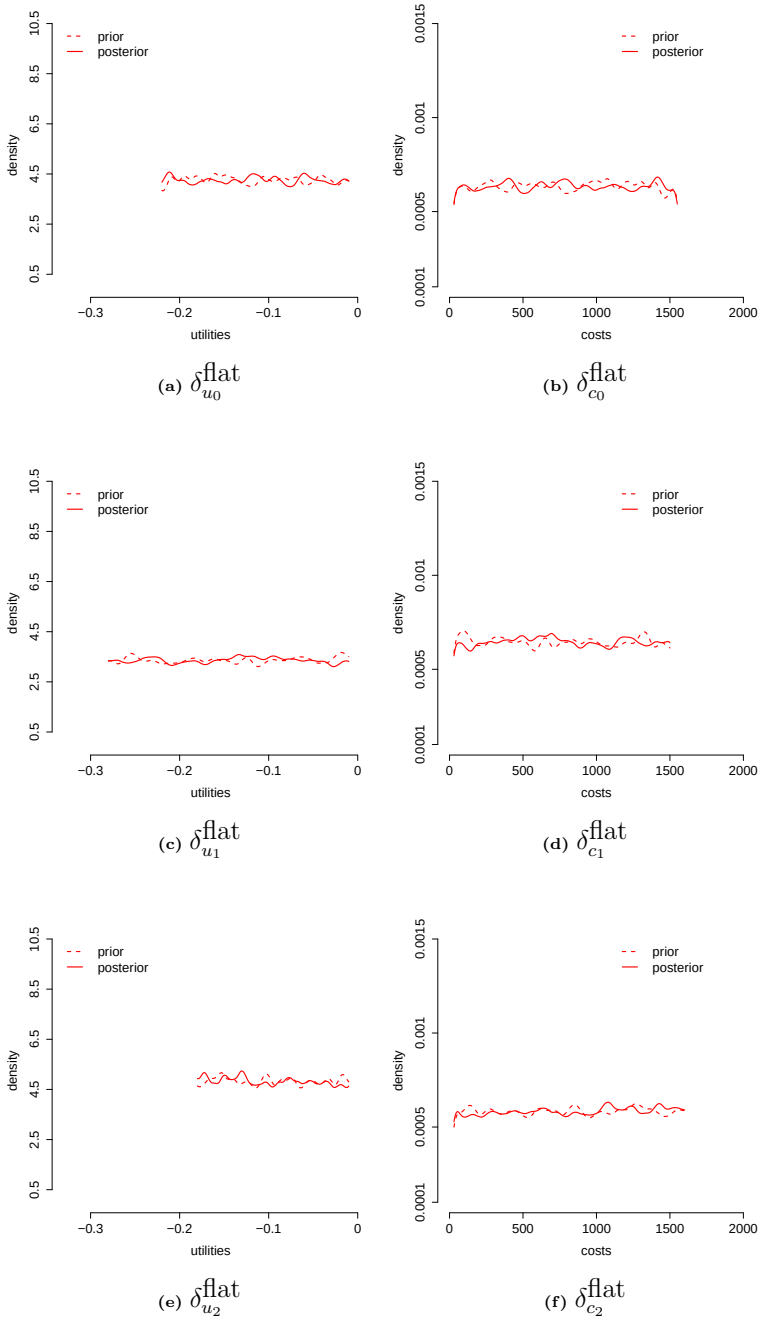
(c) control **(d)** intervention

Figure C.14: Mean and 95% credible interval estimates of the mean utilities μ^u_{jt} (panels a and b) and costs μ^c_{jt} (panels c and d) in the control (red dots and lines) and intervention (blue dots and lines) group of the PBS study. Estimates are obtained for the model $\boldsymbol{\delta} = \mathbf{0}$ and are shown under either the Dirichlet prior for $\boldsymbol{\lambda}^r_t$ described in Section 6.2.1 (Original) or the alternative specification Dirichlet$(1, \ldots, 1)$, which assumes the same weight for each pattern (Equal weights).

C.4.2 Priors and posteriors for the sensitivity parameters

Figure C.15, Figure C.16 and Figure C.17 compare the prior (dashed lines) and posterior (solid lines) densities of the sensitivity parameters δ_j used in the model in Chapter 6 for all utility and cost marginal mean parameters in the control group of the PBS trial, under the nonignorable scenarios $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$, respectively.

Figure C.18, Figure C.19 and Figure C.20 compare the prior (dashed lines) and posterior (solid lines) densities of the sensitivity parameters δ_j used in the model in Chapter 6 for all utility and cost marginal mean parameters in the intervention group of the PBS trial, under the nonignorable scenarios $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$, respectively.

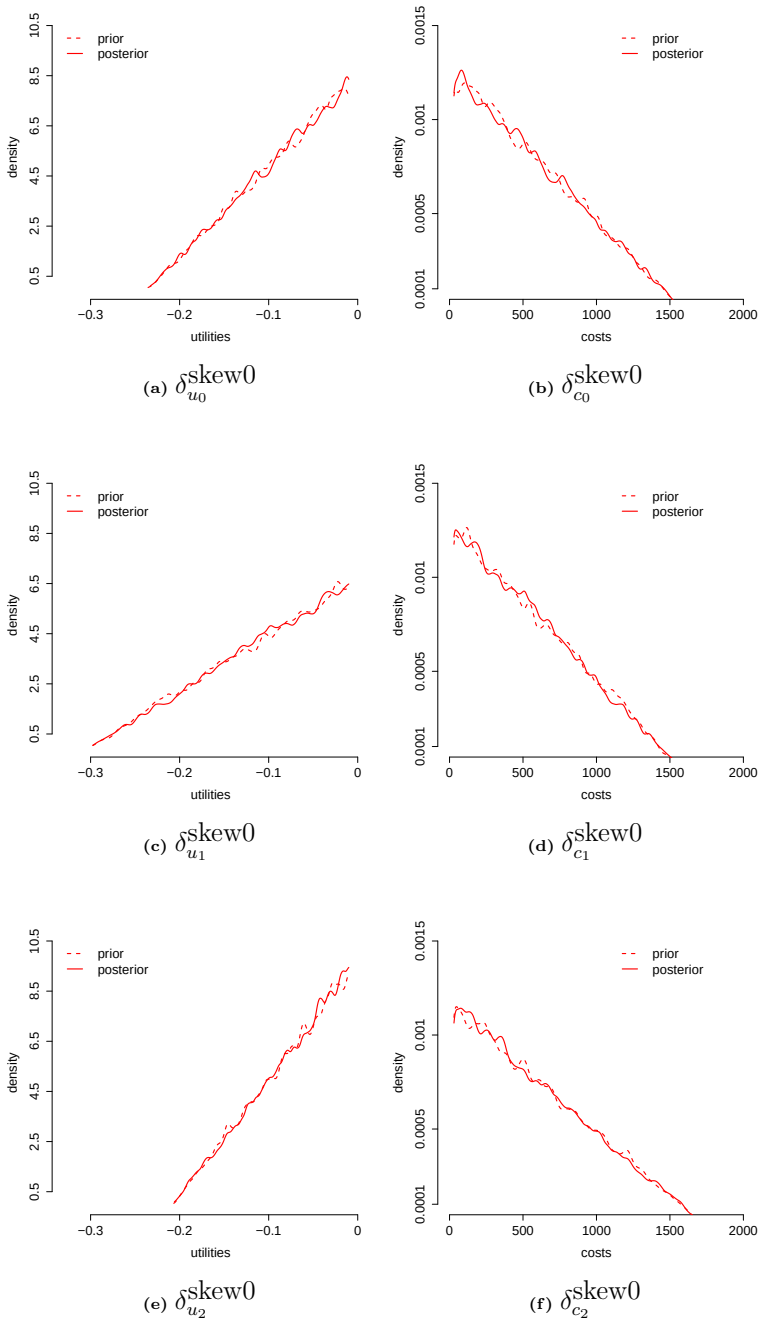


(a) $\delta^{\text{flat}}_{u_0}$ (b) $\delta^{\text{flat}}_{c_0}$

(c) $\delta^{\text{flat}}_{u_1}$ (d) $\delta^{\text{flat}}_{c_1}$

(e) $\delta^{\text{flat}}_{u_2}$ (f) $\delta^{\text{flat}}_{c_2}$

Figure C.15: Prior densities (dashed lines) and posterior densities (solid lines) of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{flat}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the control group.

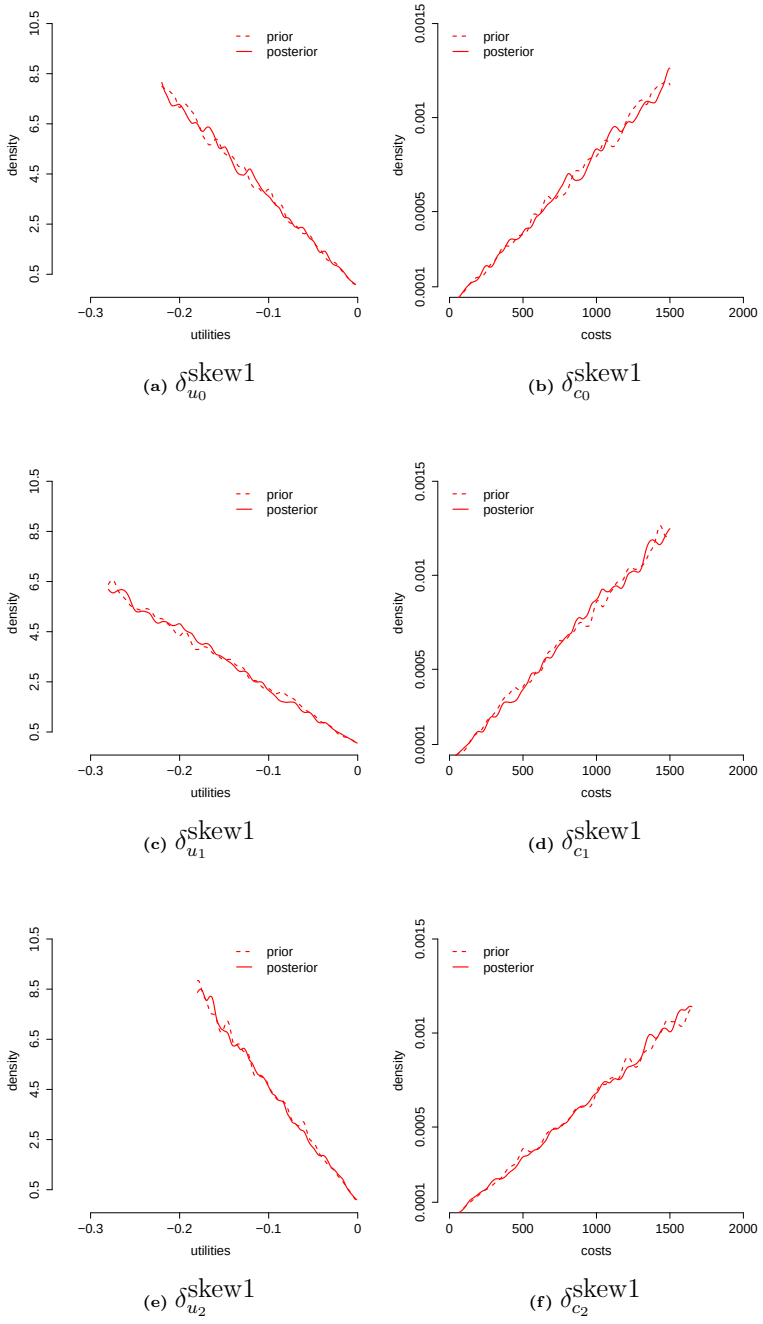


(a) $\delta_{u_0}^{\text{skew0}}$ (b) $\delta_{c_0}^{\text{skew0}}$

(c) $\delta_{u_1}^{\text{skew0}}$ (d) $\delta_{c_1}^{\text{skew0}}$

(e) $\delta_{u_2}^{\text{skew0}}$ (f) $\delta_{c_2}^{\text{skew0}}$

Figure C.16: Prior densities (dashed lines) and posterior densities (solid lines) of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{skew0}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the control group.

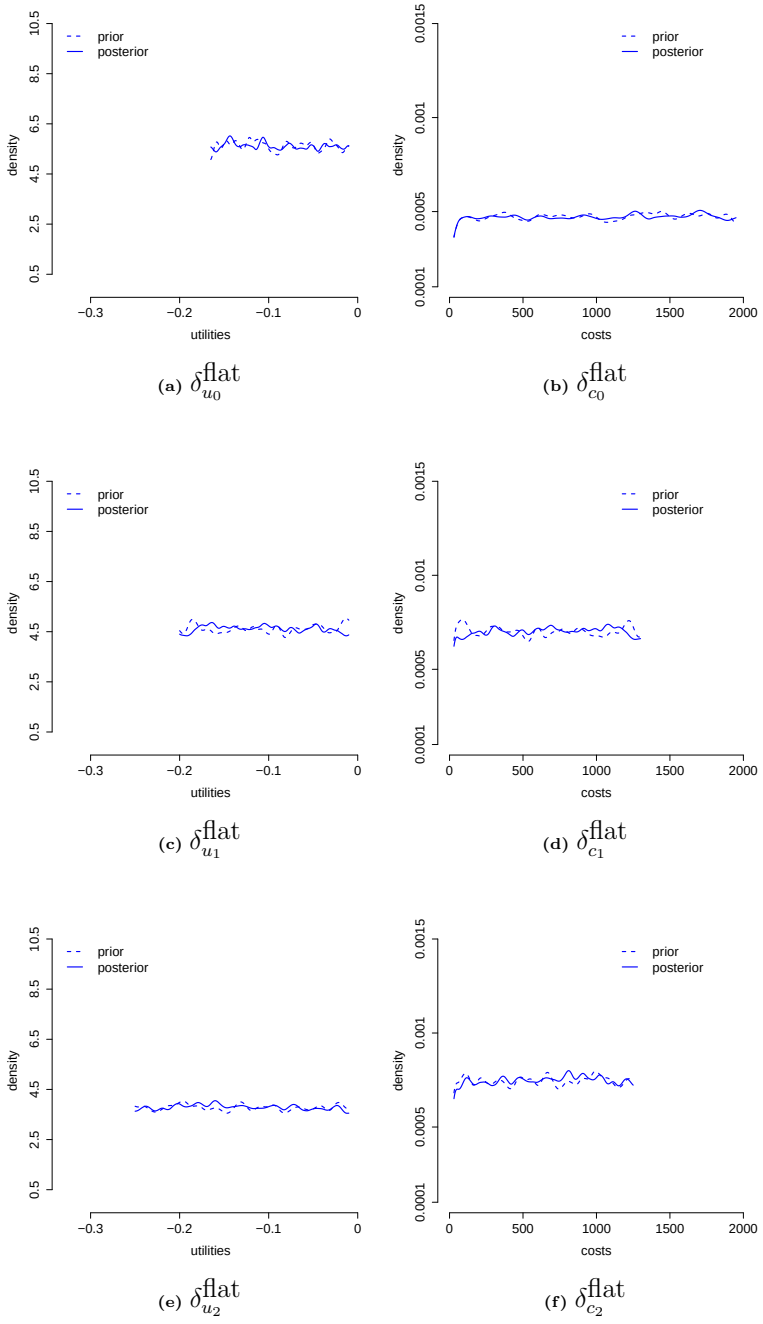


(a) $\delta^{\text{skew1}}_{u_0}$ (b) $\delta^{\text{skew1}}_{c_0}$

(c) $\delta^{\text{skew1}}_{u_1}$ (d) $\delta^{\text{skew1}}_{c_1}$

(e) $\delta^{\text{skew1}}_{u_2}$ (f) $\delta^{\text{skew1}}_{c_2}$

Figure C.17: Prior densities (dashed lines) and posterior densities (solid lines) of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{skew1}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the control group.

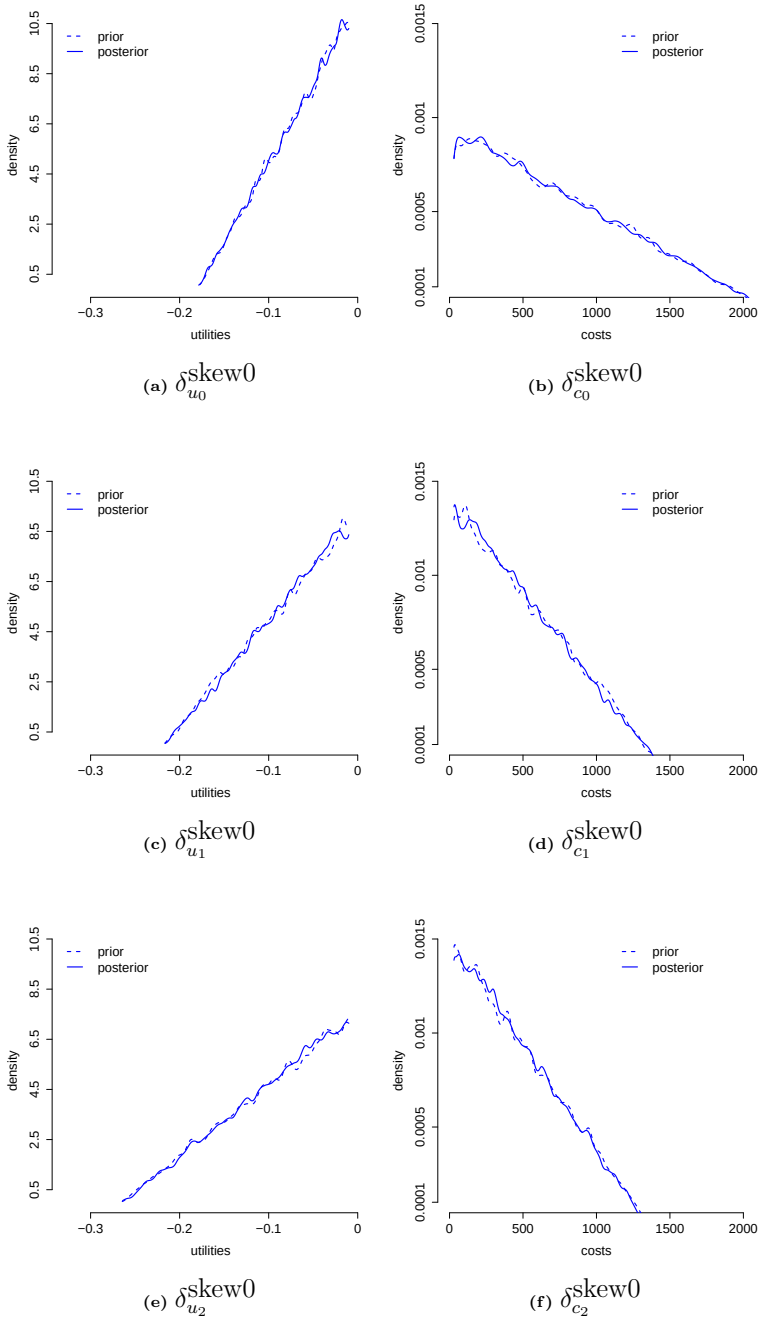


(a) $\delta^{\text{flat}}_{u_0}$ (b) $\delta^{\text{flat}}_{c_0}$

(c) $\delta^{\text{flat}}_{u_1}$ (d) $\delta^{\text{flat}}_{c_1}$

(e) $\delta^{\text{flat}}_{u_2}$ (f) $\delta^{\text{flat}}_{c_2}$

Figure C.18: Prior and posterior densities of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{flat}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the intervention group.

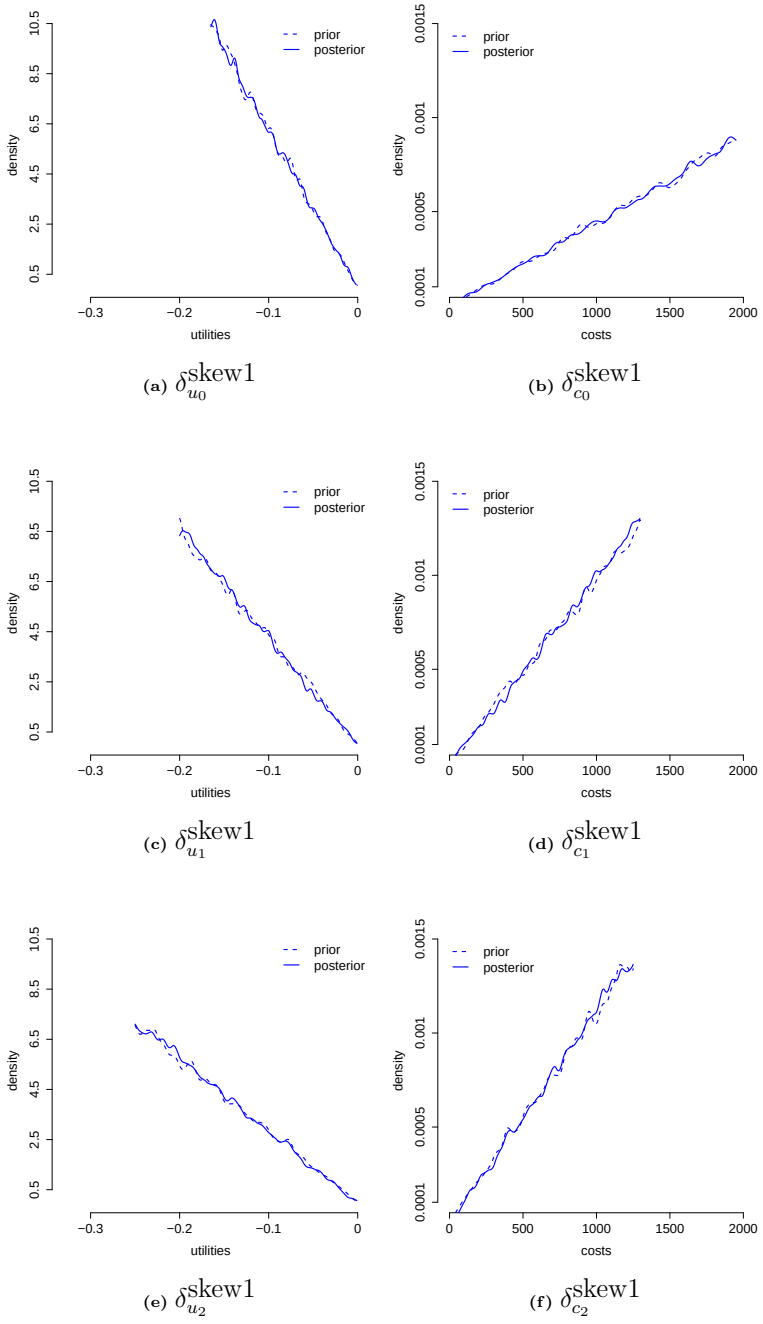


(a) $\delta^{\text{skew0}}_{u_0}$ (b) $\delta^{\text{skew0}}_{c_0}$

(c) $\delta^{\text{skew0}}_{u_1}$ (d) $\delta^{\text{skew0}}_{c_1}$

(e) $\delta^{\text{skew0}}_{u_2}$ (f) $\delta^{\text{skew0}}_{c_2}$

Figure C.19: Prior and posterior densities of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{skew0}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the intervention group.



(a) $\delta_{u_0}^{\text{skew1}}$ (b) $\delta_{c_0}^{\text{skew1}}$

(c) $\delta_{u_1}^{\text{skew1}}$ (d) $\delta_{c_1}^{\text{skew1}}$

(e) $\delta_{u_2}^{\text{skew1}}$ (f) $\delta_{c_2}^{\text{skew1}}$

Figure C.20: Prior and posterior densities of the distributions of the sensitivity parameters under the scenario $\boldsymbol{\delta}^{\text{skew1}}$ for both utilities (panels a,c and e) and costs (panels b,d and f) at $j = 0, 1, 2$ in the intervention group.

C.4.3 Posterior Estimates

Table C.15 compares the point and 95% CI estimates for the mean QALYs and costs and other incremental parameters (e.g. net benefit), in the two treatment groups of the PBS trial between eight alternative specifications of the modelling framework described in Section 6.1 and shown in Table 6.3. These are: CS-CC, CS-ALL, L-CC, L-ALL, $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$.

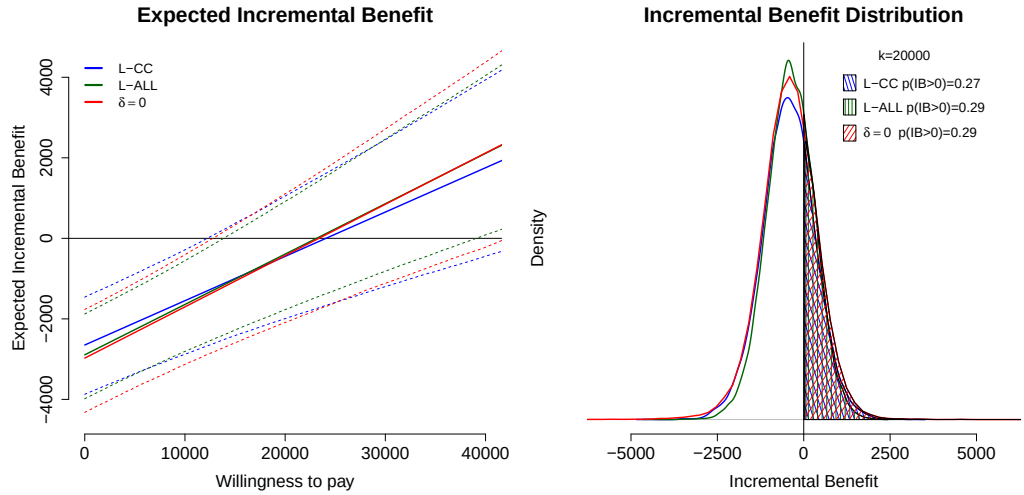
C.4.4 Alternative Missingness Scenarios

The results associated with the Expected Incremental Benefit and the Incremental Benefit distributions (evaluated at $k = £20,000$) under the three alternative nonignorable scenarios described

Parameter	CS-CC		CS-ALL		L-CC		L-ALL		$\delta = 0$		δ^{flat}		δ^{skew0}		δ^{skew1}	
	Mean	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**	**Mean**	**95% interval**
Control $(t = 1)$																
mean QALY (μ_{e1})	0.487	(0.452;0.524)	0.488	(0.457;0.521)	0.477	(0.447;0.507)	0.462	(0.434;0.491)	0.461	(0.428;0.494)	0.417	(0.364;0.467)	0.432	(0.386;0.478)	0.403	(0.353;0.452)
mean cost (μ_{c1})	3073	(2188;3915)	3284	(2459;4261)	3382	(2583;4246)	3126	(2484;3846)	3233	(2490;4042)	3401	(2616;4196)	3346	(2606;4173)	3457	(2678;4250)
Intervention $(t = 2)$																
mean QALY (μ_{e2})	0.611	(0.570;0.651)	0.595	(0.564;0.628)	0.587	(0.554;0.620)	0.587	(0.555;0.618)	0.588	(0.556;0.622)	0.570	(0.533;0.607)	0.576	(0.541;0.612)	0.564	(0.526;0.601)
mean cost (μ_{c2})	5768	(5115;6413)	5919	(5385;6506)	6031	(5282;6889)	6018	(5315;6807)	6208	(5364;7143)	6319	(5463;7272)	6282	(5409;7201)	6356	(5522;7332)
Incremental																
QALY differential (Δ_e)	0.12	(0.07;0.18)	0.11	(0.06;0.15)	0.11	(0.07;0.16)	0.13	(0.08;0.17)	0.13	(0.08;0.17)	0.15	(0.09;0.22)	0.14	(0.09;0.21)	0.16	(0.10;0.22)
Cost differential (Δ_c)	2694	(1609;3785)	2635	(1416;3616)	2649	(1459;3872)	2893	(1878;3984)	2975	(1764;4321)	2918	(1696;4290)	2936	(1710;4300)	2898	(1672;4268)
IB (at $k = 20000$)	-227	(-1904;1442)	-507	(-1843;935)	-447	(-1993;1051)	-391	(-1770;914)	-428	(-2090;1106)	135	(-1746;1932)	-58	(-1859;1677)	321	(-1565;2086)
ICER	21843		24761		24070		23130		23362		19116		20403		18008	

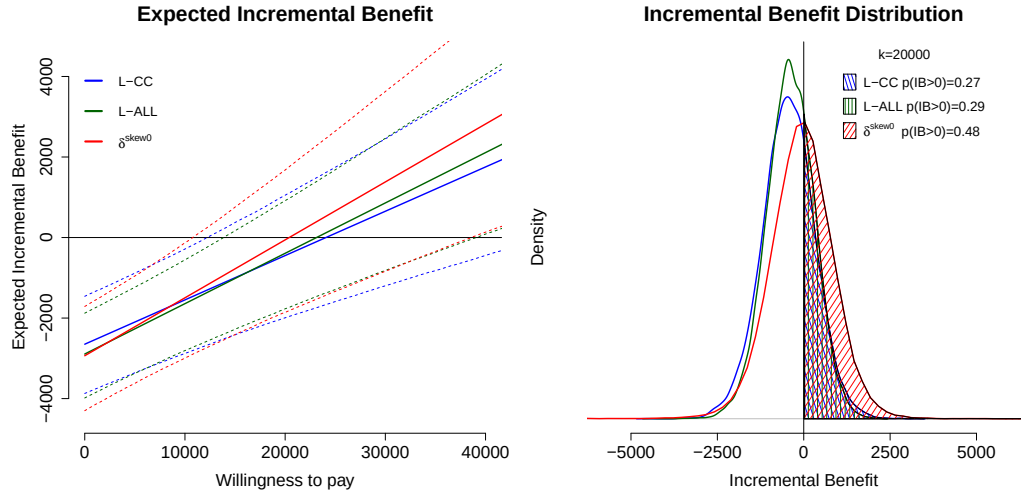
Table C.15: Means and 95% credible interval estimates of the mean QALYs and cost parameters for the control ($t = 1$) and intervention ($t = 2$) group in the PBS trial obtained from 8 different scenarios: CS-CC, CS-ALL,L-CC, L-ALL, $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\delta}^{\text{flat}}$, $\boldsymbol{\delta}^{\text{skew0}}$ and $\boldsymbol{\delta}^{\text{skew1}}$. Mean QALYs and cost differentials, the incremental benefits at $k = 20000$ and the ICERs are also reported. Cost values are expressed in £.

in Chapter 6 are separately compared with those from the L-CC and L-ALL scenarios, respectively indicated with red, blue and green lines and shaded areas in the graphs. The three scenarios are : the benchmark scenario $\boldsymbol{\delta} = \mathbf{0}$ (Figure C.21), $\boldsymbol{\delta}^{\text{skew0}}$ (Figure C.22) and $\boldsymbol{\delta}^{\text{skew1}}$ (Figure C.23).



(a) Expected Incremental Benefit **(b)** Incremental Benefit Distribution

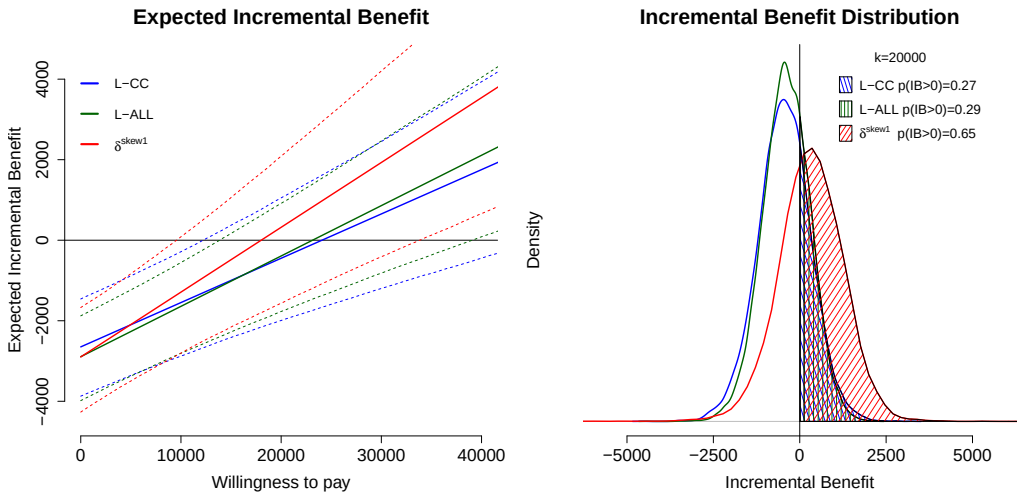
Figure C.21: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with L-CC (blue solid and dashed lines), L-ALL (green solid and dashed lines) and $\boldsymbol{\delta} = \mathbf{0}$ (red solid and dashed lines) models fitted to the data from the PBS study.



(a) Expected Incremental Benefit **(b)** Incremental Benefit Distribution

Figure C.22: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with L-CC (blue solid and dashed lines), L-ALL (green solid and dashed lines) and $\boldsymbol{\delta}^{\text{skew0}}$ (red solid and dashed lines) models fitted to the data from the PBS study.

The results associated with the Cost-Effectiveness Planes under the three alternative non-ignorable scenarios described in Chapter 6 are separately compared with those from the L-CC and L-ALL scenarios, respectively indicated with red, blue and green coloured dots in the graphs. The three scenarios are : the benchmark scenario $\boldsymbol{\delta} = \mathbf{0}$ (Figure C.24), $\boldsymbol{\delta}^{\text{skew0}}$ (Figure C.25) and $\boldsymbol{\delta}^{\text{skew1}}$ (Figure C.26). In each graph, The ICERs from each model are indicated with corresponding darker coloured dots. Compared with the benchmark $\boldsymbol{\delta} = \mathbf{0}$, the position of the points in the plane under $\boldsymbol{\delta}^{\text{skew0}}$ and, especially $\boldsymbol{\delta}^{\text{skew1}}$, is shifted to the right. At a willingness to pay of $k = £25,000$, the ICERs under these two scenarios fall in the sustainability area and, on average,



(a) Expected Incremental Benefit

(b) Incremental Benefit Distribution

Figure C.23: Expected Incremental Benefit (panel a) and Incremental Benefit distribution at $k = 20000$ (panel b) associated with L-CC (blue solid and dashed lines), L-ALL (green solid and dashed lines) and $\boldsymbol{\delta}^{\text{skew1}}$ (red solid and dashed lines) models fitted to the data from the PBS study.

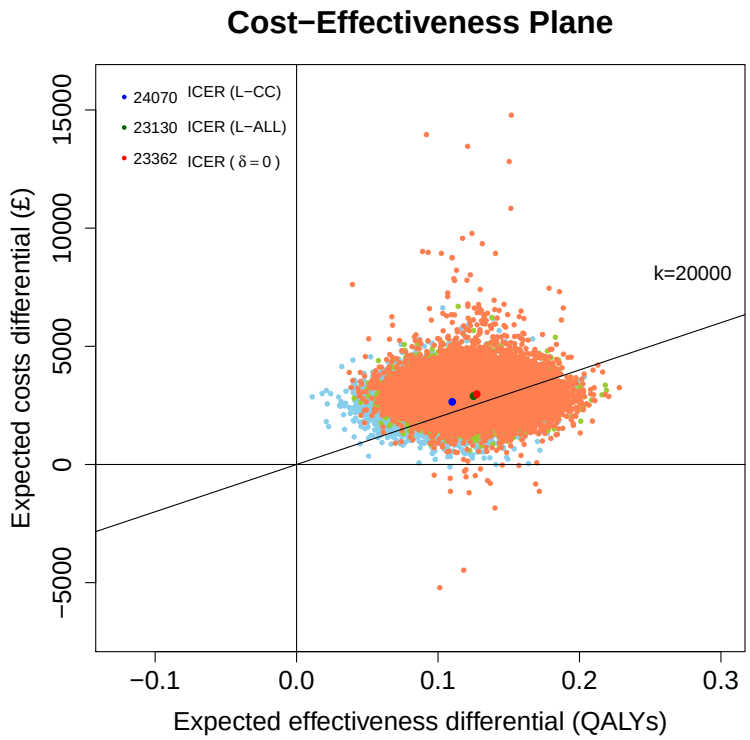


Figure C.24: CEPs associated with L-CC, L-ALL and $\boldsymbol{\delta} = \mathbf{0}$ scenarios, respectively represented with blue, green and red coloured dots.

indicate a higher cost-effectiveness for the new intervention compared with the results under CC, MAR and $\boldsymbol{\delta} = \mathbf{0}$.

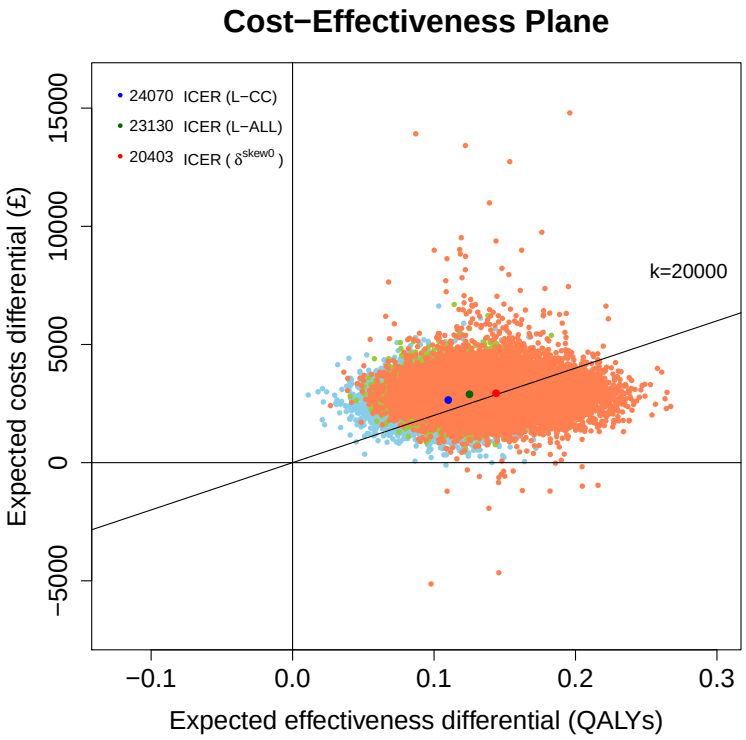


Figure C.25: CEPs associated with L-CC, L-ALL and δ^{skew0} scenarios, respectively represented with blue, green and red coloured dots.

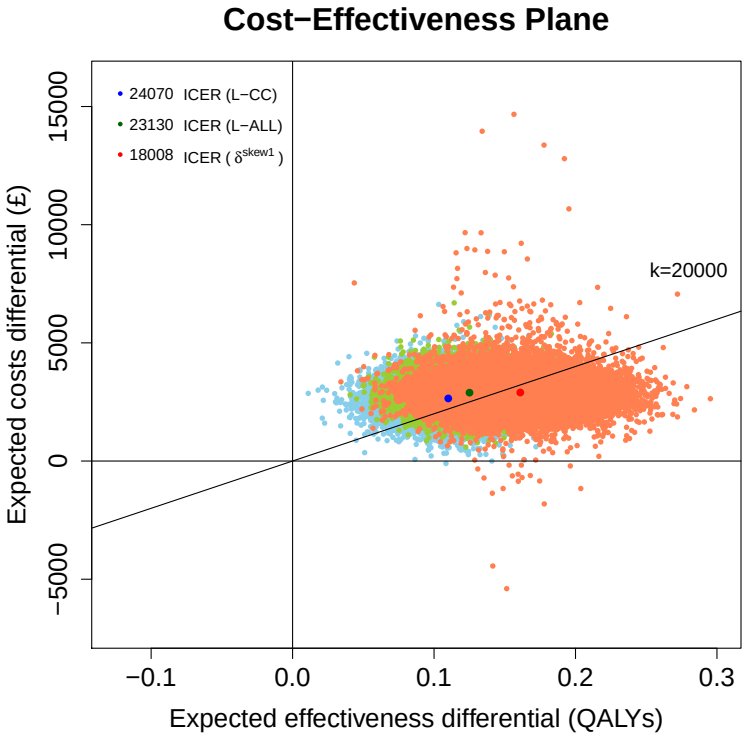


Figure C.26: CEPs associated with L-CC, L-ALL and δ^{skew1} scenarios, respectively represented with blue, green and red coloured dots.

Appendix D

missingHE: A R Package to Handle Missing Data in Economic Evaluations

D.1 Package Overview

missingHE is a package designed to perform Bayesian health economic analysis in the presence of missing outcome data. The structure of the package is based on the modelling framework presented in Chapter 5. In the current version of missingHE there are two main functions to handle missingness in either or both the effectiveness and cost variables. These are selection and hurdle, which allow to implement selection and hurdle models, respectively.

Selection models are nonignorable models that factor the full data distribution $(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{\omega})$ into the product of the full data response model $p(\boldsymbol{y} \mid \boldsymbol{\omega})$ and the missing data mechanism $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\omega})$. Selection models are attractive as they allow a direct specification of the distribution of the response and straightforwardly formulate assumptions about the nonresponse mechanism. The drawback is how we can translate these assumptions into assumptions on the distribution of the missing data. Indeed, model identification depends on assumptions on the distribution of $\boldsymbol{y}$ (often difficult to check) and on the form of the missingness model (on which unverifiable assumptions have to be made). The best way to assess the impact of missingness on the results for selection models is to vary both the distributional assumptions in the response model and the form of the missing data mechanism, and assess the robustness of the conclusions to a range of plausible scenarios.

For the purpose of this thesis, here we only focus on the function hurdle, which allows to implement hurdle models, and describe its main features and purpose. Figure D.1 shows a schematic representation of the package.

In the figure, the rectangular box indicates the main function hurdle of the package, which returns as output an object of class missingHE. Oval boxes indicate generic functions (such as print or plot) that can be applied to an object of class missingHE to generate the desired output. The diamond boxes identify the functions that are specific to missingHE. These functions can be used to compute suitable measures that assess the performance of the model either in terms of predictive accuracy, using different types of predictive information criteria (pic), or in terms of MCMC convergence, using different types of diagnostic tools (diagnostic).

The current version of missingHE can be installed in R from the online GitHub repository (https://github.com/AnGabrio/missingHE). For example, the function install_github, included in the

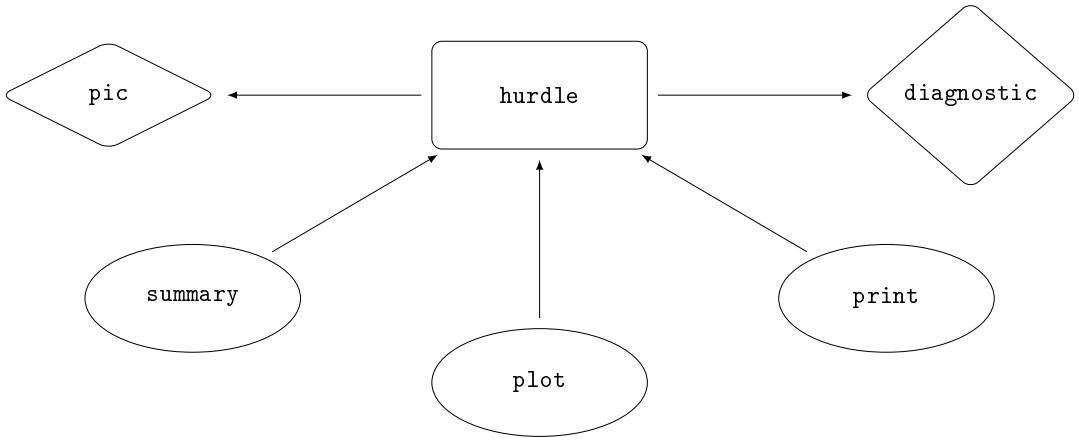


Figure D.1: A schematic representation of the `missingHE` package.

package `devtools` (which should be loaded), can be used to install the package. Once installed, the package can be loaded into the `R` workspace using the function `require` or `library`.

Suppose that the user has a suitable dataset, perhaps obtained from a RCT, in which data for each individual are recorded for the effectiveness and cost variables as well as for an arm indicator specifying whether the individual to whom the data refer belongs in the control or the intervention group. Other variables may be observed, e.g. relevant covariates such as sex, age or co-morbidity. Both outcome variables can have missing values while no unobserved values are allowed for the covariates as `missingHE` can only deal with missingness in the outcomes.

We now provide a brief summary of the `hurdle` function along with an explanation of the main inputs that need to be provided by the user and the outputs generated by the function.

D.2 The `hurdle` Function

Assume that the data are available in the `R` workspace as a data frame, say `data`. Hurdle models are implemented in `missingHE` using the `hurdle` function which calls `JAGS`, through the `R` package `R2jags`, to fit the model. This function processes the data frame `data` so that the model output is in the correct form for other functions in the package `missingHE`. Therefore, when this function is called it should be assigned to an object to create an object of class `missingHE`. For example, we can generate the object `model.hur` using the following command.

```
R> model.hur <- hurdle(data = data, dist_e = "norm", dist_c = "norm",
+ model.eff = e ~ 1, model.cost = c ~ e,
+ model.se = se ~ 1, model.sc = sc ~ 1, se = 1, sc = 0,
+ type = "SCAR", n.iter = 10000, prior = "default", d_e = my.d_e)
```

The arguments of the function have the following interpretations:

- `data`: a data frame that must contain the data to analyse, specified in a data frame format
- `dist_e` and `dist_c`: assumed effectiveness and cost distributions, specified as character names among a set of pre-defined choices. Current available choices are: Normal (`"norm"`) for both outcomes, Beta (`"beta"`) for the effectiveness and Gamma (`"gamma"`) or LogNormal (`"lnorm"`) for the costs.
- `model.eff` and `model.cost`: formulae that specify which variables should be included in the effectiveness and cost models as covariates (among those available in `data`). A joint bivariate distribution can be assumed by placing `e` on the right-hand side of the formula

for the costs. By default both formulae do not contain any covariate (indicated with `1`) and assume independence between the outcomes.

- `model.se` and `model.sc`: formulae that specify which variables should be included in the effectiveness and cost structural value models as covariates (among those available in `data`). By default no covariate is included (indicated with `1`).

- `type`: type of model for the structural values, either independent of any other variable (called Structural Completely At Random, `"SCAR"`) or including some covariates in the model (called Structural At Random, `"SAR"`).

- `se` and `sc`: values in the effectiveness and cost data should be treated as structural by the model (e.g. 1 for `e` and 0 for `c`). If structural values are observed only for one outcome it is possible to set either `se = NULL` or `sc = NULL`. In this case, no hurdle model is assumed for that outcome, which is directly modelled using the distribution specified in `dist_e` or `dist_c`.

- `prior`: prior distributions to be assumed for the parameters of the model (by default vague priors). The default priors can be overwritten by the user. In this case, the new hyperpriors for each parameter in the model can be provided by creating a list object that contains the new values. A description of the parameterisation of the model and a list of the character names to be used to change each parameter's prior can be accessed by typing `help(hurdle)`.

- `d_e` and `d_c`: (optional) vectors of the structural value indicators to be used in the model for the effectiveness and costs. If not provided, `missingHE` internally computes these vectors based on the observed cases (`NA` if the cases are missing). When provided, the argument of `d_e` and `d_c` must be vectors of length equal to the number of rows in `data`, and should take value 1 or 0 to respectively associate each (missing) case with the (assumed) structural or non-structural component in the hurdle model.

- Other additional arguments that may be provided are: the burnin period to be discarded (`n.burnin`), the number of the chains (`n.chains`), the thinning interval (`n.thin`), the initialised values for the parameters in each chain (`inits`), the upper and lower bounds of the credible intervals for describing the uncertainty around the imputed values (`prob`) and whether the model text file should be saved in the current working directory (`save_model`).

The object `model.hur` contains the following elements which are then used as inputs to the other functions in the `missingHE` package.

- `model_output`: a list storing the output of the `JAGS` model. Depending on the type of model, the elements in this list can vary as they contain the posterior samples of the parameters of interest based on the model specification assumed. In the list, a summary of the posterior estimates of the `JAGS` model is also available, taken directly from the output of the function `jags` in the package `R2jags`.

- `cea`: another list that stores the output of the economic evaluation based on the posterior samples of the marginal mean effectiveness and cost parameters. This object can be analysed using tailored functions in the package `BCEA` to visually represent standard CEA outputs, such as the CEP and the CEAC.

- `type`: a character name that reports the type of model assumed for the structural values.

A more detailed explanation of the inputs of each function in the package as well as a running example on how to perform economic evaluations using `missingHE` can be found in the package documentation, which is available at the GitHub repository `https://github.com/AnGabrio/missingHE/blob/master/inst/doc/missingHE.pdf`.

An area of possible extension of the current version of `missingHE` is the incorporation of a wider range of choice for the distributions of the effectiveness and cost variables. This could provide the users with more flexibility when fitting the model and compare the performance of alternative specifications of the model in terms of the fit to the observed data.

In addition, more functions for fitting different types of missing data models will be included in the package. Specifically, the implementation of alternative approaches to deal with nonignorbale missingness (e.g. pattern mixture models) will provide greater flexibility for conducting sensitivity analysis to the missingness assumptions and assess their impact on both the inferences and the cost-effectiveness conclusions.

Appendix E

Literature Review Articles

Aasa, M., Henriksson, M., Dellborg, M., Grip, L., Herlitz, J., Levin, L., Svensson, L., and Janzon, M. (2010). Cost and health outcome of primary percutaneous coronary intervention versus thrombolysis in acute st-segment elevation myocardial infarction results of the swedish early decision reperfusion study (swedes) trial. *American Heart Journal*, 160:322–328.

Andersson, E., Ljotsson, B., Hedman, E., Mattson, S., Enander, J., Andersson, G., Kaldor, V., Lindefors, N., and Ruck, C. (2015). Cost-effectiveness of an internet-based booster program for patients with obsessive compulsive disorder: Results from a randomized controlled trial. *Journal of Obsessive-Compulsive and Related Disorders*, 4:14–19.

Araya, R., Flynn, T., Rojas, G., Fritsch, R., and Simon, G. (2006). Cost-effectiveness of a primary care treatment program for depression in low-income women in Santiago, Chile. *American Journal of Psychiatry*, 163:1379–1387.

Asha, S., Chan, A., Walter, E., Kelly, P., Morton, R., Ajami, A., Wilson, R., and Honneyman, D. (2014). Impact from point-of-care devices on emergency department patient processing times compared with central laboratory testing of blood samples: a randomised controlled trial and cost-effectiveness analysis. *Emerg Med J*, 31:714–719.

Atthobari, J., Asselbergs, F., Boersma, C., de Vries, R., Hillege, H., van Gilst, W., Gansevoort, R., de Jong, P., de Jong-van den Berg, L., and Postma, M. (2006). Cost-effectiveness of screening for albuminuria with subsequent fosinopril treatment to prevent cardiovascular events: a pharmacoeconomic analysis linked to the prevention of renal and vascular endstage disease (prevend) study and the prevention of renal and vascular endstage disease intervention trial (prevend it). *Clinical Therapeutics*, 28:432–444.

Barrett, B., Byford, S., Crawford, M., Patton, R., Drummond, C., Henry, J., and Touquet, R. (2006). Costeffectiveness of screening and referral to an alcohol health worker in alcohol misusing patients attending an accident and emergency department: a decision-making approach. *Drug and Alcohol Dependence*, 81:47–54.

Barton, G., Fairall, L., Bachmann, M., Uebel, K., Timmerman, V., Lombard, C., and Zwarenstein, M. (2013). Cost-effectiveness of nurse-led versus doctor-led antiretroviral treatment in south africa: pragmatic cluster randomised trial. *Tropical Medicine and International Health*, 18:769–777.

Barton, G., Hodjekins, J., Mugford, M., Jones, P., Croudace, T., and Fowler, D. (2009a). Cognitive behaviour therapy for improving social recovery in psychosis: Cost-effectiveness analysis. *Schizophrenia Research*, 112:158–163.

Barton, G., Sach, T., Jenkinson, C., Doherty, M., Avery, A., and Muir, K. (2009b). Lifestyle interventions for knee pain in overweight and obese adults aged over 45: economic evaluation of randomised controlled trial. *BMJ*, 339:2273.

Berkhof, F., Hesselink, A., Vaessen, D., Uil, S., Kerstjens, H., and van der Berg, J. (2014). The effect of an outpatient care on-demand system on health status and costs in patients with COPD. A randomized trial. *Respiratory Medicine*, 108:1163–1170.

Bijen, C., Vermeulen, K., Mourits, M., Arts, H., ter Brugge, H., van der Sijde, R., Wijma, J., Bongers, M., van der Zee, A., and de Bock, G. (2011). Cost effectiveness of laparoscopy versus laparotomy in early stage endometrial cancer: A randomised trial. *Gynecologic Oncology*, 121:76–82.

Bos, I., Hoving, J., van Tulder, M., Molken, M., Ader, H., de Vet, H., Koes, B., Vondeling, H., Bouter, L., and Mullner, M. (2003). Cost effectiveness of physiotherapy, manual therapy, and general practitioner care for neck pain: economic evaluation alongside a randomised controlled trial commentary: Bootstrapping simplifies appreciation of statistical inferences. *BMJ*, 326:911.

Brouwers, E., Bruijne, M., Terluin, B., and Verhaak, P. (2007). Cost-effectiveness of an activating intervention by social workers for patients with minor mental disorders on sick leave: a randomized controlled trial. *The European Journal of Public Health*, 17:214–220.

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.

Byford, S., Barrett, B., Roberts, C., Clark, A., Edwards, V., Smethurst, N., and Gowers, S. (2007a). Economic evaluation of a randomised controlled trial for anorexia nervosa in adolescents. *British Journal of Psychiatry*, 191:436–440.

Byford, S., Barrett, B., Roberts, C., Wilkinson, P., Dubicka, B., Kelvin, R., White, L., Ford, C., Breen, S., and Goodyer, I. (2007b). Cost-effectiveness of selective serotonin reuptake inhibitors and routine specialist care with and without cognitive behavioural therapy in adolescents with major depression. *British Journal of Psychiatry*, 191:521–527.

Campbell, A., Nunes, E., Miele, G., Metthews, A., Polsky, D., Ghitza, U., Turrigiano, E., Bailey, G., VanVeldhuisen, P., Chapdelaine, R., Froias, A., Stitzer, M., Carroll, K., Winhusen, T., Clingerman, S., Perez, L., McClure, E., Goldman, B., and Crowell, A. (2012). Design and

methodological considerations of an effectiveness trial of a computer-assisted intervention: An example from the nida clinical trials network. *Contemporary Clinical Trials*, 33:386–395.

Carr, A., Cooper, C., Campbell, M., Rees, J., Moser, J., Beard, D., Fitzpatrick, R., Gray, A., Dawson, J., Murphy, J., Bruhn, H., Cooper, D., and Ramsay, C. (2015). Clinical effectiveness and cost-effectiveness of open and arthroscopic rotator cuff repair [the uk rotator cuff surgery (ukuff) randomised trial]. *Health Technology Assessment*, 19.

CLOTS Trials Collaboration (2014). Effect of intermittent pneumatic compression on disability, living circumstances, quality of life, and hospital costs after stroke: secondary analyses from clots 3, a randomised trial. *Lancet Neurol*, 13:1186–1192.

Coast, J., Noble, A., Horrocks, S., Asim, O., Peters, T., and Salisbury, C. (2005). Economic evaluation of a general practitioner with special interests led dermatology service in primary care. *BMJ*, 331:1444–1449.

Costa, M., Achten, J., Parsons, N., Edlin, R., Foguet, P., Prakash, U., and Griffin, D. (2012). Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial. *BMJ*, 344.

Coupe, V., Veenhof, C., van Tulder, M., Dekker, J., Bjlsma, J., and Van den Ende, C. (2007). The cost effectiveness of behavioural graded activity in patients with osteoarthritis of hip and/or knee. *Annals of the Rheumatic Diseases*, 66:215–221.

Cuthbertson, B., Rattray, J., Campbell, M., Gager, M., Roughton, S., Smith, A., Hull, A., Breeman, S., Norrie, J., Jenkinson, D., Hernandez, R., Johnston, M., Wilson, E., and Waldmann, C. (2009). The practical study of nurse led, intensive care follow-up programmes for improving long term outcomes fromcritical illness: a pragmatic randomised controlled trial. *BMJ*, 339.

Dakin, H., Wordsworth, S., Rogers, C., Abangma, G., Raftery, J., Harding, S., Lotery, A., Downes, S., Chakravarthy, U., and Reeves, B. (2014). Cost-effectiveness of ranibizumab and bevacizumab for age-related macular degeneration: 2-year findings from the ivan randomised trial. *BMJ*, 4.

D'Amico, F., Rehill, A., Knapp, M., Aguirre, E., Donovan, H., Hoare, Z., Hoe, J., Russell, I., Spector, A., Streater, A., Whitaker, C., Woods, R., and Orrell, M. (2015). Maintenance cognitive stimulation therapy: An economic evaluation within a randomized controlled trial. *JAMDA*, 16:63–70.

De Beurs, D., Bosmans, J., de Groot, M., de Keijser, J., van Duijn, E., de Winter, R., and Kerkhof, A. (2015). Training mental health professionals in suicide practice guideline adherence:cost-effectiveness analysis alongside a randomized controlled trial. *Journal of Affective Disorders*, 186:203–210.

Delaney, B., Qume, M., Moayyedi, P., Logan, R., Ford, A., Elliott, C., McNulty, C., Wilson, S., and Hobbs, F. (2008). Helicobacter pylori test and treat versus proton pump inhibitor in initial management of dyspepsia in primary care: multicentre randomised controlled trial (mrc-cube trial). *BMJ*, 336:651–654.

Dennis, M., Godley, S., Diamond, G., Tims, F., Babor, T., Donaldons, J., Liddle, H., Titus, J., Kaminer, Y., Webb, C., Hamilton, N., and Funk, R. (2004). The cannabis youth treatment (cyt) study: main findings from two randomized trials. *Journal of Substance Abuse Treatment*, 27:197–213.

Djkgraaf, M., van der Zanden, B., de Borgie, C., Blanken, P., van Ree, J., and van den Brink, W. (2005). Cost utility analysis of co-prescribed heroin compared with methadone maintenance treatment in heroin addicts in two randomised trials. *BMJ*, 330:1297.

Domino, M., Foster, E., Vitiello, B., Kratochvil, C., Burns, B., Silva, S., Reinecke, M., and March, J. (2009). Relative cost-effectiveness of treatments for adolescent depression: 36-week results from the tads randomized trial. *J. AM. ACAD. CHILD ADOLESC. PSYCHIATRY*, 48.

Dornelas, E., Magnavita, J., Beazoglou, T., Fischer, E., Oncken, C., Lando, H., Greene, J., Barbagallo, J., Stepnowski, R., and Gregonis, E. (2006). Efficacy and cost-effectiveness of a clinic-based counseling intervention tested in an ethnically diverse sample of pregnant smokers. *Patient Education and Counseling*, 64:342–349.

Drummond, M., Becker, D., Hux, M., Chancellor, J., Duprat-Lomon, I., Kubin, R., and Sagnier, P. (2003). An economic evaluation of sequential iv/po moxifloxacin therapy compared to iv/po co-amoxiclav with or without clarithromycin in the treatment of community-acquired pneumonia. *Chest*, 124:526–535.

Duarte, A., Walker, J., Walker, S., Richardson, G., Hansen, C., Martin, P., Murray, G., Sculpher, M., and Sharpe, M. (2015). Cost-effectiveness of integrated collaborative care for comorbid major depression in patients with cancer. *Journal of Psychosomatic Research*, 79:465–470.

Edwards, R., Ceilleachair, A., Bywater, T., Hughes, D., and Hutchings, J. (2007). Parenting programme for parents of children at risk of developing conduct disorder: cost effectiveness analysis. *BMJ*, 334:682.

Emmons, K., Puleo, E., Park, E., Gritz, E., Butterfield, R., Weeks, J., Mertens, A., and Li, F. (2005). Peer-delivered smoking counseling for childhood cancer survivors increases rate of cessation: the partnership for health study. *Journal of Clinical Oncology*, 23:6516–6523.

Fals-Stewart, W., Klostermann, K., Yates, B., O'Farrell, T., and Birchler, G. (2005). Brief relationship therapy for alcoholism: a randomized clinical trial examining clinical efficacy and cost-effectiveness. *Psychology of Addictive Behaviors*, 19:363–371.

Fals-Stewart, W. and Lam, W. (2008). Brief behavioral couples therapy for drug abuse: a randomized clinical trial examining clinical efficacy and cost-effectiveness. *Families, Systems and Health*, 26:377–392.

Felker, G., Ahmad, T., Anstrom, K., Adams, K., Cooper, L., Ezekowitz, J., Fiuzat, M., Houston-Miller, N., Januzzi, J., Leifer, E., Mark, D., Desvigne-Nickens, P., Paynter, G., Pina, I., Whellan, D., and O'Connor, C. (2014). Rationale and design of the guide-it study guiding evidence based

therapy using biomarker intensified treatment in heart failure. *JACC : HEART FAILURE*, 2:457–465.

Forster, A., Dickerson, J., Young, J., Patel, A., Kalra, A., Nixon, J., Smithard, D., Knapp, M., Holloway, I., Anwar, S., and Farrin, A. (2013). A structured training programme for caregivers of inpatients after stroke (tracs): a cluster randomised controlled trial and cost-eff ectiveness analysis. *Lancet*, 382:2069–2076.

Fuller, N., Colagiuri, S., Schofield, D., Olson, A., Shrestha, R., Holzapfel, C., Wolfenstretter, S., Holle, R., Ahern, A., Hauner, H., Jebb, S., and Caterson, I. (2013). A within-trial cost-effectiveness analysis of primary care referral to a commercial provider for weight loss treatment, relative to standard care—an international randomised controlled trial. *International Journal of Obesity*, 37:828–834.

Furze, G., Dumviille, J., Miles, J., Irvine, K., Thompson, D., and Lewin, R. (2009). 'prehabilitation' prior to cabg surgery improves physical functioning and depression. *International Journal of Cardiology*, 132:51–58.

Gilbert, F., Grant, A., Gillan, M., Vale, L., Campbell, M., Scott, N., Knight, D., and Wardlaw, D. (2004). Low back pain: influence of early mr imaging or ct on treatment and outcome—multicenter randomized trial. *Radiology*, 231:343–351.

Gillespie, P., O'Shea, E., Casey, D., Murphy, K., Devane, D., Cooney, A., Mee, L., Kirwan, C., CmCarthy, B., and Newell, J. (2013). The cost-effectiveness of a structured education pulmonary rehabilitation programme for chronic obstructive pulmonary disease in primary care: the prince cluster randomised trial. *BMJ*, 3.

Gillett, M., Dallosso, H., Dixon, S., Brennan, A., Carey, M., Campbell, M., Heller, S., Khunti, K., Skinner, T., and Davies, M. (2010). Delivering the diabetes education and self management for ongoing and newly diagnosed (desmond) programme for people with newly diagnosed type 2 diabetes: cost effectiveness analysis. *BMJ*, 341.

Godley, S., Garner, B., Passetti, L., Funk, R., Dennis, M., and Godley, M. (2010). Adolescent outpatient treatment and continuing care: Main findings from a randomized clinical trial. *Drug and Alcohol Dependence*, 110:44–54.

Goodacre, S., Nicholl, J., Dixon, S., Cross, E., Angelini, K., Arnold, J., Revill, S., Locker, T., Capewell, S., Quinney, D., Campbell, S., and Morris, F. (2004). Rct and economic evaluation of a chest pain observation unit compared with routine care. *BMJ*, 328:254.

Graff, M., Adang, E., Vernooij-Dassen, M., Dekker, J., Jonsson, L., Thijssen, M., Hoefnagels, W., and Rikkert, M. (2008). Community occupational therapy for older patients with dementia and their care givers: cost effectiveness study. *BMJ*, 336:134–138.

Green, J., Wood, A., Kerfoot, M., Trainor, G., Roberts, C., Rothwell, J., Woodham, A., Ayodeji, E., Barrett, B., Byford, S., and Harrington, R. (2011). Group therapy for adolescents with repeated self harm: randomised controlled trial with economic evaluation. *BMJ*, 342.

Group, A. C. (2004). Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000): randomised double-blind trial. *The Lancet*, 363:2105–2115.

Group, P. M. C. (2014). Long-term eff ectiveness of dopamine agonists and monoamine oxidase b inhibitors compared with levodopa as initial treatment for parkinson's disease (pd med): a large, open-label, pragmatic randomised trial. *Lancet*, 384:1196–1205.

Group, T. (2009). Options for managing low grade cervical abnormalities detected at screening: cost effectiveness study. *BMJ*, 339.

Haddock, G., Barrowclough, C., Tarrier, N., Moring, J., O'Brien, R., Schofield, N., Quinn, J., Palmer, S., Davies, L., Lowens, I., McGovern, J., and Lewis, S. (2003). Cognitive-behavioural therapy and motivational intervention for schizophrenia and substance misuse: 18-month outcomes of a randomised controlled trial. *British Journal of Psychiatry*, 183:418–426.

Hartman, M., van Ede, A., Severens, J., Laan, R., van de Putte, L., and van der Wilt, G. (2004). Economic evaluation of folate supplementation during methotrexate treatment in rheumatoid arthritis. *The Journal of Rheumatology*, 31:902–908.

Hedman, E., El Alaoui, S., Lindefors, N., Andersson, E., Ruck, C., Ghaderi, A., Kaldo, V., Lekander, M., Andersson, G., and Ljotsson, B. (2014). Clinical effectiveness and cost-effectiveness of internet- vs. groupbased cognitive behavior therapy for social anxiety disorder: 4-year follow-up of a randomized trial. *Behaviour Research and Therapy*, 59:20–29.

Heliovaara-Peippo, S., Hurskainen, R., Teperi, J., Aalto, A., Grenman, S., Halmesmaki, K., Jokela, M., Kivela, A., Tomas, E., Tuppurainen, M., and Paavonen, J. (2013). Quality of life and costs of levonorgestrel-releasing intrauterine system or hysterectomy in the treatment of menorrhagia: a 10-year randomized controlled trial. *American Journal of Obstetrics and Gynecologyl*, 535.

Henderson, C., Knapp, M., Fernandez, J., Beecham, J., Hirani, S., Beynon, M., Cartwright, M., Rixon, L., Doll, H., Bower, P., Steventon, A., Rogers, A., Fitzpatrick, R., Barlow, J., Bardsley, M., and Newman, S. (2013). Cost effectiveness of telehealth for patients with long term conditions (whole systems demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. *BMJ*, 346.

Henderson, C., Knapp, M., Fernandez, J., Beecham, J., Hirani, S., Beynon, M., Cartwright, M., Rixon, L., Doll, H., Bower, P., Steventon, A., Rogers, A., Fitzpatrick, R., Barlow, J., Bardsley, M., and Newman, S. (2014). Cost-effectiveness of telecare for people with social care needs: the whole systems demonstrator cluster randomised trial. *Age and Ageing*, 0:1–7.

Higginson, I., McCrone, P., Burman, R., Silber, E., and Edmonds, P. (2009). Is short-term palliative care cost-effective in multiple sclerosis? a randomized phase ii trial. *Journal of Pain and Symptom Management*, 38:816–826.

Hollinghurst, S., Redmond, N., Costelloe, C., Montgomery, A., Fletcher, M., Peters, T., and Hay, A. (2008a). Paracetamol plus ibuprofen for the treatment of fever in children (pitch): economic evaluation of a randomised controlled trial. *BMJ*, 337:1490.

Hollinghurst, S., Sharp, D., Ballard, K., Barnett, J., Beattie, A., Evans, M., Lewith, G., Middleton, K., Oxford, F., Webley, F., and Little, P. (2008b). Randomised controlled trial of alexander technique lessons, exercise, and massage (ateam) for chronic and recurrent back pain: economic evaluation. *BMJ*, 337:2656.

Hollis, J., McAfee, T., Fellows, J., Zbikowski, S., Stark, M., and Riedlinger, K. (2007). The effectiveness and cost effectiveness of telephone counselling and the nicotine patch in a state tobacco quitline. *Tobacco Control*, 16:53–59.

Honkoop, P., Loijmans, R., Termeer, E., Snoeck-Stroband, J., van den Hout, W., Bakker, M., Assendelft, W., ter Riet, G., Sterk, P., Schermer, T., and Sont, J. (2014). Symptom- and fraction of exhaled nitric oxide–driven strategies for asthma control: A cluster-randomized trial in primary care. *J ALLERGY CLIN IMMUNOL*, 135:683–688.

Hurskainen, R., Teperi, J., Rissanen, P., Aalto, A., Grenman, S., Kivela, A., Kujansuu, E., Vuorma, S., Yliskoski, M., and Paavonen, J. (2004). Clinical outcomes and costs with the levonorgestrel-releasing intrauterine system or hysterectomy for treatment of menorrhagia: randomized trial 5-year followup. *The Journal of the American Medical Association*, 291:1456–1463.

Jones, K., Colson, P., Holter, M., Lin, S., Valencia, E., Susser, E., and Wyatt, R. (2003). Cost-effectiveness of critical time intervention to reduce homelessness among persons with mental illness. *Psychiatric Services*, 54:884–890.

Jones, L., FitzGerald, G., Leurent, B., Round, J., Eades, J., Davis, S., Gishen, F., Holman, A., Hopkins, K., and Tookman, A. (2013). Rehabilitation in advanced, progressive, recurrent cancer: A randomized controlled trial. *Journal of Pain and Symptom Management*, 46:315–325.

Katon, W., Schoenbaum, M., Fan, M., Callahan, C., Williams, J., Hunkeler, E., Harpole, L., Zhou, X., Langston, C., and Unutzer, J. (2006a). Cost-effectiveness of improving primary care treatment of late-life depression. *Archives of General Psychiatry*, 62:1313–1320.

Katon, W., Unutzer, J., Fan, M., Williams, J., Schoenbaum, M., Lin, E., and Hunkeler, E. (2006b). Costeffectiveness and net benefit of enhanced treatment of depression for older adults with diabetes and depression. *Diabetes Care*, 29:265–270.

Kattan, M., Stearns, S., Crain, E., Stout, J., Gergen, P., EvansIII, R., Visness, C., Gruchalla, R., Morgan, W., O'Connor, G., Mastin, J., and Mitchell, H. (2005). Cost-effectiveness of a home-based environmental intervention for inner-city children with asthma. *Journal of Allergy and Clinical Immunology*, 116:1058–1063.

Kendrick, T., Peveler, R., Longworth, L., Baldwin, D., Moore, M., Chatwin, J., Thornett, A., Goddard, J., Campbell, M., Smith, H., Buxton, M., and Thompson, C. (2006a). Cost-effectiveness and cost-utility of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine: randomised controlled trial. *British Journal of Psychiatry*, 188:337–345.

Kendrick, T., Simons, L., Mynors-Wallis, L., Gray, A., Lathlean, J., Pickering, R., Harris, S., Rivero-Arias, O., Gerard, K., and Thompson, C. (2006b). Cost-effectiveness of referral for generic

care or problem-solving treatment from community mental health nurses, compared with usual general practitioner care for common mental disorders: randomised controlled trial. *British Journal of Psychiatry*, 189:50–59.

Kilonzo, M., Sambrook, A., Cook, J., Campbell, M., and Cooper, K. (2010). A cost-utility analysis of microwave endometrial ablation versus thermal balloon endometrial ablationvhe. *Value in Health*, 13:528–534.

Kilonzo, M., Vale, L., Cook, J., Milne, A., Stephen, A., and Avenell, A. (2007). A cost-utility analysis of multivitamin and multimineral supplements in men and women aged 65 years and over. *Clinical Nutrition*, 26:364–370.

Knapp, M., King, D., Romeo, R., Schehl, B., Barber, J., Griffin, M., Rapaport, P., Livingston, D., Mummery, C., Walker, Z., Hoe, J., Sampson, E., Cooper, C., and Livingston, G. (2013). Cost effectiveness of a manual based coping strategy programme in promoting the mental health of family carers of people with dementia (the start (strategies for relatives) study): a pragmatic randomised controlled trial. *BMJ*, 347.

Kolu, P., Raitanen, J., Rissanen, P., and Luoto, R. (2013). Cost-effectiveness of lifestyle counselling as primary prevention of gestational diabetes mellitus: Findings from a cluster-randomised trial. *PLOS ONE*, 8.

Krist, M., van Beijsterveldt, A., Backx, F., and de Wit, G. (2013). Preventive exercises reduced injury-related costs among adult male amateur soccer players: a cluster-randomised trial. *Journal of Physiotherapy*, 59:15–23.

Kuyken, W., Byford, S., Taylor, R., Watkins, E., Holden, E., White, K., Barrett, B., Byng, R., Evans, A., Mullan, E., and Teasdale, J. (2008). Mindfulness-based cognitive therapy to prevent relapse in recurrent depression. *Journal of Consulting and Clinical Psychology*, 76:966–978.

Kuyken, W., Hayes, R., Barrett, B., Byng, R., Dalgleish, T., Kessler, D., Lewis, G., Watkins, E., Brejcha, C., Cardy, J., Causley, A., Cowderoy, S., Evans, A., Gradinger, F., Kaur, S., Lanham, P., Morant, N., Richards, J., Shah, P., Sutton, H., Vicary, R., Weaver, A., Wilks, J., Williams, M., Taylor, R., and Byford, S. (2015). Eff ectiveness and cost-eff ectiveness of mindfulness-based cognitive therapy compared with maintenance antidepressant treatment in the prevention of depressive relapse or recurrence (prevent): a randomised controlled trial. *Lancet*, 386:63–73.

Ladapo, J., Elliott, M., Bogart, L., Kanouse, D., Vestal, K., Klein, D., Ratner, J., and Schuster, M. (2013). Cost of talking parents, healthy teens: A worksite-based intervention to promote parenteadolescent sexual health communication. *Journal of Adolescent Health*, 53:595–601.

Lall, R., Hamilton, P., Young, D., Hulme, C., Hall, P., Shah, S., MacKenzie, I., Tunnicliffe, W., Rowan, K., Cuthbertson, B., McCabe, C., and Lamb, S. (2015). A randomised controlled trial and cost-effectiveness analysis of high-frequency oscillatory ventilation against conventional artificial ventilation for adults with acute respiratory distress syndrome. the oscar (oscillation in ards) study. *Health Technology Assessment*, 19.

Lam, D., McCrone, P., Wright, K., and Kerr, N. (2005). Cost-effectiveness of relapse-prevention cognitive therapy for bipolar disorder. *British Journal of Psychiatry*, 186:500–506.

Lamb, S., Gates, S., Williams, M., Williamson, E., Mt-Isa, S., Withers, E., Castelnuovo, E., Smith, J., Ashby, D., Cooke, M., Petrou, S., and Underwood, M. (2013). Emergency department treatments and physiotherapy for acute whiplash: a pragmatic, two-step, randomised controlled trial. *Lancet*, 381:546–556.

Lamb, S., Hansen, Z., Lall, R., Castelnuovo, E., Withers, E., Nichols, V., Potter, R., and Underwood, M. (2010). Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-eff ectiveness analysis. *Lancet*, 375:916–923.

Lambeek, L., Bosmans, J., Van Royen, B., Van Tulder, M., Van Mechelen, W., and Anema, J. (2010). Effect of integrated care for sick listed patients with chronic low back pain: economic evaluation alongside a randomised controlled trial. *BMJ*, 341.

Lewis, M., James, M., Stokes, E., Hill, J., Sim, J., Hay, E., and Dziedzic, K. (2007). An economic evaluation of three physiotherapy treatments for non-specific neck disorders alongside a randomized trial. *Rheumatology*, 46:1701–1708.

Luik, A., Merkel, M., Hoeren, D., Riexinger, T., Kieser, M., and Schmitt, C. (2010). A randomized controlled noninferiority trial comparing isolation of the pulmonary veins with the cryoballoon catheter versus open irrigated radiofrequency ablation in patients with paroxysmal atrial fibrillation. *American Heart Journal*, 159:555–560.

Lynch, F., Dickerson, J., Saldana, L., and Fisher, P. (2014). Incremental net benefit of early intervention for preschool-aged children with emotional and behavioral problems in foster care. *Children and Youth Services Review*, 36:213–219.

Mak, S., Lee, M., Cheung, J., Choi, K., Chung, T., Wong, T., Lam, K., and Lee, D. (2015). Pressurised irrigation versus swabbing method in cleansing wounds healed by secondary intention: A randomised controlled trial with cost-effectiveness analysis. *International Journal of Nursing Studies*, 52:88–101.

Maljanen, T., Kenkt, P., Lindfors, O., Virtala, E., Tillman, P., and Harkanen, T. (2015). The cost-effectiveness of short-term and long-term psychotherapy in the treatment of depressive and anxiety disorders during a 5-year follow-up. *Journal of Affective Disorders*, 190:254–263.

Manca, A., Asseburg, C., Vergel, Y., Seymour, M., Meade, A., Stephens, R., Parmar, M., and Sculpher, M. (2012). The cost-effectiveness of different chemotherapy strategies for patients with poor prognosis advanced colorectal cancer (mrc focus). *Value in Health*, 15:22–31.

Manca, A., Dumville, J., Toregerson, D., Klaber Moffett, J., Mooney, M., Jackson, D., and Eaton, S. (2007). Randomized trial of two physiotherapy interventions for primary care back and neck pain patients: cost effectiveness analysis. *Rheumatology*, 46:1495–1501.

Manca, A., Sculpher, M., Ward, K., and Hilton, P. (2003). A cost-utility analysis of tension-free vaginal tape versus colposuspension for primary urodynamic stress incontinence. *An International Journal of Obstetrics and Gynaecology*, 110:255–262.

Mandelblatt, J., Cullen, J., Lawrence, W., Stanton, A., Yi, B., Kwan, L., and Ganz, P. (2008). Economic evaluation alongside a clinical trial of psycho-educational interventions to improve adjustment to survivorship among patients with breast cancer. *Journal of Clinical Oncology*, 26:1684–1690.

Maniadakis, N., Dafni, U., Fragoulakis, V., Grimani, I., Galani, E., Fragkoulidi, A., and Fountzilas, G. (2009). Economic evaluation of taxane-based first-line chemotherapy in the treatment of patients with metastatic breast cancer in greece: an analysis alongside a multicenter, randomized phase iii clinical trial. *Annals of Oncology*, 20:278–285.

Marson, A., Al-Kharusi, A., Alwaidh, M., Appleton, R., Baker, G., Chadwick, D., Cramp, C., Cockerell, O., Cooper, P., Doughty, J., Eaton, B., Gamble, C., Goulding, P., Howell, S., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G., Leach, J., Licolaides, P., Roberts, R., Shackley, P., Shen, J., Smith, D., Smith, P., Smith, C., Vanoli, A., and Williamson, P. (2003a). The sanad study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet*, 363:1016–1026.

Marson, A., Al-Kharusi, A., Alwaidh, M., Appleton, R., Baker, G., Chadwick, D., Cramp, C., Cockerell, O., Cooper, P., Doughty, J., Eaton, B., Gamble, C., Goulding, P., Howell, S., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G., Leach, J., Licolaides, P., Roberts, R., Shackley, P., Shen, J., Smith, D., Smith, P., Smith, C., Vanoli, A., and Williamson, P. (2003b). The sanad study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet*, 363:1000–1015.

Mary Davies, L., Anne Fargher, E., Tricker, K., Dawnes, P., Scott, D., and Symmons, D. (2007). Is shared care with annual hospital review better value for money than predominantly hospital-based care in patients with established stable rheumatoid arthritis? *Annals of the Rheumatic Diseases*, 66:658–663.

McCollister, K., Yang, X., and McKay, J. (2015). Cost-effectiveness analysis of a continuing care interventionfor cocaine-dependent adults. *Drug and Alcohol Dependence*, 158:38–44.

McCrone, P., Knapp, M., Proudfoot, J., Ryden, C., Cavanagh, K., Shapiro, D., Ilson, S., Gray, J., Goldberg, D., Mann, A., Marks, I., Everitt, B., and Tylee, A. (2004). Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. *British Journal of Psychiatry*, 185:55–62.

McKenna, C., Bojke, L., Manca, A., Adebajo, A., Dickson, J., Helliwell, P., Morton, V., Russell, I., Torgerson, D., and Watson, J. (2009). Shoulder acute pain in primary health care: is retraining gps effective? the sapphire randomized trial: a cost-effectiveness analysis. *Rheumatology*.

Melis, R., Adang, E., Teerenstra, S., van Eijken, M., Wimo, A., Achterberg, T., Lisdonk, E., and Rikkert, M. (2008). Multidimensional geriatric assessment: back to the future cost-effectiveness of a multidisciplinary intervention model for community-dwelling frail older people. *Journals of Gerontology Series A: Biological and Medical Sciences*, 63:275–282.

Meuldijk, D., Carlier, I., van Vliet, I., van der Akker-Marle, M., and Zitman, F. (2012). A randomized controlled trial of the efficacy and cost-effectiveness of a brief intensified cognitive behavioral therapy and/or pharmacotherapy for mood and anxiety disorders: Design and methods. *Contemporary Clinical Trials*, 33:983–992.

Munro, J., Nicholl, J., Brazier, J., Davey, R., and Cochrane, T. (2004). Cost effectiveness of a community based exercise programme in over 65 year olds: cluster randomised trial. *Journal of Epidemiology and Community Health*, 58:1004–1010.

Najafzadeh, M., Marra, C., Sadatsafavi, M., Aaron, S., Sullivan, S., Vandemheen, K., Jones, P., and Fitzgerald, J. (2008). Cost effectiveness of therapy with combinations of long acting bronchodilators and inhaled steroids for treatment of copd. *Thorax*, 63:962–967.

Nathoe, H., van Dijk, D., Jansen, E., Suyker, W., Diephuis, J., van Boven WJ., de la Riviere, A., Borst, C., Kalkman, C., Grobbee, D., Buskens, E., and de Jaegere PPT. (2003). A comparison of on-pump and off-pump coronary bypass surgery in low-risk patients. *New England Journal of Medicine*, 348:394–402.

Ninot, G., Moullec, G., Picot, M., Jaussent, A., Hayot, M., Desplan, M., Brun, J., Mercier, J., and Prefaut, C. (2011). Cost-saving effect of supervised exercise associated to copd self-management education program. *Respiratory Medicine*, 105:377–385.

Noben, C., Smit, F., Nieuwenhuijsen, K., Ketelaar, S., Gartner, F., Boon, B., Sluiter, J., and Evers, S. (2014). Comparative cost-effectiveness of two interventions to promote work functioning by targeting mental health complaints among nurses: Pragmatic cluster randomised trial. *International Journal of Nursing Studies*, 51:1321–331.

Noyes, K., Dick, A., and Holloway, R. (2004). Pramipexole v. levodopa as initial treatment for parkinson's disease: a randomized clinical-economic trial. *Medical Decision Making*, 24:472–485.

Nyman, M., Gustafsson, M., Langius-Eklof, A., Johansson, J., Norlin, R., and Hagberg, L. (2013). Intermittent versus indwelling urinary catheterisation in hip surgery patients: A randomised controlled trial with cost-effectiveness analysis. *International Journal of Nursing Studies*, 50:1589–1598.

Olmstead, T., Sindelar, J., and Petry, N. (2007). Cost-effectiveness of prize-based incentives for stimulant abusers in outpatient psychosocial treatment programs. *Drug and Alcohol Dependence*, 87:175–182.

Olsson, A., Casciano, R., Stern, L., and Svangren, P. (2004). A pharmacoeconomic evaluation of aggressive cholesterol lowering in sweden. *International Journal of Cardiology*, 96:51–57.

Oosternbrink, J., Rutten-van Molken, M., Al, M., Van Noord, J., and Vincken, W. (2004). One-year costeffectiveness of tiotropium versus ipratropium to treat chronic obstructive pulmonary disease. *European Respiratory Journal*, 23:241–249.

O'Reilly, J., Lowson, K., Young, J., Forster, A., Green, J., and Small, N. (2006). A cost effectiveness analysis within a randomised controlled trial of post-acute care of older people in a community hospital. *BMJ*, 333:228.

Parry, G., Cooper, C., Moore, J., Yadegarfar, G., Campbell, M., Esmonde, L., Morice, A., and Hutchcroft, B. (2012). Cognitive behavioural intervention for adults with anxiety complications of asthma: Prospective randomised trial. *Respiratory Medicine*, 106:802–810.

Patel, A., Knapp, M., Evans, A., Parez, I., and Kalra, L. (2004). Training care givers of stroke patients: economic evaluation. *BMJ*, 328:1102.

Peek, G., Mugford, M., Tiruvoipati, R., Wilson, A., Allen, E., Thalanany, M., Hibbert, C., Truesdale, A., Clemens, F., Cooper, N., Firmin, R., and Elbourne, D. (2009). Efficacy and economic assessment of conventional ventilatory support versus extracorporeal membrane oxygenation for severe adult respiratory failure (cesar): a multicentre randomised controlled trial. *Lancet*, 374:1351–1363.

Pennington, M., Grieve, R., Sekhon, J., Gregg, P., Black, N., and van der Meulen, J. (2013). Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis. *BMJ*, 346.

Petrou, S., Bischof, M., Bennett, C., Elbourne, D., Field, D., and McNally, H. (2006). Cost-effectiveness of neonatal extracorporeal membrane oxygenation based on 7-year results from the united kingdom collaborative ecmo trial. *Pediatrics*, 117:1640–1649.

Petrou, S., Dakin, H., Abangma, G., Benge, S., and Williamson, I. (2010). Cost–utility analysis of topical intranasal steroids for otitis media with effusion based on evidence from the gnometrial. *Value in Health*, 13:543–551.

Pickard, R., Starr, K., MacLennan, G., Kilonzo, M., Lam, T., Thomas, R., Burr, J., Norrie, J., McPherson, G., McDonald, A., Shearer, K., Gillies, K., Anson, K., Boachie, C., N'Dow, J., Burgess, N., Clark, T., Cameron, S., and McClinton, S. (2015). Use of drug therapy in the management of symptomatic ureteric stones in hospitalised adults: a multicentre, placebo-controlled, randomised controlled trial and cost-effectiveness analysis of a calcium channel blocker (nifedipine) and an alpha-blocker (tamsulosin) (the suspend trial). *Health Technology Assessment*, 19.

Prestmo, A., Hagen, G., Sletvold, O., Helbostad, J., Thingstad, P., Taraldsen, K., Lydersen, S., Halsteinli, V., Saltnes, T., Lamb, S., Johnsen, L., and Saltvedt, I. (2015). Comprehensive geriatric care for patients with hip fractures: a prospective, randomised, controlled trial. *Lancet*, 385:1623–1633.

Prinssen, M., Buskens, E., de Jong, S., Buth, J., Mackaay, A., Sambeek, M., and Blankensteijn, J. (2007). Costeffectiveness of conventional and endovascular repair of abdominal aortic aneurysms: results of a randomized trial. *Journal of Vascular Surgery*, 46:883–890.

Raftery, J., Yao, G., Murchie, P., Campbell, N., and Ritchie, L. (2005). Cost effectiveness of nurse led secondary prevention clinics for coronary heart disease in primary care: follow up of a randomised controlled trial. *BMJ*, 330:707.

Ratcliffe, J., Thomas, K., MacPherson, H., and Brazier, J. (2006). A randomised controlled trial of acupuncture care for persistent low back pain: cost effectiveness analysis. *BMJ*, 333:626.

Reed, S., Radeva, J., Glendenning, G., Saad, F., and Schulman, K. (2004). Cost-effectiveness of zoledronic acid for the prevention of skeletal complications in patients with prostate cancer. *The Journal of Urology*, 171:1537–1542.

Regier, D., Petrou, S., Henderson, J., Eddama, O., Patel, N., Strohm, B., Brocklehurst, P., and Edwards, AD.Azzopardi, D. (2010). Cost-effectiveness of therapeutic hypothermia to treat neonatal encephalopathyvhe. *Value in Health*, 13:695–702.

Revicki, D., Siddique, J., Frank, L., Chung, J., Green, B., Krupnick, J., Prasad, M., and Miranda, J. (2005). Costeffectiveness of evidence-based pharmacotherapy or cognitive behavior therapy compared with community referral for major depression in predominantly low-income minority women. *Archives of General Psychiatry*, 62:868–875.

Richardson, G., Bloor, K., Williamns, J., Russell, I., Durai, D., Cheung, W., Farrin, A., and Coulton, S. (2009). Cost effectiveness of nurse delivered endoscopy: findings from randomised multi-institution nurse endoscopy trial (minuet). *BMJ*, 338:270.

Richardson, G., Kennedy, A., Reeves, D., Bower, P., Lee, V., Middleton, E., Gardner, C., Gately, C., and Rogers, A. (2008). Cost effectiveness of the expert patients programme (epp) for patients with chronic conditions. *Journal of Epidemiology and Community Health*, 62:361–367.

Richardson, G., Sculpher, M., Kennedy, A., Nelson, E., Reeves, D., Roberts, C., Robinson, A., Rogers, A., and Thompson, D. (2006). Is self-care a cost-effective use of resources? evidence from a randomized trial in inflammatory bowel disease. *Journal of Health Services Research and Policy*, 11:225–230.

Rocca, H., Kaiser, C., Bernheim, A., Zellweger, M., Jeger, R., Buser, P., Osswald, S., and Pfisterer, M. (2003). Cost-effectiveness of drug-eluting stents in patients at high or low risk of major cardiac events in the basel stent kosteneffektivitsts trial (basket): an 18-month analysis. *The Lancet*, 370:1552–1559.

Roijen, L., Van Straten, A., Al, M., Rutten, F., and Donker, M. (2006). Cost-utility of brief psychological treatment for depression and anxiety. *British Journal of Psychiatry*, 188:323–329.

Rosenheck, R., Kasprow, W., Frisman, L., and Liu-Mares, W. (2003). Cost-effectiveness of supported housing for homeless persons with mental illness. *Archives of General Psychiatry*, 60:940–951.

Ryan, D., Price, D., Musgrave, S., Malhotra, S., Lee, A., Ayansina, D., Sheikh, A., Tarassenko, L., Pagliari, C., and Pinnock, H. (2012). Clinical and cost effectiveness of mobile phone supported self monitoring of asthma: multicentre randomised controlled trial. *BMJ*, 344.

Sabes-Figuera, R., McCrone, P., Hurley, M., King, M., Donaldson, A., and Risdale, L. (2012). Cost-effectiveness of zoledronic acid for the prevention of skeletal complications in patients with prostate cancer. *BMC Health Services Research*, 12:264.

Schweikert, B., Jacobi, E., Seitz, R., Cziske, R., Ehlert, A., Knab, J., and Leidl, R. (2006). Effectiveness and costeffectiveness of adding a cognitive behavioral treatment to the rehabilitation of chronic low back pain. *The Journal of Rheumatology*, 33:2519–2526.

Scott, D., Ibrahim, F., Farewell, V., O'Keeffe, A., Walker, D., Kelly, C., Birrell, F., Chakravarty, K., Maddison, P., Heslin, M., Patel, A., and Kingsley, G. (2015). Tumour necrosis factor inhibitors versus combination intensive therapy with conventional disease modifying anti-rheumatic drugs in established rheumatoid arthritis: Tacit non-inferiority randomised controlled trial. *BMJ*, 350.

Scott, J., Palmer, S., Paykel, E., Teasdale, J., and Hayhurst, H. (2003). Use of cognitive therapy for relapse prevention in chronic depression: cost-effectiveness study. *British Journal of Psychiatry*, 182:221–227.

Seivewright, H., Green, J., Salkovskis, P., Barrett, B., Nur, U., and Tyrer, P. (2008). Cognitive-behavioural therapy for health anxiety in a genitourinary medicine clinic: randomised controlled trial. *British Journal of Psychiatry*, 193:332–337.

Severens, J., Prins, J., van der Wilt, G., van der Meer, J., and Bleijenberg, G. (2004). Cost-effectiveness of cognitive behaviour therapy for patients with chronic fatigue syndrome. *The Quarterly Journal of Medicine*, 97:153–161.

Sevick, M., Napolitano, M., Papandonatos, G., Gordon, A., Reiser, L., and Marcus, B. (2007). Costeffectiveness of alternative approaches for motivating activity in sedentary adults: results of project stride. *Preventive Medicine*, 45:54–61.

Simon, J., Gray, A., Clarke, P., Wade, A., Neil, A., and Farmer, A. (2008). Cost effectiveness of self monitoring of blood glucose in patients with non-insulin treated type 2 diabetes: economic evaluation of data from the digem trial. *BMJ*, 336:1177–1180.

Smeets, R., Severens, J., Beelen, S., Vlaeyen, J., and Knottnerus, J. (2009). More is not always better: costeffectiveness analysis of combined, single behavioral and single physical rehabilitation programs for chronic low back pain. *European Journal of Pain*, 13:71–81.

Smit, F., Willemse, G., Koopmanschap, M., Onrust, S., Cuijpers, P., and Beekman, A. (2006). Cost-effectiveness of preventing depression in primary care patients: randomised trial. *British Journal of Psychiatry*, 188:330–336.

Smit, F., Willemse, G., Meulenbeek, P., Koopmanschap, M., van Balkom, A., Spinhoven, P., and Cuijpers, P. (2009). Preventing panic disorder: cost-effectiveness analysis alongside a pragmatic randomised trial. *Cost Effectiveness and Resource Allocation*, 7:8.

Stoddart, A., Hanley, J., Wild, S., Pagliari, C., Peterson, M., Lewis, S., Sheikh, A., Krishan, A., Padfield, P., and McKinstry, B. (2013). Telemonitoring-based service redesign for the management of uncontrolled hypertension (hits): cost and cost-effectiveness analysis of a randomised controlled trial. *BMJ*, 3.

Sullivan, S., Buxton, M., Andersson, L., Lamm, C., Liljas, B., Chen, Y., Pauwels, R., and Weiss, K. (2003). Cost-effectiveness analysis of early intervention with budesonide in mild persistent asthma. *Allergy and Clinical Immunology*, 112:1229–1236.

Taimela, S., Justen, S., Aronen, P., Sintonen, H., Laara, E., Malmivaara, A., Tiekso, J., and Aro, T. (2008). An occupational health intervention programme for workers at high risk for sickness absence. cost effectiveness analysis based on a randomised controlled trial. *Occupational and Environmental Medicine*, 65:242–248.

Taylor, A., Thompson, T., Greves, C., Taylor, R., Green, C., Warren, F., Kandiyali, R., Aveyard, P., Ayres, R., Byng, R., Campbell, J., Ussher, H., Michie, S., and West, R. (20). A pilot randomised trial to assess the methods and procedures for evaluating the clinical effectiveness and cost-effectiveness of exercise assisted reduction then stop (ears) among disadvantaged smokers. *Health Technology Assessment*, 18.

Teng, J., Mayo, N., Latimer, E., Hanley, J., Wood-Dauphinee, S., Cote, R., and Scott, S. (2003). Costs and caregiver consequences of early supported discharge for stroke patients. *Stroke*, 34:528–536.

Thompson, S., Ashton, H., Gao, L., and Scott, R. (2009). Screening men for abdominal aortic aneurysm: 10 year mortality and cost effectiveness results from the randomised multicentre aneurysm screening study. *BMJ*, 338.

Turner, D., Little, P., Raftery, J., Turner, S., Smith, H., Rumsby, K., and Mullee, M. (2010). Cost effectiveness of management strategies for urinary tract infections: results from randomised controlled trial. *BMJ*, 340.

Tyrer, P., Cooper, S., Salkovkis, Tyrer, H., Crawford, M., Byford, S., Dupont, S., Finnis, S., Green, J., McLaren, E., Murphy, D., Reid, S., Smith, G., Wang, D., Warwick, H., Petkova, H., and Barrett, B. (2014). Clinical and cost-eff ectiveness of cognitive behaviour therapy for health anxiety in medical patients: a multicentre randomised controlled trial. *Lancet*, 383:219–225.

Underwood, M., Lamb, S., Eldridge, S., Sheehan, B., Slowther, A., Spencer, A., Thorogood, M., Atherton, N., Bremner, S., Devine, A., Diaz-Ordaz, K., Ellard, D., Potter, R., Spanjers, K., and Taylor, S. (2013). Exercise for depression in care home residents: a randomised controlled trial with cost-effectiveness analysis (opera). *Health Technology Assessment*, 196:319–325.

Van Rossem, C., Spigt, M., Smit, E., Viechtbauer, W., Mijnheer, K., van Schayck, C., and Kotz, D. (2015). Combining intensive practice nurse counselling or brief general practitioner advice with varenicline for smoking cessation in primary care: Study protocol of a pragmatic randomized controlled trial. *Contemporary Clinical Trials*, 41:298–312.

Van Wijk, S., van Asselt, A., Rickli, H., Estlinbaum, W., Erne, P., Rickenbacher, P., Vuillomenet, A., Peter, M., Pfisterer, M., and Brunner-La Rocca, H. (2013). Cost-effectiveness of n-terminal pro-b-type natriuretic-guided therapy in elderly heart failure patients. *JACC*, 1:64–71.

Viksveen, P. and Relton, C. (2014). Depression treated by homeopaths: a study protocol for a pragmatic cohort multiple randomised controlled trial. *Homeopathy*, 103:147–152.

Wagner, T., Hattler, B., Bishawi, M., Baltz, J., Collins, J., Quin, J., Grover, F., and Shroyer, A. (2013). On-pump versus off-pump coronary artery bypass surgery: Cost-effectiveness analysis alongside a multisite trial. *Ann Thorac Surg*, 96:770–777.

Wake, M., Baur, L., Gerner, B., Gibbons, K., Gold, L., Gunn, J., Levickis, P., McCallum, Z., Naughton, G., Sanci, L., and Ukoumunne, O. (2009). Outcomes and costs of primary care surveillance and intervention for overweight or obese children: the leap 2 randomised controlled trial. *Preventive Medicine*, 339.

Ward, S., Wang, K., Serlin, R., Peterson, S., and Murray, L. (2009). A randomized trial of a tailored barriers intervention for cancer information service (cis) callers in pain. *Pain*, 144:49–56.

Witt, C., Jena, S., Selim, D., Brinkhaus, B., Reinhold, T., Wruck, K., Liecker, B., Linde, K., Wegscheider, K., and Willich, S. (2006). Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *American Journal of Epidemiology*, 164:487–496.

Witt, C., Reinhold, T., Brinkhaus, B., Roll, S., Jena, S., and Willich, S. (2008). Acupuncture in patients with dysmenorrhea: a randomized study on clinical effectiveness and cost-effectiveness in usual care. *American Journal of Obstetrics and Gynecology*, 198:166.

Wolfs, C., Dirksen, C., Kessels, A., Severens, J., and Verhey, F. (2009). Economic evaluation of an integrated diagnostic approach for psychogeriatric patients: results of a randomized controlled trial. *Archives of General Psychiatry*, 66:313–323.

Wonderling, D., Vickers, A., Grieve, R., and McCarney, R. (2004). Cost effectiveness analysis of a randomised trial of acupuncture for chronic headache in primary care. *BMJ*, 328:747.

Wu, E., Birnbaum, H., Mareva, M., Le, T., Robinson, R., Rosen, A., and Gelwicks, S. (2006). Costeffectiveness of duloxetine versus routine treatment for u.s. patients with diabetic peripheral neuropathic pain. *The Journal of Pain*, 7:399–407.

Yardley, L., Barker, F., Muller, I., Turner, D., Kirby, S., Mullee, M., Morris, A., and Little, P. (2012). Clinical and cost effectiveness of booklet based vestibular rehabilitation for chronic dizziness in primary care: single blind, parallel group, pragmatic, randomised controlled trial. *BMJ*, 344.

Yoo, S., Nyman, J., Cheville, A., and Kroenke, K. (2014). Cost effectiveness of telecare management for pain and depression in patients with cancer: results from a randomized trial. *General Hospital Psychiatry*, 36:599–606.

Zwanziger, J., Hall, W., Dick, A., Zhao, H., Mushlin, A., Hahn, R., Wang, H., Andrews, M., Mooney, C., Wang, H., and Moss, A. (2006). The cost effectiveness of implantable cardioverter-defibrillators: results from the multicenter automatic defibrillator implantation trial (madit)-ii. *Journal of the American College of Cardiology*, 47:2310–2318.

Zwerink, M., Am Kerstjens, H., van der Palen, J., van der Valk, P., Brusse-Keizer, M., Zielhuis, G., and Effing, T. (2015). (cost-)effectiveness of self-treatment of exacerbations in patients with copd: 2 years follow-up of a rct. *Respirology*.

Bibliography

Agbla, S. and Diaz-Ordaz, K. (2018). Reporting non-adherence in cluster randomised trials: A systematic review. *Clinical Trials*, 15:294–304.

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. (2013). Package "BCEs0". `www.statistica.it/gianluca`.

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

Baio, G. (2017). Statistical modeling for health economic evaluations. *Annual Review of Statistics and Its Application*, 5:289–309.

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. (2004). Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research and Policy*, 9:197–204.

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.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H. Dai, B., Scheipl, F., Grothendieck, G., Green, P., and Fox, J. (2019). Package "lme4". `https://cran.r-project.org/web/packages/lme4/`.

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

Briggs, A. (1999). A bayesian approach to stochastic cost-effectiveness analysis. *Health Economics*, 8:257–261.

Briggs, A., Clark, T., Wolstenholme, J., and Clarke, P. (2003). Missing ..., presumed at random: cost-analysis of incomplete data. *Health Economics*, 12:377–392.

Briggs, A. and Gray, A. (1998). Power and sample size calculations for stochastic cost-effectiveness analysis. *Medical Decision Making*, 18:81–92.

Briggs, A. and Gray, A. (1999). Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technology Assessment*, 3:1–134.

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

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Buck, S. (1960). A method of estimation for missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, 22:302–306.

Campbell, M., Fayers, P., and Grimshaw, J. (2005). Determinants of the intracluster correlation coefficient in cluster randomised trials: the case of implementation research. *Clinical Trials*, 2:99–107.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32.

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

Celeux, G., Forbes, S., Robert, C., and Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1:651–674.

Chambless, L. and Roeback, J. (1993). Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation. *Statistics in Medicine*, 12:1213–1237.

Chan, S., Macaskil, P., Irwig, L., and Walter, S. (2004). Adjustment for baseline measurement error in randomized controlled trials induces bias. *Controlled Clinical Trials*, 25:408–416.

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., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P., and Sculpher, M. (2015). Methods for the estimation of the national institute for health and care excellence cost-effectiveness threshold. *Health Technology Assessment*, 19.

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.

Cooper, N., Sutton, A., Mugford, M., and Abrams, K. (2003). Use of bayesian markov chain monte carlo methods to model cost-of-illness data. *Medical Decision Making*, 23:38–53.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK.

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall, New York.

Dempster, A. (1973). Statistics and computing. *The Statistician*, 7:247–252.

Dempster, A., Laird, L., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Diaz-Ordaz, K., Franchini, J., and Grieve, R. (2018). Methods for estimating complier average causal effects for cost-effectiveness analysis. *Journal of the Royal Statistical Society: Series A*, 181:277–297.

Diaz-Ordaz, K., Kenward, M., Cohen, A., Coleman, C., and Eldridge, S. (2014a). Are missing data adequately handled in cluster randomised trials? a systematic review and guidelines. *Clinical Trials*, 11:590–600.

Diaz-Ordaz, K., Kenward, M., Gomes, M., and Grieve, R. (2016). Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Statistics in Medicine*, 35:3482–3496.

Diaz-Ordaz, K., Kenward, M., and Grieve, R. (2014b). 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.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society*, 43:49–93.

Dodd, S., White, I., and Williamson, P. (2012). Non-adherence to treatment protocol in published randomised controlled trials: a review. *Trials*, 13:84.

Dolan, P. and Gutex, C. (1995). Time preference, duration and health state valuations. *Heath Economics*, 4:289–299.

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.

Eekhout, I., de Boer, M., Twisk, J., de Vet, H., and Heymans, M. (2012). A systematic review of how they are reported and handled. *Epidemiology*, 23:729–732.

Eldridge, S., Kerry, S., and Torgerson, D. (2009). Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*, 339.

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.

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.

Gaskins, J., Daniels, M., and Marcus, B. (2016). Bayesian methods for nonignorable dropout in joint models in smoking cessation studies. *Journal of the American Statistical Association*, 111:1454–1465.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis - 3rd edition*. Chapman and Hall, New York, NY.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis - 2nd edition*. Chapman and Hall, New York, NY.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Stat. Comput.*, 24:997–1016.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741.

Gilks, W., Thomas, A., and Spiegelhalter, D. (1994). A language and program for complex bayesian modelling. *The Statistician*, 43:169–177.

Glick, H. (2011). Sample size and power for cost-effectiveness analysis (part 1). *Pharmacoeconomics*, 29:190–198.

Gomes, M., Diaz-Ordaz, K., Grieve, R., and Kenward, M. (2013). Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: An application to cluster randomised trials. *Medical Decision Making*, 33:1051–1063.

Gomes, M., Grieve, R., and Nixon, R. (2011). Statistical methods for cost-effectiveness analyses that use data from cluster randomized trials: A systematic review and checklist for critical appraisal. *Medical Decision Making*, 32:209–220.

Gomes, R., Grieve, R., Nixon, R., and Edmunds, W. (2012a). Statistical methods for cost-effectiveness analyses that use data from cluster randomized trials. *Medical Decision Making*, 32:209–220.

Gomes, R., Grieve, R., Nixon, R., NG, E., Carpenter, J., and Thompson, S. (2012b). Methods for covariate adjustment in cost-effectiveness analysis that use cluster randomised trials. *Health Economics*, 21:1101–1118.

Gomes, R., Ng, E., Grieve, R., Nixon, R., NG, E., Carpenter, J., and Thompson, S. (2012c). Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Medical Decision Making*, 32:350–361.

Green, W. (2003). *Econometric Analysis*. Prentice Hall:Upper Saddle River, UK.

Grieve, R., Nixon, R., Simon, G., and Thompson, S. (2010). Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. *Medical Decision Making*, 30:163–175.

Grieve, R., Nixon, R., Thompson, S., and Normand, C. (2005). Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Economics*, 14:185–196.

Grigore, B., Peters, J., Hyde, C., and Stein, K. (2016). A comparison of two methods for expert elicitation in health technology assessments. *BMC Medical Research Methodology*, 16.

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.

Hassiotis, A., Poppe, M., Strydom, A., Vickerstaff, V., Hall, I., Crabtree, J., Omar, R., King, M., Hunter, R., Biswas, A., Cooper, V., Howie, W., and Crawford, M. (2018). Clinical outcomes of staff training in positive behaviour support to reduce challenging behaviour in adults with intellectual disability: cluster randomised controlled trial. *The British Journal of Psychiatry*, 212:161–168.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Henningsen, A. and Hamann, J. (2018). Package "systemfit". `https://cran.r-project.org/web/packages/systemfit/`.

Hoch, J., Briggs, A., and Willan, A. (2002). Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics*, 11:415–430.

Hogan, J. and Laird, N. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, 16:239–257.

Huelin, R., Iheanacho, I., Payne, K., and Sandman, K. (2015). What's in a Name? Systematic and Non-Systematic Literature Reviews, and Why the Distinction Matters. https://www.evidera.com/wp-content/uploads/2015/06/Whats-in-a-Name-Systematic-and-Non-Systematic-Literature-Reviews-and-Why-the-Distinction-Matters.pdf.

Hughes, D., Charles, J., Dawoud, D., Edwards, R., Holmes, E., Jones, C., Parham, P., Plumpton, C., Ridyard, C., Lloyd-Williams, H., Wood, E., and Yeo, S. (2016). Conducting economic evaluations alongside randomised trials: Current methodological issues and novel approaches. *PharmacoEconomics*, 34:447–461.

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.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley and Sons, New York, NY.

Jackson, C. (2008). Displaying uncertainty with shading. *The American Statistician*, 62:340–347.

Kenward, M. and Molenberghs, G. (2010). Last observation carry-forward: a crystal ball? *J Biopharm Stat*, 19:872–888.

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.

Kroese, P., Taimre, T., and Zdravko, Z. (2013). *Handbook of Monte Carlo Methods*. John Wiley and Sons.

Laska, E., Meisner, M., and C, S. (1999). Power and sample size in cost-effectiveness analysis. *Medical Decision Making*, 8:203–211.

Lee, P. (2012). *Bayesian statistics: an introduction*. John Wiley and Sons.

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

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., and Carpenter, J. (2018b). Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: A tutorial. *PharmacoEconomics*, pages 1–13.

Linero, A. and Daniels, M. (2015). A flexible Bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial. *Journal of the American Statistical Association*, 110:45–55.

Linero, A. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika*, 104:327–341.

Linero, A. and Daniels, M. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, 33:198–213.

Little, R., D'Agostino, R., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., Neaton, J., Rotnitzky, A., Scharfstein, D., Shih, W., Siegel, J., , and Stern, H. (2010). The prevention and treatment of missing data in clinical trials. panel on handling missing data in clinical trials. *Committee on National Statistics, Division of Behavioral and Social Sciences and Education*.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134.

Little, R. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471–483.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data, Second Edition*. John Wiley and Sons, New York.

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.

Marshall, A., Billingham, L., and Bryan, S. (2009). Can we afford to ignore missing data in cost-effectiveness analyses? *Eur J Health Econ*, 10:1–3.

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.

Mason, A., Richardson, S., and Best, N. (2012a). Two-pronged strategy for using dic to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7:109–146.

Mason, A., Richardson, S., Plewis, I., and Best, N. (2012b). Strategy for modelling nonrandom missing data mechanisms in observational studies using bayesian methods. *Journal of Official Statistics*, 28:279–302.

Meltzer, M. (2001). Introduction to health economics for physicians. *Lancet*, 358:993–998.

Meng, X. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22:1142–1160.

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.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, T., and Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4.

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

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*. John Wiley and Sons, Chichester, UK.

Molenberghs, G., Kenward, M., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random drop-out. *Biometrika*, 84:33–44.

Mooney, G. (1989). Qalys: are they enough? a health economist's perspective. *Journal of Medical Ethics*, 15:148–152.

Neumann, P. and Greenberg, D. (2009). Is the united states ready for qalys? *Health Affairs*, 28:1366–1371.

Ng, E., Diaz-Ordaz, K., Grieve, R., Nixon, R., Thompson, S., and Carpenter, J. (2016). Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice. *Statistical Methods in Medical Research*, 25:2036–2052.

Ng, E., Grieve, R., and Carpenter, J. (2013). Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data. *Stata J*, 13:141–164.

NICE (2008). *Guide to the Methods of Technological Appraisal*. NICE, London.

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

Nixon, R. and Thompson, S. (2004). Parametric modelling of cost data in medical studies. *Statistics in Medicine*, 23:1311–1331.

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.

OECD (2015). *Fiscal Sustainability of Health Systems: Bridging Health and Finance Perspectives*. OECD Publishing, Paris.

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.

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

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

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

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

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.

Rascati, K., Smith, L., and Neilands, T. (2001). Dealing with skewed data: An example using asthma-related costs of medicaid clients. *Health Economics*, 23:481–498.

Robins, J. and Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable-missing data. *Statistics in Medicine*, 16:39–56.

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

Rubin, D. (1988). An overview of multiple imputation. *Proceedings of the Survey Research Section American Statistical Association*, pages 79–84.

Rubin, D. (1996). Multiple imputation after 18 years. *Journal of the American Statistical Association*, 91:473–489.

Sadinle, M. and Reiter, J. (2017). Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika*, 104:207–220.

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

Schafer, J. and Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177.

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

Shao, J. and Zhong, B. (2010). Last observation carry-forward and last observation analysis. *Statistics in Medicine*, 22:2429–2441.

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.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64:583–639.

Stevens, J. and O'Hagan, A. (2002). Incorporation of genuine prior information in cost-effectiveness analysis of clinical trial data. *International Journal of Technology Assessment in Health Care*, 18:782–790.

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

Su, Y. and Yajima, M. (2015). Package "R2jags". `https://cran.r-project.org/web/packages/R2jags/`.

Tchetgen Tchetgen, E., Wang, L., and Sun, B. (2016). Discrete choice models for non-monotone nonignorable missing data: Identification and inference. `https://arxiv.org/pdf/1607.02631.pdf`.

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

Tooze, J., Grunwald, G., and Jones, K. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, 211:341–355.

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 Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67.

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.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under non-ignorable non-monotone non-response. *Biometrika*, 94:841–860.

Vazquez Polo, F., Hernandez, M., and Lopez-Valcarcel, B. (2005). Using covariates to reduce uncertainty in the economic evaluation of clinical trial data. *Health Economics*, 14:545–557.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross validation and waic. *Stat. Comput.*, 27:1413–1432.

Wang, C. and Daniels, M. (2011). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. *Biometrics*, 67:810–818.

White, I. and Carlin, J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29:2920–2931.

Willan, A. (2001). Analysis, sample size, and power for estimating incremental net health benefit from clinical trial data. *Control Clinical Trial*, 22:228–237.

Willan, A. and Briggs, A. (2006). *Statistical Analysis of Cost-Effectiveness Data*. Wiley, Chichester, UK.

Willan, A., Briggs, A., and Hock, J. (2004). 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.

Wu, M. and Carroll, R. (1994). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44:175–188.

Yanez, N., Krnomal, N., and Shemanski, L. (1998). The effects of measurement error in response variables and tests of association of explanatory variables in change models. *Statistics in Medicine*, 17:2597–2606.

Zhang, Z., Peluso, M., Gross, C., Viscoli, C., and Kernan, W. (2014). Adherence reporting in randomized controlled trials. *Clinical Trials*, 11:195–204.