The
University
Of
Sheffield.

Access to Electronic Thesis

Author: Mark Strong

Thesis title: Managing Structural Uncertainty in Health Economic Decision Models

Qualification: PhD

This electronic thesis is protected by the Copyright, Designs and Patents Act 1988. No reproduction is permitted without consent of the author. It is also protected by the Creative Commons Licence allowing Attributions-Non-commercial-No derivatives.

If this electronic thesis has been edited by the author it will be indicated as such on the title page and in the text.

The
University
Of
Sheffield.

Managing Structural Uncertainty in Health Economic Decision Models

Mark Strong

Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

School of Health and Related Research

March 2012

Acknowledgements

Firstly, I would like to thank Jeremy Oakley for his excellent supervision throughout my PhD study. I would also like to thank Jim Chilcott, Ravi Maheswaran and Matt Stevenson who have played important roles in supervising my work, and Mike Campbell who kindly acted as examiner for my upgrade from MPhil to PhD.

I have benefited greatly over the past three years from the many interesting and challenging discussions I have had with colleagues in ScHARR, and I would particularly like to thank Alan Brennan, Ben van Hout, Hazel Squires, John Stevens, Paul Tappenden and Allan Wailoo.

This PhD was funded by the Medical Research Council as part of a Health Services Research / Health of the Public Research Training Fellowship (grant number G0601721).

Summary

Health economic models are representations of judgements about the relationships between the model's input parameters and the costs and health effects that the model aims to predict. We recognise that we can rarely define with certainty a 'true' model for a particular decision problem. Building an 'incorrect' model will result in an uncertain prediction error, which we denote 'structural uncertainty'. The absence of observations on the total costs and health effects under each decision option limits the use of data driven approaches to managing structural uncertainty, such as model averaging.

We therefore propose a discrepancy based approach in which we make judgements about structural error at the *sub-function* level within the model and introduce a series of terms to 'correct' the errors. This is deemed to be easier than making meaningful statements about the error at the level of the model output. The specification of discrepancy terms *within* the model also allows us to use sensitivity analysis methods to determine the relative importance of the different structural uncertainties in driving output and decision uncertainty.

Following the computation of either the main effect index or the partial expected value of perfect information for each discrepancy term, we can review the structure of those parts of the model where structural uncertainty is an important source of model output or decision uncertainty. We interpret the overall expected value of perfect information for all the discrepancy terms as an upper bound on the expected value of model improvement (EVMI).

We illustrate the sub-function discrepancy method in two case studies: a simple decision tree, and a more complex Markov model. Finally, we propose an efficient method for computing the main effect index and the partial expected value of perfect information when inputs and/or discrepancies are correlated.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis overview 3

2 Economic Evaluation in Health **5**

2.1 Introduction 5

2.2 The problem of allocating scarce resources 6

2.3 Decision Theory 7

2.4 Utility functions 10

2.5 What are the outcomes under each decision option? 11

2.6 Measuring resource inputs 12

2.7 Measuring health outputs 13

2.8 Types of economic evaluation analysis 16

2.9 Discounting 19

2.10 Mathematical modelling in economic evaluation 19

2.11 How do health economic evaluation models work? 22

2.12 Categorising health economic decision models 23

2.13 The general cost effectiveness model 30

2.14 Sources of error 32

2.15 The health care decision context and NICE 35

2.16 Conclusion 37

3 Managing Uncertainty in Computer Models **38**

3.1 Introduction 38

3.2 Computer models 39
3.3 Uncertainty 47
3.4 Managing uncertainty in computer models 50
3.5 Constructing a single framework 70
3.6 Conclusion 71

4 Elicitation 72
4.1 Introduction 72
4.2 What is elicitation? 72
4.3 How does elicitation work? 73
4.4 Elicitation in health services and medical research 80
4.5 Conclusion 84

5 Case study 1 86
5.1 Introduction 86
5.2 Base case model 87
5.3 A discrepancy approach 94
5.4 Sub-function discrepancy assessment 99
5.5 Discrepancy analysis results 110
5.6 Conclusion 118

6 Case study 2 119
6.1 Introduction 119
6.2 Case study model 120
6.3 Discrepancy analysis 124
6.4 Results 137
6.5 Could we have added fewer discrepancy terms? 141
6.6 Conclusion 142

7 Efficient Computation 144
7.1 Introduction 144
7.2 Method for partial EVPI 146
7.3 Method for main effect index 150

7.4 Theoretical justification 153
7.5 Sample size considerations 155
7.6 Case study 158
7.7 Conclusion 162

8 Conclusion 167

8.1 Summary 167
8.2 Model complexity and parsimony 170
8.3 How might this work in practice? 171

Chapter 1

Introduction

1.1 Motivation

Mathematical models are routinely used to aid healthcare resource allocation decisions, with the majority of such models falling into the broad category of 'cost-effectiveness' models. These models aim to predict the costs and health consequences associated with a range of competing decisions. Although decision theory tells us that we only need to know the expectation of the net benefit of the resource costs and health effects under each decision option in order to make the optimum choice, we will usually also want to quantify our uncertainty about the costs and health effects. If we are sufficiently uncertain then we may wish to gather more information before embarking on a decision that is costly to reverse, or committing resources that are potentially unrecoverable (McKenna and Claxton, 2011). A cost-effectiveness model will be most helpful then if it allows us to properly specify all of our uncertainty about the quantity we wish to predict.

There are two primary sources of uncertainty when basing statements about costs and health effects on a mathematical model: uncertainty about the model *inputs* and uncertainty about the model *structure*. Methods for quantifying the first source of uncertainty are well established, but understanding and managing the second is rather more difficult to since it involves making judgements about a model's ability to faithfully represent a (possibly highly complex) real life decision problem.

The problem of uncertainty in deterministic mathematical model (or 'computer model') predictions is common to many disciplines, and has been the subject of much research (see for example Santner et al., 2003; Bayarri et al., 2009). Methods for analysing the effects of *input* uncertainty are particularly well developed (Oakley and O'Hagan, 2004; Saltelli et al., 2008), however the problem of *structural* uncertainty has received much less attention. There are two broad approaches to quantifying structural uncertainty: via *model averaging* (for example Draper, 1995; Kass and Raftery, 1995; Kadane and Lazar, 2004; Bojke et al., 2009; Jackson et al., 2009) and via specification of *model discrepancy* (Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2009).

There are, however, difficulties with both methods when applied in the context of health economic models. There is usually an absence of direct observations on the costs and health effects under the set of decision options being considered in the economic evaluation. This means that in the model averaging approach model weights cannot be based on some likelihood based measure, and in the discrepancy approach there are no data that will directly inform beliefs about the difference between the model output and the expected population costs and benefits. Elicitation could in theory provide a means of specifying the necessary weights or distributions, but it is not clear whether this is feasible in practice. How would a modeller (or a decision maker) weight a set of competing computer models? How would he or she make meaningful judgements about how wrong the predictions of a given model might be?

The motivation then for this research is to develop a method for managing structural uncertainty in health economic models that does not rely on past observations on the output quantities (the costs and health effects) that are predicted by the model. We use the term 'managing uncertainty' to include the goal of understanding better the sources of structural uncertainty such that it can be reduced through model revision, as well as the goal of formally quantifying the uncertainty. We also consider the problem of deciding when it is worth spending resources to improve a model. The output of a computer model is not a free good. Considerable resources may be required to gather the necessary evidence to inform the input distributions and model structure, to program, debug and validate the model, and to report the results in a meaningful fashion. Running a complex model may also require expensive computer software and/or hardware, and require long run times. Given a decision where we have considerable uncertainty, how much is a computer model worth? Given a computer model that we know is not perfect, how much should we pay to make it better?

Throughout the thesis we adopt a subjectivist Bayesian perspective since this provides a sensible and coherent framework within which to think about decision making under uncertainty (de Finetti, 1974; Dowie, 2006; Smith, 2010). In health care resource management, a decision maker will choose from competing options based on *her own subjective* uncertainty specification of the net benefits of the various options open to her. This specification may well be informed by the results of a mathematical model, but ultimately, it is the decision maker's own personal probabilities that will guide the decision.

1.2 Thesis overview

We begin in chapter 2 by introducing some basic decision theory that underpins the economic evaluation methods that are used to inform resource allocation decisions in health care. We go on to discuss health economic evaluation methods and the use of computer models, and end the chapter by thinking about how models used for economic evaluation might be 'wrong'. Chapter 3 begins with a general introduction to computer models, followed by a discussion of the management of the different sources of uncertainty that are evident when we wish to make predictions based on a computer model output. We recognise that we will often need to quantify aspects of health economic model uncertainty in the absence of data to inform the uncertainty distribution, and this motivates the use of formal elicitation methods. We introduce some basic ideas about elicitation in chapter 4 and review the use of elicitation in health services research and health economic evaluation.

In chapters 5 and 6 we consider the problem of health economic model uncertainty and introduce a novel approach to managing this source of uncertainty. We illustrate the method in two case studies, the first concerning a simple decision tree model, and the second a more complex Markov model. Sensitivity analysis techniques are used to help us understand the implications of our uncertainty about model structure, and chapter 7 contains the description of an efficient method for calculating those sensitivity measures (the main effect index and the partial expected value of perfect information) when model inputs are correlated. The thesis ends with a discussion in chapter 8 of the main themes along with implications for future research and practice.

Chapter 2

Economic Evaluation in Health

2.1 Introduction

In this chapter we introduce the basic decision theory that underpins the economic evaluation framework used to inform health care resource allocation decisions. We see that economic evaluation methods seek to generate predictions about the costs and health effects that will follow some set of competing decision options open to the decision maker. For such predictions to be *evidence based* we must have some mechanism for linking observations of the world that have taken place in the past, to the predictions of costs and health effects that will take place in the future under each decision option. It is this linking of past observations to future predictions that motivates the building of a *computer model.*

We describe some types of computer model that are commonly found in health economic analysis, and show that these are all particular cases of a general form of model. Given this general model, we are able to identify a number of important basic criteria that must be fulfilled in order to avoid error in the model predictions. Towards the end of the chapter we begin to unpick the meaning of 'structure' in the context of health economic decision models, and look ahead to methods for quantifying uncertainty in this structure.

2.2 The problem of allocating scarce resources

Healthcare resources are scarce. Demand for consultations, treatments, services and health care programmes exceeds our ability to supply them within the resources committed by society for this purpose. Scarcity necessitates resource allocation choices (Weinstein and Stason, 1977).

We assume the existence of a single decision maker, whose responsibility it is to allocate some set of scarce resources for the benefit of some population. Each choice open to the decision maker will result in an outcome: a set of health (and other) consequences, along with some degree of resource usage. Taken together, the consequences and the size and nature of the resources used to achieve those consequences has some net value to the decision maker. Following the description given in Smith (2010) our decision maker wishes to act in a *logical*, *coherent* and *honest* fashion; has *responsibility for* and the *authority to enact* the decision; and is answerable to some higher 'auditor' for her actions. We recognise that in reality decisions are rarely taken by single individuals acting in isolation, and that in many circumstances the responsibility for making a decision belongs very explicitly *not* to an individual, but instead to a group or committee (the committees of NICE[1] being a good example).

So, to give an example of a health resource allocation problem, a decision maker may be faced with the choice to either recommend or not recommend the use of a new drug treatment for some disease. Using the drug will result in costs (i.e. an allocation of the scarce resources) and consequences, which include (but are not necessarily limited to) the health outcomes for those who will take the drug. By not allowing the drug to be used, other costs (perhaps those of an existing drug treatment) and consequences (the health outcomes related to the existing drug) will result. Given these two options, how should a decision maker choose between them?

Formal methods for economic evaluation to inform health care resource allocation decisions have been routinely applied for several decades in many developed countries, and there are now a considerable number of standard texts that de-

[1] The National Institute for Health and Clinical Excellence (`http://www.nice.org.uk/`).

scribe these methods (e.g. Drummond et al., 2005; Gold et al., 1996; Neumann, 2005). Underpinning the economic evaluation approach is the theory of decision making under uncertainty (Raiffa, 1968), and it is from this perspective that we will review methods in the remainder of this chapter, and indeed this perspective underpins the thesis as a whole.

2.3 Decision Theory

For any decision problem, the decision maker is faced with a range of competing decision options. Each of these options will lead to a set of outcomes. The decision maker is able to express preferences for the outcomes through a *utility function* that describes the value to her of any particular set of outcomes. We write the set of possible decisions as $d = 1, \dots, D$, and we denote the (vector of) relevant outcomes following decision d to be $Z(d)$. The decision maker has a utility function, $U\{Z(d)\}$, either implicit or explicit, and a desire to choose the option d^* that maximises utility,

$$d^* = \arg\max_d U\{Z(d)\}. \tag{2.1}$$

Decision makers are, however, faced with a problem. The outcomes that will occur under each decision option are almost always unknown before the decision is made. If this is the case then we can say that the decision maker has *uncertainty* about the outcomes, and this uncertainty may (but not always) result in uncertainty about which of the choices that are available has the greatest utility.

We now write the vector of outcomes as a function not only of the decision, but of a vector of unknowns, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Our optimal decision is now that which maximises *expected* utility,

$$d^* = \arg\max_d E[U\{Z(\boldsymbol{\theta}, d)\}] = \arg\max_d \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} U\{Z(\boldsymbol{\theta}, d)\} p(\boldsymbol{\theta}|d) d\boldsymbol{\theta}. \tag{2.2}$$

We note in passing here that $E_{\boldsymbol{\theta}}[U\{Z(\boldsymbol{\theta}, d)\}] = U[Z\{E(\boldsymbol{\theta}), d\}]$ if and only if $U\{Z(\boldsymbol{\theta}, d)\}$ is linear in $\boldsymbol{\theta}$ (or multilinear in $U\{Z(\boldsymbol{\theta}, d)\}$ with independence in the components of $\boldsymbol{\theta}$). If not, then we must somehow evaluate the integral on the RHS of (2.2). If we have used a computer model to define $U\{Z(\boldsymbol{\theta}, d)\}$ then we may find that there is no closed form solution to this integral. If this is the case we typically use Monte Carlo integration. In health economic evaluation this approach is called 'probabilistic sensitivity analysis' (Griffin et al., 2006). We return to this topic in Chapter 3.

To show why the optimal decision is that which maximises *expected* utility we first define two reference states, S_0 and S_1, that describe our least preferred and most preferred outcomes with utilities $U(S_0) = u_0$ and $U(S_1) = u_1$. From the definition of utility we have

$$U\{Z(\boldsymbol{\theta}, d)\} = qu_1 + (1 - q)u_0, \tag{2.3}$$

for some value q. Equation (2.3) implies that we have equal preference for $Z(\boldsymbol{\theta}, d)$ and a state S_q in which we will move to state S_1 with probability q, and to state S_0 with probability $1 - q$.

If we were to learn the value of $\boldsymbol{\theta}$, then our conditional utility would be

$$U\{Z(\boldsymbol{\theta}, d)|\boldsymbol{\theta}\} = P_\theta u_1 + (1 - P_\theta)u_0, \tag{2.4}$$

for some probability P_θ, following the same argument as above. We now write q in equation (2.3) in terms of an integral over $\boldsymbol{\theta}$ as follows:

$$\begin{aligned} q &= P(\text{move to state } S_1) && (2.5)\\ &= \int_{\boldsymbol{\theta}\in\Theta} P(\text{move to state } S_1|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} && (2.6)\\ &= \int_{\boldsymbol{\theta}\in\Theta} P_\theta p(\boldsymbol{\theta})d\boldsymbol{\theta}. && (2.7)\end{aligned}$$

Our unconditional utility is therefore

$$
\begin{aligned}
U\{Z(\boldsymbol{\theta},d)\} &= qu_1 + (1-q)u_0 && (2.8)\\
&= \left\{\int_{\boldsymbol{\theta}\in\Theta} P_\theta p(\boldsymbol{\theta})d\boldsymbol{\theta}\right\}u_1 + \left\{1 - \int_{\boldsymbol{\theta}\in\Theta} P_\theta p(\boldsymbol{\theta})d\boldsymbol{\theta}\right\}u_0 && (2.9)\\
&= \int_{\boldsymbol{\theta}\in\Theta} \{P_\theta u_1 + (1-P_\theta)u_0\}\, p(\boldsymbol{\theta})d\boldsymbol{\theta} && (2.10)\\
&= E_{\boldsymbol{\theta}}[U\{Z(\boldsymbol{\theta},d)|\boldsymbol{\theta}\}]. && (2.11)
\end{aligned}
$$

Therefore the utility of the outcome $Z(\boldsymbol{\theta},d)$ for uncertain $\boldsymbol{\theta}$, is its expectation $E_{\boldsymbol{\theta}}[U\{Z(\boldsymbol{\theta},d)\}]$. In order to maximise utility under uncertainty, the decision maker simply chooses the decision option with maximum expected utility.

In an important sense the decision maker acts *rationally* by choosing d^* in this way to be the optimal choice: expected gains in utility are maximised and conversely expected losses are minimised. This is shown by Raiffa (1968) in his seminal text on decision theory, and comprehensively discussed in the explicitly *subjectivist* introduction to decision analysis by Smith (2010).

We note that finding the optimal decision does *not* depend on the ‘statistical significance’ of any measure of difference between the utilities of the different decisions (or indeed the Bayesian posterior probability of equivalence), and as such there is an ‘irrelevance of inference’ (Claxton, 1999). If we wish to make the decision *now*, only the expectations (equation 2.2) are required to rationally choose between options.

If the uncertainty in costs and consequences is such that there is uncertainty as to which decision option has greatest net utility, then by making a choice the decision maker is taking a risk. The choice may be the wrong choice. Choosing a different option may have resulted in greater utility. The decision maker may therefore also wish to quantify the probability that d^* is the decision option that maximises utility, given uncertainty about $\boldsymbol{\theta}$,

$$P[U\{Z(\boldsymbol{\theta},d^*)\} \geq U\{Z(\boldsymbol{\theta},d)\}\ \forall d]. \tag{2.12}$$

If new evidence $\boldsymbol{\theta}'$ becomes available at some point in the future we may find that the optimal decision d^* is no longer optimal, i.e. that

$$\arg\max_d E[U\{Z(\boldsymbol{\theta}', d)\}] \neq \arg\max_d E[U\{Z(\boldsymbol{\theta}, d)\}]. \tag{2.13}$$

This is of little concern if a decision maker can costlessly switch between decision options each time new evidence implies an optimal choice different to that currently adopted. However, this is usually not the case in health care. Adopting or reimbursing a new intervention or service almost always implies certain 'sunk' costs; irrecoverable costs associated with the change in practice (Eckermann and Willan, 2008). Identifying one of the decision options as optimal may also lead to another important irreversibility; a trial or study that could gather additional evidence to support the decision may be deemed unnecessary or unethical (Griffin et al., 2011). If a decision maker anticipates that the adoption of the optimal decision option (under current information, $\boldsymbol{\theta}$) will be associated with sunk costs and/or other irreversibilities, then she may wish to quantify the value of reducing uncertainty about $\boldsymbol{\theta}$ first. We will consider methods for computing the value of information in detail in our discussion of computer model uncertainty in chapter 3.

2.4 Utility functions

For brevity we denote the uncertain outcomes under decision d as the random variable vector $\boldsymbol{Z}_d$ where $\boldsymbol{Z}_d = Z(\boldsymbol{\theta}, d)$, and the vector of outcomes under all decisions as $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_D)$. Using this notation we can re-express (2.2) and (2.12) respectively as

$$d^* = \arg\max_d E\{U(\boldsymbol{Z}_d)\} = \arg\max_d \int_{\boldsymbol{Z}_d} U(\boldsymbol{Z}_d)p(\boldsymbol{Z}_d)d\boldsymbol{Z}_d, \tag{2.14}$$

and

$$P\{U(\boldsymbol{Z}_{d^*}) \geq U(\boldsymbol{Z}_d)\ \forall d\}. \tag{2.15}$$

Each choice available to the decision maker results in a vector of outcomes, which we have denoted $\boldsymbol{Z}_d$. For example, this vector may comprise costs (of drugs, primary and secondary care and so on), health outcomes, and a range of non-health outcomes such as those related to the ability of the patient to work. Let there be $j = 1, \ldots, J$ outcomes of interest, and the vector of outcomes under decision d be $\boldsymbol{Z}_d = \{o_{1,d}, \ldots, o_{J,d}\}$. Note that the decision that maximises one outcome may not be the same as the decision that maximises another. We treat the outcomes here (and in the rest of this section) as known with certainty to simplify the notation.

Given J outcomes of interest, the decision maker must choose which of the D J-dimensional vectors of outcomes that she prefers. Even if we ignore uncertainty, this is not a trivial problem and may involve complex value tradeoffs between outcomes (Keeney and Raiffa, 1976). In order to proceed the decision maker must express a *utility function* $U(\cdot)$, which can be thought of as a projection of the J-dimensional outcome space onto the real line. The function must result in $U(\boldsymbol{Z}_{d'}) > U(\boldsymbol{Z}_d)$ if the outcome $\{o_{1,d'}, \ldots, o_{J,d'}\}$ is preferred to $\{o_{1,d}, \ldots, o_{J,d}\}$, $U(\boldsymbol{Z}_{d'}) < U(\boldsymbol{Z}_d)$ if $\{o_{1,d}, \ldots, o_{J,d}\}$ is preferred to $\{o_{1,d'}, \ldots, o_{J,d'}\}$ and $U(\boldsymbol{Z}_{d'}) = U(\boldsymbol{Z}_d)$ if there is indifference between the two vectors of outcomes.

We note at this point that any decision maker is likely in reality to have a complex *implicit* utility function that includes not only costs and health consequences, but also may include for example the political or personal (for example in terms of career advancement) consequences of each decision option. The existence of some of these aspects of the decision problem may be quite hidden and we will not consider this problem further, other to recognise that decisions are rarely made solely on the result of an economic evaluation analysis.

2.5 What are the outcomes under each decision option?

If we consider health as a measurable set of 'health outputs' that is produced at the individual level in part as a result of 'resource inputs' in the form of health care then we can divide the outcomes under each decision option into two sets:

'inputs' which represent the allocation of resources, and 'outputs' which represent health. We may also want to consider non-health outputs such as ability to work, but in the simple case we will consider outcomes under each decision option as just comprising a set of resource inputs (which we will often refer to just as 'costs'), and a set of health outputs (for which we will also use the term health effects).

2.6 Measuring resource inputs

Implementing a decision option will result in the allocation of resources. These allocations are termed 'inputs' to whatever process that is generating health 'outputs'. The *relevant* inputs are those that are expected to differ systematically between the different decision options within the decision problem, since these differences will inform the choice between the decision options. The decision maker should therefore choose to include in their utility function all resource allocation inputs that are expected to differ between the decision options. For example, if we are considering a decision problem that relates to treatments for heart disease then relevant inputs would include the competing treatments themselves, but also resources committed for primary and secondary care if we believe that these will differ between decision options. We would probably *not* be interested in resource inputs relating to ophthalmic care, since these would not be expected to differ between the decision options (though ophthalmic services may still be used by individuals who have heart disease).

Inputs are usually valued in monetary units. For any health economy to function inputs to the health care process must be assigned monetary values to allow accounting of resources.

Resource allocation decisions almost always relate to populations rather than individuals, and the decision maker is therefore rarely interested in resource inputs only for a single individual, but rather in the total or per person 'average' input in some defined population. The decision maker is also likely to be interested in the inputs committed over some period of time, rather than at a single point in time. For decisions that have outcomes only in the short term, the mean per person cost of resources allocated may be of interest. Where the decision maker is interested in outcomes over extended time periods, inputs might be expressed as mean per person, per year, costs.

2.7 Measuring health outputs

There is no single accepted definition of health, and therefore no single measure of health output. There is an established literature on the measurement of health (e.g. Bowling, 1997; McDowell, 2006; Streiner and Norman, 1995). For our purposes 'health' is the set of measurable characteristics that the decision maker herself would define as comprising those aspects of an individual's health relevant to the decision problem. These characteristics may be objectively measurable using an instrument or test, such as blood pressure or cholesterol level, clinically measurable (i.e. based on the opinion of a health care professional) such as the presence or absence of a disease, or subjectively measurable, such as pain.

Given the set of measurable characteristics relevant to the decision problem, an individual can be said to exist in a 'health state' comprising the set of measured values of the characteristics. So, for example, the relevant health characteristics for some decision problem might be chest pain (measured on a four level scale: none, mild, moderate, severe) diagnostic status for angina (yes or no as diagnosed by a cardiologist) and blood pressure (measured on a continuous scale in mmHg). At some time t the health state of individual A might therefore be {no chest pain; no angina; 120mmHg} while the health state for individual B might be {mild chest pain; angina; 130mmHg}.

Again, the decision maker is rarely interested in the health of a single individual, but rather in the 'average' health experience within some defined population. The decision maker is also likely to be interested in the health experience of the population over some period of time, rather than at a single point in time, and be specifically interested in changes in health experience that differ between decision options. The decision maker therefore needs to consider meaningful measures of health output that reflect the aggregated health experience of a population of individuals over time.

In a simple case in which there is a single health characteristic of interest, alive or dead, then a sensible aggregate measure of population health output over time could be mean survival time. If the health characteristic of interest is chest pain on a four point scale (none, mild, moderate, severe), the aggregate measure might be the mean number of days on which moderate or severe pain is experienced. It is then be up to the decision maker to value this population aggregated health output under each of the decision options, along with costs and other relevant outcomes.

In the above examples health is measured in 'natural' units. 'Natural units' does not have a precise definition, but loosely speaking, a natural unit defines some measure of health that we can measure directly (but which is not a measure of health state preference or utility). So, for example for a set of decision options relating to the treatment of cancer the 'natural unit' outcome of interest may be the length of survival post treatment. For the treatment of hypertension the outcome might be blood pressure; and for decision options concerning a screening test the outcome may be numbers of cases of disease detected.

This may present a problem if there are different natural units for the different decision options within some decision problem. So, for example, if we were faced with the rather artificial decision problem: fund hip replacements versus asthma treatment, then we might find it difficult to decide the relative value of an increases in mobility versus a reduction in breathlessness. For this reason, health measured in natural units is often transformed onto a *health state preference* scale where a value of 1 represents perfect health, 0 represents death, and negative values represent states worse than death. Given two health states h_1 and h_2 the preference scale transformation $u(\cdot)$ is defined such that if h_1 is preferred to h_2 then $u(h_1) > u(h_2)$, if h_2 is preferred to h_1 that $u(h_2) > u(h_1)$ and if there is indifference that $u(h_2) = u(h_1)$. The *valuation* of health in this way is not straightforward and as for health *measurement* there is a large literature. A comprehensive recent text on the subject is Brazier et al. (2007).

The preference for a health state can be elicited from the individual who is experiencing the health state, or from the general population who are asked to imagine that they are in that health state. A commonly used method for transforming health states to preferences elicited from the general population is via a generic health outcome measurement instrument such as the SF36 (Brazier, 1993) or EQ5D (Brooks, 1996). These instruments measure health on a number of dimensions, so for example in EQ5D there are five: degree of mobility impairment, ability to self-care, ability to engage in usual activities, level of pain or discomfort, and level of anxiety or depression, with three levels for each dimension. Each of the resulting 243 health states is then associated with a measure of preference derived from the responses of study participants drawn from the general population who were asked to value each of the states, relative to full health and to death. The preferences are elicited using methods such as time trade off or standard gamble (see Drummond et al., 2005, for basic details).

Health state preference values can be considered to be measures of 'quality of life' (Bowling, 1997). Quality of life values are aggregated over time to create Quality Adjusted Life Years (QALYs). This generic measure of health outcome, Q, measured in QALYs is defined as

$$Q = \int_{t_0}^{t_1} u\{h(t)\}dt, \tag{2.16}$$

where $h(t)$ describes the health state of an individual at time t and $u(\cdot)$ is the health state preference transformation. Time t is measured in years within some interval of interest (t_0, t_1). If time is considered discrete rather than continuous then we replace the integral with a summation,

$$Q = \sum_{t=t_0}^{t_1} u(h_t). \tag{2.17}$$

2.7.1 Which health measures should the decision maker choose?

Let us imagine a large randomised controlled trial where patients in each arm of the trial are exposed to one of a set of decision options. Each individual in the trial will experience 'health' over the subsequent course of the trial. As we have discussed above, this experience of health is multifaceted and it is up to the decision maker to choose the appropriate measures to capture the 'relevant' aspects. What are the 'relevant' aspects? If our trial has randomised individuals to various treatments for heart disease then we can see that health experience related to heart disease is likely to be relevant. In this case we might measure level of chest pain, breathlessness and blood pressure say. We could also measure visual acuity or hair loss or knee joint flexibility, but we would probably not consider these relevant. Why? Because we don't expect there to be systematic differences in these measures between the trial arms.

The relevant measures are therefore those that capture health experience that is expected to differ systematically between those exposed to the different decision options. We only expect to see systematic differences in those aspects of health that are associated causally with some aspect of the decision option.

Before we describe the various approaches to mathematical modelling in health economic evaluation, we briefly review four broad types of economic analysis.

2.8 Types of economic evaluation analysis

In the health economics literature, methods for economic evaluation are categorised according to the measurement units of the output of the analysis. Classically, four types of analysis are described: cost-minimisation analysis, cost-effectiveness analysis, cost-utility analysis and cost-benefit analysis (Drummond et al., 1997).

2.8.1 Cost-minimisation analysis

A cost-minimisation analysis is applicable if all outcomes, except costs, are identical under all decision options. So, if we denote costs over the relevant time period (expressed as mean per person costs, say) as $o_{1,d}$ and the remaining $J-1$ outcomes of interest as $o_{2,d}, \dots, o_{J,d}$ then the analysis is applicable if $o_{j,d} = o_{j^*,d} \quad \forall d, j > 1, j^* > 1$. The results of this analysis are the costs $o_{1,d}$ for each of the decision options. The decision maker usually then has a utility function of the simple form

$$U(o_{1,d}) = -o_{1,d}, \tag{2.18}$$

i.e. negative costs. The optimal decision is that which maximises utility, and therefore minimises costs, i.e.

$$d^* = \arg\max_{d \in \mathcal{D}} -o_{1,d}, \tag{2.19}$$

hence the name cost-minimisation analysis. Given that we need to determine that outcomes except costs really *are* equal before a cost-minimisation analysis is applicable, a cost-minimisation analysis is better seen as just a special case of the next analysis type, cost-effectiveness analysis.

2.8.2 Cost-effectiveness analysis

A cost-effectiveness analysis is applicable if health outcome is easily measurable in 'natural units' under each decision option. The results of this analysis are the (population mean per person) inputs in cost units, $o_{1,d}$, and the (population mean per person) health outcome in natural units, $o_{2,d}$, for each of the decision options. The decision maker must then value the health outcome on the monetary scale via their 'willingness to pay' for one unit of this health outcome, λ, resulting in the *net benefit* (or *net monetary benefit*) of decision option d,

$$NB_d = \lambda o_{2,d} - o_{1,d}. \tag{2.20}$$

The decision maker's utility for the outcomes (in monetary units) is usually then assumed to be the net benefit itself, i.e.

$$\begin{aligned} U(o_{1,d}, o_{2,d}) &= NB_d & (2.21) \\ &= \lambda o_{2,d} - o_{1,d}. & (2.22) \end{aligned}$$

Net benefits can equally be expressed in health rather than monetary units, giving the *net health benefit* (Stinnett and Mullahy, 1998),

$$NB_d = o_{2,d} - o_{1,d}/\lambda. \tag{2.23}$$

2.8.3 Cost-utility analysis

A cost-utility analysis extends the cost effectiveness method to the case where there is no single common 'natural' measure of health outcome for all decision options. Instead, health states are transformed onto a health state preference scale and aggregated over time to generate population mean numbers of QALYs. The results of this analysis are the population mean per person inputs in cost units, $o_{1,d}$, and the population mean per person health outcome on the QALY scale, $o_{2,d}$, for each of the decision options. Given the decision maker's willingness to pay for one QALY unit, λ, the net monetary benefit is

$$NB_d = \lambda o_{2,d} - o_{1,d}, \tag{2.24}$$

and as above, utility in monetary units is assumed to equal the net benefit, i.e. $U(o_{1,d}, o_{2,d}) = NB_d$. Net benefit can be expressed in QALY rather than monetary units,

$$NB_d = o_{2,d} - o_{1,d}/\lambda. \tag{2.25}$$

2.8.4 Cost-benefit analysis

A cost-benefit analysis is a more general approach to economic evaluation in which all outcomes are transformed within the analysis onto the monetary scale. This allows decision options that have outcomes across multiple sectors, e.g. health, education, and transport to be compared. We again denote resource inputs, measured in cost units, as $o_{1,d}$. If there are $j = 2, \ldots, J$ health and other outcomes of interest for each decision d, and the monetary value of one unit of each outcome $o_{j,d}$ is λ_j then the result of this analysis is $\{o_{1,d}, \lambda_2 o_{2,d}, \ldots, \lambda_J o_{J,d}\}$ for each decision option. The net monetary benefit is simply

$$NB_d = \sum_{j=2}^{J} \lambda_n o_{j,d} - o_{1,d}, \tag{2.26}$$

and again, the utility $U(o_{1,d}, \ldots, o_{J,d})$ is assumed to equal NB_d itself. A cost benefit analysis, which results in outcomes on the monetary scale, implies that all outcomes are valued within the analysis itself.

2.9 Discounting

We allow costs and health outcomes to be time dependent in part because costs and outcomes may change over time, but also to allow discounting. The decision maker may wish to *value* resource inputs and health (and other) outputs that are accrued at different points in time differently. This commonly takes the form of discounting, where outcomes that are accrued in the future are reduced in value to reflect society's preference for rewards now rather than later. Costs and consequences that are accrued now are valued more highly than those accrued in the future (Krahn and Gafni, 1993). Discounting usually takes the form

$$U\{o_{j,d}(t)\} = U\{o_{j,d}(t_0)\}(1 + r_j)^{-t}, \tag{2.27}$$

where r_j is the discount rate per unit time for outcome j.

2.10 The role of mathematical modelling in economic evaluation

In section §2.3 we defined $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_D)$ as the vector of outcomes (resource inputs, health outputs, and possibly other outputs) under all decision options in some population of interest, aggregated over some period of time. Some or all of the components of $\boldsymbol{Z}$ are likely to be uncertain, and in order to determine the decision option that maximises the expected utility, $d^* = \arg\max_d E\{U(\boldsymbol{Z}_d)\}$, we must specify $p(\boldsymbol{Z})$.

It is perfectly possible for a decision maker to derive her own distribution $p(\boldsymbol{Z})$, and therefore evaluate $\arg\max_d E\{U(\boldsymbol{Z}_d)\}$ with little or no reference to any external evidence. It is unlikely, however, that these statements will be useful, and the resulting decision may easily be challenged. For a robust decision to be made, meaningful, evidence informed statements must be made about $\boldsymbol{Z}$, and it is this that provides the primary motivation for the building of a *computer model*.

Building a model is usually seen as a necessary alternative to conducting a study to determine the costs and consequences of the competing decision options. Given a decision problem it is, at least in theory, possible to allocate the population of interest randomly to the various decision options, follow up the groups for some period of time judged to be adequate, and count all the relevant costs and consequences in each of the study arms. Although conducting such a trial would arguably provide the means of deriving $p(\boldsymbol{Z})$, such a trial is usually not feasible. Firstly, the decision maker may be interested in costs and health outcomes accrued over a time period of the order of years or decades. Running a trial of such a length is likely to be extremely expensive, and results will only be known at some point far into the future. This is of little benefit if the decision needs to be be made now. Secondly, it may be unethical to randomise participants to all the available decision options if some decision options are already known to result in better clinical outcomes than others (but perhaps at greater cost). Lastly, a trial, once set up does not have the same flexibility of a mathematical model in its ability to explore large numbers of alternative scenarios or sets of assumptions.

Within the context of health economic evaluation, models built to predict $\boldsymbol{Z}$ are variously referred to as 'health economic models', 'decision models', 'decision analytic models', 'economic evaluation models' or 'cost-effectiveness models' (note that the term 'cost-effectiveness' is used here more loosely than in section §2.8.2 to encompass any economic evaluation analysis).

The primary purpose then of a health economic decision model is to provide the decision maker with an 'evidence based' judgement about outcomes, i.e. the model allows the decision maker to evaluate $p(\boldsymbol{Z}|\boldsymbol{D})$ for some data $\boldsymbol{D}$, rather than just their prior beliefs $p(\boldsymbol{Z})$. Models therefore link evidence (in the form of parameters estimated from data along with other judgements about the world) to the outcomes $\boldsymbol{Z}$. Buxton et al. (1997) define a number of specific functions that models perform: extrapolating beyond the data observed in a trial; linking intermediate clinical endpoints to final outcomes; generalising from one setting to another; and synthesising head-to-head comparisons where relevant trials do not exist. These roles can all be thought of as falling under the banner of 'evidence synthesis' in the sense that in each case we combine data from various sources, perhaps with other evidence in the form of any elicited parameter values, and evidence that is encoded in the model structure itself.

Brennan and Akehurst (2000) cite two additional roles for models in addition to their use in synthesising evidence in order to make healthcare resource allocation decisions: to inform research strategy and design, and to make explicit the evidence supporting a decision. The first of these acknowledges that a decision maker always has the option of delaying a decision until more information is available. Models can be used to quantify the expected value of any additional information, and this can be combined with the costs of obtaining the information (say, from a new study) to predict the net utility of delaying the decision to gather data (see Bernardo and Smith (1994) and Howard (1966) for theoretical discussions regarding the value of information, and Claxton and Posnett (1996), Felli and Hazen (1998) and Chilcott et al. (2003) for discussions in the context of health economic decision modelling).

Viewing models as vehicles for making explicit the evidence supporting a decision is attractive, but potentially problematic. For this to be strictly true, the decision would need to lie with the model, and not with the decision maker. There would be no possibility that a decision maker could weigh the evidence from the output of a model with other evidence or considerations from elsewhere, since the model would then no longer be making explicit the evidence behind the decision! By placing the model output in a wider 'political' context, say, the decision maker is combining the model output with other evidence, using some implied conceptual model and a utility function that is almost certainty *not* explicitly stated (or even understood).

2.11 How do health economic evaluation models work?

The majority of models that are used to aid healthcare resource allocation decisions fall into the broad category of 'cost-effectiveness' models. Such a model is, in one sense, just a representation of the modeller's judgements about the functional relationships between the inputs and outputs of the model, which in turn is informed by their views about the processes by which a population utilises healthcare resources and the causal chain that links health care utilisation to health related events at the individual level. For example, if the decision was whether or not to allow a new drug for diabetes, then the causal pathways linking the costs and consequences to the use or not of the drug would be embedded within the structure of the model.

Operationally, cost-effectiveness models synthesise information in order to quantify the costs and consequences associated with each competing decision option such that the optimum decision option can be chosen. The sources of information synthesised for, say, a diabetes model may be: estimates of treatment effects for the new drug and any alternatives, the costs of the new drug, its alternatives, and the other healthcare costs associated with diabetes, the epidemiology and natural history of diabetes and related diseases, and the utility valuations of the health care costs and health states associated with diabetes. These sources of information become the model inputs. Cost-effectiveness models are typically deterministic, but because we are uncertain about the model inputs, this uncertainty is propagated through to the model output.

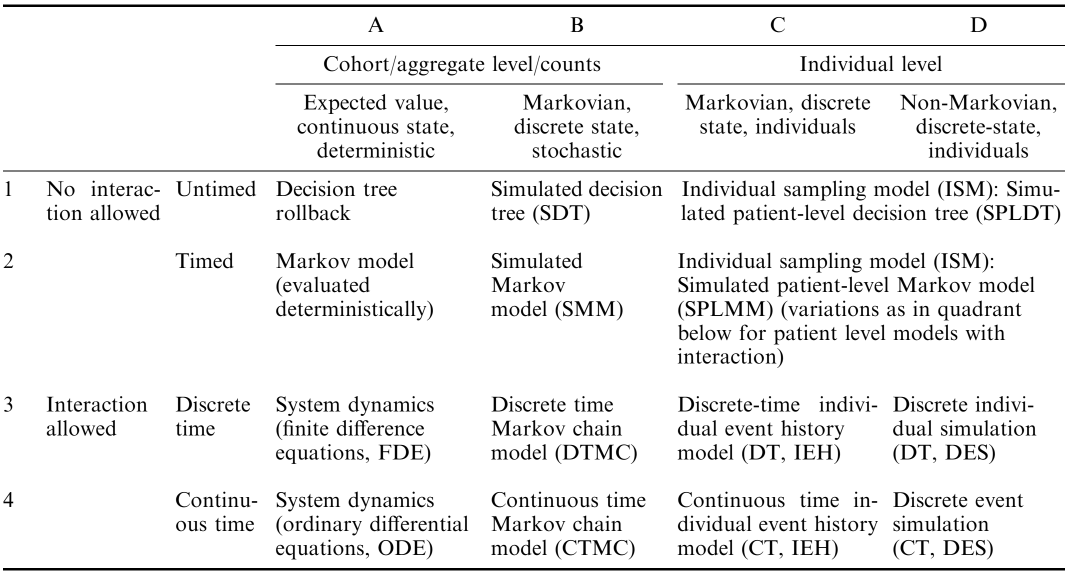
Unless there has been a study that assesses the effect of the different decision options in a population that is relevant to our decision, then we can not usually derive the distribution on the outcomes under each decision directly. Instead, we estimate the outcome (i.e. the vector of resource inputs and health outputs) for each of a series of well defined states at each time point, along with the proportion of our population of interest who occupy that state at that time point.

2.12 Categorising health economic decision models

Models for economic evaluation can be categorised on a number of different dimensions. Brennan et al. (2006) suggest a taxonomy based on the following criteria:

- whether individuals are explicitly modelled,
- whether interaction between individuals is modelled,
- whether time is modelled, and if so, whether it is treated as discrete or continuous,
- whether transitions between states are assumed to be Markovian,
- whether the model is purely deterministic, or whether Monte Carlo sampling is used to compute expectations.

This categorisation of the various economic models that were in use at the time of the study led Brennan et al. (2006) to produce the following table of model types (table 2.1).



			A	B	C	D
			Cohort/aggregate level/counts		Individual level	
			Expected value, continuous state, deterministic	Markovian, discrete state, stochastic	Markovian, discrete state, individuals	Non-Markovian, discrete-state, individuals
1	No interaction allowed	Untimed	Decision tree rollback	Simulated decision tree (SDT)	Individual sampling model (ISM): Simulated patient-level decision tree (SPLDT)	
2		Timed	Markov model (evaluated deterministically)	Simulated Markov model (SMM)	Individual sampling model (ISM): Simulated patient-level Markov model (SPLMM) (variations as in quadrant below for patient level models with interaction)	
3	Interaction allowed	Discrete time	System dynamics (finite difference equations, FDE)	Discrete time Markov chain model (DTMC)	Discrete-time individual event history model (DT, IEH)	Discrete individual simulation (DT, DES)
4		Continuous time	System dynamics (ordinary differential equations, ODE)	Continuous time Markov chain model (CTMC)	Continuous time individual event history model (CT, IEH)	Discrete event simulation (CT, DES)

Figure 2.1: Taxonomy of model structures. (From Brennan et al., 2006)

Although the range of model types used in economic analysis is broad, the majority of models fall into one of the following three categories: decision trees, Markov models and individual level simulation models.

2.12.1 Decision tree models

The simplest model commonly used in economic evaluation is the 'decision tree'. Figure 2.2 shows a decision tree designed to inform whether or not to fund an exercise promoting intervention. The left-most (square) node represents the choice between the two decision options, $d = 1$, no intervention, and $d = 2$, the exercise promoting intervention.

The subsequent circular nodes are 'chance nodes'. If we imagine a single individual at the left most chance node, we expect them to proceed to an 'exercise' state with some probability p_1 and to a 'sedentary' state with some probability $1 - p_1$. The next chance node then represents the options open to an individual, conditional on them being in the 'exercise' state: they will proceed to the 'exercise maintained' state with probability p_2 and to the 'exercise not maintained' state with probability $1-p_2$. The third node represents the probability of eight mutually exclusive 'terminal' states, conditional on each of the three outcomes from the first two nodes: exercise that is maintained, exercise that is not maintained, and no exercise (sedentary lifestyle).

We can calculate the probability of each of the 'terminal' states by multiplying out the conditional probabilities along each of the possible paths through the tree to the terminal state. Given the terminal state probabilities and a vector of outcomes associated with each state (usually costs and some measure of health output such as QALYs) we can analytically calculate the expected outcomes under each decision. This process is sometimes described as 'rolling back' the decision tree.

Decision tree models are typically 'cohort' models in that the values for the conditional probabilities and the outcomes are population 'average' measures of these quantities, and as such that the expected costs and health outputs relate to the population, not any single individual. In this sense single individuals are not

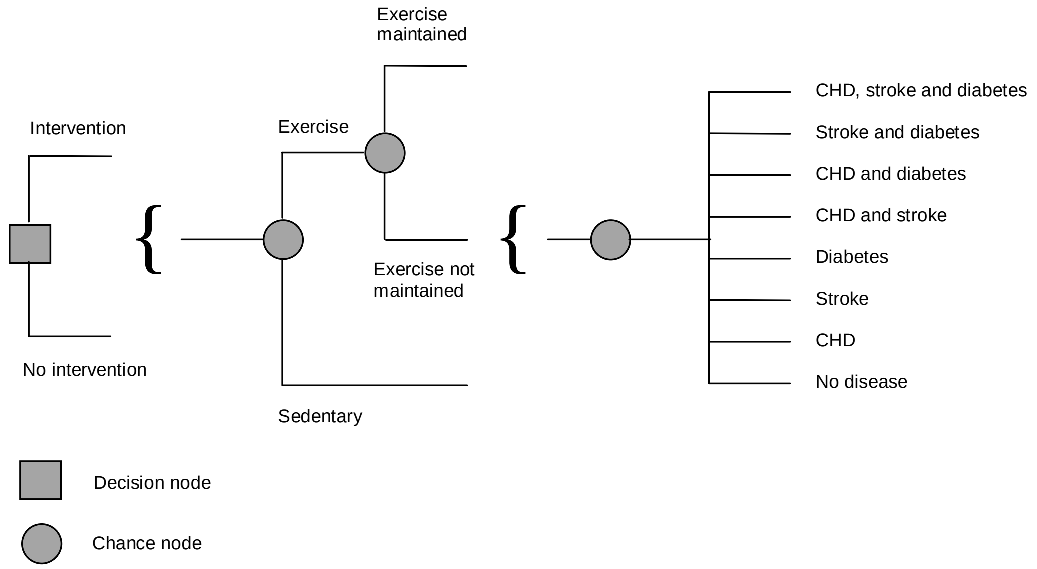


Figure 2.2: A simple decision tree designed to inform whether or not to fund an exercise promoting intervention.

explicitly modelled in this form of simple decision tree, and neither is there any explicit modelling of interaction between individuals.

Within a decision tree model as expressed in figure 2.2, time is implied from left to right, in that the probabilities in the model relate to moving from some state A to a state on the right of A, conditional on being in A (and in this sense the model is Markovian, though the intermediate nodes in a decision tree may not be obviously thought of as states). However, there is no *explicit* modelling of time within the model. For the purposes of the computation all events can be considered to occur simultaneously.

In a decision tree model population level expected outcomes can be computed analytically without the need for Monte Carlo sampling, and as such these models are entirely deterministic. Note that by defining the model as deterministic we are not saying that the values of the model inputs (and hence the model outputs) are known with certainty, just that each set of model inputs uniquely determines a single model output set.

2.12.2 Markov models

We may decide that in order to inform a particular decision we need to predict outcomes over some period of time. Within this period of time (say the five years following the implementation of the decision option), we expect that individuals will move through various states, each state having different outcomes in terms of resource inputs and health outputs. So, for example, if our decision problem concerns a choice between a number of different cardiovascular drug treatments we might define a set of states for each treatment option as ‘stable angina’, ‘unstable angina’, ‘myocardial infarction’ and ‘dead’. We expect resource inputs and health outputs (measured say in terms of quality of life) to differ between the states.

In order to compute the total resource inputs and health outputs over the time period of interest under each decision option we wish to predict the proportion of the population in each health state at each time point t, where we treat time as a discrete variable. If we have data that allow us to estimate the probabilities of transition to each state j at time point $t+1$ conditional on being in state i at time t, one approach would be to construct a decision tree that recursively branches at each time point. Assuming D decision options and a time period of interest that is divided into T time steps, this will result in a tree with $D \times 4^T$ branches - potentially a very ‘bushy’ tree indeed. It will be difficult to graphically represent a decision tree of this type for any T that is not trivially small. In these circumstances it is more convenient to specify a ‘Markov’ model rather than a decision tree (Beck and Pauker, 1983; Sonnenberg and Beck, 1993).

In a Markov model we specify a ‘state vector’, $\boldsymbol{\pi}_d(t) = \{\pi_{1,d}(t), \ldots, \pi_{I_d,d}(t)\}'$ that describes the proportions of the population that exist in each of the $i_d = 1, \ldots, I_d$ states defined in the model at time point t under decision option d; and a matrix of transition probabilities, $\boldsymbol{M}_d(t)$, that describes the probability of transition from state i at time point t to state j at time period $t+1$. Given the proportion of the population in each state at time point zero (under each decision

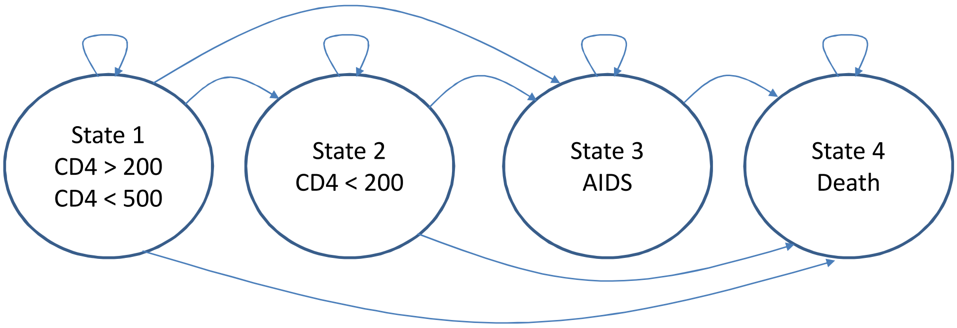


Figure 2.3: Graphical representation of a Markov model for a decision concerning treatment for HIV/AIDS

option) the proportion of the population in each state at time point t is given by

$$\boldsymbol{\pi}'_d(t) = \boldsymbol{\pi}'_d(0) \prod_{n=1}^{t} \boldsymbol{M}_d(t). \tag{2.28}$$

Figure 2.3 shows a graphical representation of a four state Markov model first described in Chancellor et al. (1997) and subsequently used for illustrative purposes in Drummond et al. (2005) and Briggs et al. (2006). The purpose of the model is to predict costs and health outcomes (life years) under two drug treatment options in people with HIV. The directed arrows show ‘allowable’ transitions between states. Where no arrow exists between states, the probability of transition is zero (this occurs, for example, for all transitions *out* of the dead state). The $j = 1, \ldots, J$ outcomes associated with state i_d at time t are denoted $o_{i_d,j}(t)$ and the total outcome j is therefore

$$o_j = \sum_{t=t_0}^{t_1} \sum_{i_d=1}^{I_d} \pi_{i_d}(t) o_{i_d,j}(t). \tag{2.29}$$

Markov models are described as ‘time homogeneous’ if $\boldsymbol{M}_d(t) = \boldsymbol{M}_d\ \forall t$. Relaxing this assumption allows the transitions between states to differ according to, for example, age or the length of time since initial treatment with some drug.

As with decision tree models, Markov models in this simple formulation are ‘cohort’ models if the values for the conditional probabilities and the outcomes are population 'average' measures of these quantities. Single individuals are not explicitly modelled, and neither is there any explicit modelling of interaction between individuals. In this form, outcomes can be computed analytically without the need for Monte Carlo sampling, and as such Markov models are entirely deterministic.

2.12.3 Individual patient level simulation models

Individual patient level simulation models are a broad class of models in which individuals are explicitly modelled. This approach allows outcomes to be functions of individual level covariates, rather than functions just of population level covariates.

This approach is often adopted if the relationship between outcomes and individual level covariates is non-linear. If we write outcome j for individual i as a function $\eta(\cdot)$ of some individual level covariates $\boldsymbol{v} = \boldsymbol{v}_i$, as well as uncertain inputs $\boldsymbol{X}$, then $o_{ij} = \eta(\boldsymbol{v}_i, \boldsymbol{X})$. Assuming covariates vary across the population of interest, the population mean outcome, $E_{\boldsymbol{v}}\{\eta(\boldsymbol{v}, \boldsymbol{X})\}$, will only equal $\eta\{E_{\boldsymbol{v}}(\boldsymbol{v}), \boldsymbol{X}\}$, the function $\eta(\cdot)$ evaluated at the population mean value of the covariate, if $\eta(\cdot)$ is linear in $\boldsymbol{v}$. Even in a very simple case in which a single risk factor exists at two levels within a population, Zaric (2003) shows that result of a Markov cohort model is biased if transition probabilities are functions of the risk factor.

Evaluating the expectation $E_{\boldsymbol{v}}\{\eta(\boldsymbol{v}, \boldsymbol{X})\}$ in an individual level model is usually difficult or impossible analytically, hence the use of Monte Carlo simulation. Samples $\boldsymbol{v}_i$ $(i = 1, \ldots, n)$ are drawn from a joint distribution $p(\boldsymbol{v})$ that represents individual level variability in the covariates, and $\eta(\boldsymbol{v}_i, \boldsymbol{X})$ evaluated in each case. The resulting sample set $\{\eta(\boldsymbol{v}_1, \boldsymbol{X}), \ldots, \eta(\boldsymbol{v}_n, \boldsymbol{X})\}$ is then taken as a sample from the distribution of the outcome o_j across individuals in the population of interest (conditional on $\boldsymbol{X}$). From this the population level mean for the outcome can be easily estimated by the sample mean of $\{\eta(\boldsymbol{v}_1, \boldsymbol{X}), \ldots, \eta(\boldsymbol{v}_n, \boldsymbol{X})\}$, along with any other statistic of interest. This individual level variability is usually referred to as 'first order' uncertainty (e.g. see Groot Koerkamp et al., 2010). It is quite separate to any consideration of uncertainty about $\boldsymbol{X}$, the values of the inputs to the model ('second order' uncertainty in this context).

If Monte Carlo simulation is used to evaluate the expectation $E_{\boldsymbol{v}}\{\eta(\boldsymbol{v}, \boldsymbol{X})\}$ then the model is no longer deterministic. However, as the simulation size n is increased, the Monte Carlo estimate of the population mean outcome will converge to $E_{\boldsymbol{v}}\{\eta(\boldsymbol{v}, \boldsymbol{X})\}$, which is uniquely determined for each input set. In this sense individual level simulation models can be considered deterministic.

A second reason for adopting an individual level model is in the case where differences in outcomes between different population subgroups is of interest. Rather than having to run a cohort model with a new set of input parameters for each sub group, an individual level model allows the sub group analysis to be performed on the single set of results that the model generates. This may allow considerable extra flexibility in defining the sub groups.

Thirdly, an individual level model can allow the incorporation of time and history dependence in transition probabilities in an intuitive manner. Incorporation of time and history dependence is possible within a Markov cohort model, but at the expense of generating an unwieldy model with a large number of states (Karnon, 2003).

Lastly, by treating individuals within a model separately, interactions between individuals can be considered. For example, we may want to model the effect of herd immunity in the context of a decision scenario involving competing immunisation schedules. Modified cohort model approaches to this problem have been described (Bauch et al., 2009), but these are only approximate. Alternatively, interaction may arise through competition for scarce resources such as hospital beds. An individual level model can allow the probability that an individual enters some state (e.g. occupying a hospital bed) at some time t to depend on the number of other individuals already occupying that state at time t.

2.13 The general formulation of a cost effectiveness model

The various forms of cost-effectiveness model described above are all particular forms of a more general model that we can specify as follows.

We imagine that for decision option d all of the individuals within our population of interest exist at each time point $t \in (t_0, t_1)$ in one of a set of $i_d = 1, \ldots, I_d$ states. Each state is associated with a vector of $j = 1, \ldots, J$ outcomes of interest (i.e. resource inputs and health outputs) at each time point t. We write the numerical value of the outcome j associated with state i_d at time t as $o_{i_d,j}(t)$. The proportion of the population of interest that exists in state i_d at time t under decision option d is denoted by $\pi_{i_d}(t)$. Alternatively, $\pi_{i_d}(t)$ can be thought of as the probability that a single individual exists in state i_d at time t given decision option d.

Summing over the states gives us the total outcome j at time t under decision d,

$$o_{j,d}(t) = \sum_{i_d=1}^{I_d} \pi_{i_d}(t) o_{i_d,j}(t). \tag{2.30}$$

Summing over time then gives us the total outcome j under decision d,

$$\begin{aligned} o_{j,d} &= \sum_{t=t_0}^{t_1} o_{j,d}(t), && (2.31) \\ &= \sum_{t=t_0}^{t_1} \sum_{i_d=1}^{I_d} \pi_{i_d}(t) o_{i_d,j}(t). && (2.32) \end{aligned}$$

If we treat time as continuous we replace the summation over time with an integral,

$$\begin{aligned} o_{j,d} &= \int_{t_0}^{t_1} o_{j,d}(t)dt, && (2.33) \\ &= \int_{t_0}^{t_1} \sum_{i_d=1}^{I_d} \pi_{i_d}(t) o_{i_d,j}(t)dt. && (2.34) \end{aligned}$$

Given the outcomes $o_{j,d}$ the decision maker then evaluates the utility for each decision, $U(o_{1,d}, \ldots, o_{J,d})$, and (if we ignore uncertainty) chooses the decision that maximises this, i.e. $d^* = \arg\max_d U(o_{1,d}, \ldots, o_{J,d})$. If outcomes are uncertain due to uncertainties in $\pi_{i_d}(t)$ or $o_{i_d,j}(t)$ then the decision maker maximises expected utility, i.e. $d^* = \arg\max_d E\{U(o_{1,d}, \ldots, o_{J,d})\}$.

So, for example, we have a decision problem relating to whether or not to recommend a new anti-hypertensive drug ($d = 1$) versus an existing drug ($d = 2$). The outcomes of interest are costs ($j = 1$) and health effects in QALYs ($j = 2$) in some population of interest. There has been no trial of the new drug against the old that measures these outcomes. We do, however, have trials that measure the effects of the old and new drugs on the incidence of stroke and myocardial infarction (MI). We define a set of four states under decision $d = 1$ as {*well after taking new drug; stroke after taking new drug; MI after taking new drug; stroke and MI after taking new drug*}, and index these states $i_1 = 1, \ldots, 4$. We define a set of four states under decision $d = 2$ as {*well after taking existing drug; stroke after taking existing new drug; MI after taking existing drug; stroke and MI after taking existing drug*}, and index these $i_2 = 1, \ldots, 4$. The proportion of the population that experiences state i_d under decision d at time t is $\pi_{i_d}(t)$.

We also have data on the costs of the drugs and the costs of treating stoke and MI, and the population mean numbers of QALYs for those who are well, and those who have had a stroke, MI, or both. The costs and the QALYs associated with state i_d at time t are $o_{i_d,1}(t)$ and $o_{i_d,2}(t)$ respectively. We then sum over the states and over time to obtain costs and QALYs under each decision d via equation (2.31).

So the components of this general model are, $o_{i_d,j}(t)$, the outcomes associated with each state as functions of time, and $\pi_{i_d}(t)$, the proportions of the population in each health state, again as functions of time. Different types of economic evaluation models primarily differ in the 'machinery' that determines how the state vector, $\boldsymbol{\pi}_d(t) = \{\pi_{1_d}(t), \ldots, \pi_{I_d}(t)\}'$, evolves with respect to time. For example, the evolution of the state vector may be explicitly determined by a simple Markov process, or instead it may be determined indirectly as the result of an individual level discrete event simulation. It is in this 'machinery' that much of the uncertain 'structure' of the model is.

2.14 Sources of error in a health economic decision model

In chapters 5 and 6 of the thesis we will discuss in detail the management of structural uncertainty in the context of two cost-effectiveness model case studies. At this point we consider some possible causes of model error, given the general formulation of the model in the section above. This will help to provide a framework for thinking about model structure uncertainty.

2.14.1 Choice of outcomes

First we note that to avoid potential structural error all of the $j = 1, \ldots, J$ outcomes must be counted under all decision options. Any outcomes that are missing from the model specification for a subset of decision options are implicitly assumed to be zero. If in reality they are not zero, then this is a source of error. For example, if a decision is between heart transplants ($d = 1$) and hip replacements ($d = 2$), and we have counted only the costs of heart transplants for $d = 1$ and only the costs of hip replacements in $d = 2$, we are implicitly assuming that there are no costs associated with heart transplants for $d = 2$ and no costs associated with hip replacements in $d = 1$. Likewise, when we are computing health outputs, if the output is not included in the model specification it is assumed to be zero. In many circumstances this may be entirely reasonable.

The set of outcomes that the model generates has to include all *relevant* outcomes. If we are interested in a choice between cardiac drugs and in our model we count costs and health outcomes that relate to ophthalmic care, then our model might be entirely correct. However, it it will not be useful for the decision problem at hand (the right answer to the wrong question, or an 'error of the third kind', Kimball, 1957). Whether or not the model includes all relevant outcomes is a judgement that the decision maker must make before they use the results of the model to inform the decision.

2.14.2 Determining the evolution of the state vector

Next, whatever the form of the model, it must generate the correct value for the probability that an individual exists in state i_d at time t, $\pi_{i_d}(t)$, and the correct value for outcome j associated with state i_d at time t, $o_{i_d,j}(t)$. Given that states are just arbitrary constructs that allow probabilities (or population proportions) and outcomes to be associated together, what we are really saying here is that the model must correctly compute weights for a set of outcome values. If not, there is error.

We can always write down a correct model. For example, if the resource costs of interest under decision option d are $o_{d,1}$ and health effects are $o_{d,2}$ then we can write the following model for the net benefit,

$$NB_d = \lambda o_{d,2} - o_{d,1}. \tag{2.35}$$

The model inputs are $\boldsymbol{X} = \{(o_{1,1}, o_{1,2}), \ldots, (o_{D,1}, o_{D,2})\}$. If we learned the true values of these inputs, then we would know the true net benefit under each decision option, and there is no model error. The source of uncertainty about the target quantity is therefore entirely located within the inputs, and there is no structural uncertainty.

If we partition the population of interest into n sub groups we can again write down a correct model. If the proportion of the population in subgroup i under decision d is $\pi_{i,d}$, and if subgroup i has mean costs $o_{i,d,1}$ and mean health effects $o_{i,d,2}$, then the mean population net benefit is

$$NB_d = \lambda \sum_{i=1}^{n} \pi_{i,d} o_{i,d,2} - \sum_{i=1}^{n} \pi_{i,d} o_{i,d,1}. \tag{2.36}$$

The inputs are now $\boldsymbol{X} = \{(o_{1,1,1}, o_{1,1,2}), \ldots, (o_{n,D,1}, o_{n,D,2}), \pi_{1,1}, \ldots, \pi_{n,D}\}$. As above, if we learned the true values of $\boldsymbol{X}$ we would know the net benefits, and there is no structural error. Extending this argument further, we could introduce a time element and write down a correct model in which we now also sum over the time steps,

$$NB_d = \lambda \sum_{t=t0}^{t1} \sum_{i=1}^{n} \pi_{i,d}(t) o_{i,d,2}(t) - \sum_{t=t0}^{t1} \sum_{i=1}^{n} \pi_{i,d}(t) o_{i,d,1}(t). \tag{2.37}$$

The problem is how to correctly compute the population proportion terms $\pi_{i,d}(t)$. Decision trees, Markov models and discrete event type models reflect different levels of sophistication in generating the population proportion terms $\pi_{i,d}(t)$, and therefore allow different levels of complexity in terms of assumptions regarding the relationships between states in the model. We get *structural* error if the assumptions embedded in the model that link inputs to the population proportion terms $\pi_{i,d}(t)$ do not properly reflect the system. So, for example, if the Markov assumption is invalid, then there is structural error, or if in a simple decision tree there are branches missing, then we have structural error.

2.14.3 Choice of model states

The ultimate purpose of the model is to quantify $p(\boldsymbol{Z}|\boldsymbol{D})$, the conditional distribution of the outcomes given some body of evidence. The model structure needs to be chosen to facilitate the incorporation of $\boldsymbol{D}$ in such a way as to minimise uncertainty. If we have good quality data on the costs and outcomes that relate to a set of well defined states, along with good quality data from which we can estimate the population proportions in each of these states, then using this set of states in our model seems sensible. If we use a different set of states, less 'congruent' with the data, then we are likely to introduce an uncertain error into our predictions.

2.14.4 Choice of decision options

We have implicitly assumed that the set of decision options $d = 1, \dots, D$ is given, but it is worth noting that this is of course a 'structural' choice that must be made, along with many others. To some extent there is an issue of semantics, in that any single decision problem may be considered to be defined by the set of competing options chosen, and therefore that within the decision problem the choice of options is given. If, however, we consider the wider problem of how to maximise efficiency in resource allocation then the choice of decision options for each analysis *is* important. By *not* including within the competing set an option which is in fact the optimum choice, then resource allocation will be inefficient.

2.15 The health care decision context and NICE

Up to now we have discussed decision making in rather general terms without considering any specific context in which resource allocation decisions are made. In England and Wales the National Institute of Health and Clinical Excellence (NICE) has been given a remit by parliament to make recommendations on the use of new and existing interventions and programmes within the NHS, and as such acts as a 'decision maker'.

In its decision making capacity, NICE operates according to well defined processes (see for example the methods manual for technology appriasal NICE, 2008). These processes define certain aspects of the decision problem that would otherwise be considered uncertain, and so, in a sense, reduce the space of $\boldsymbol{\theta}$ in (2.2) and (2.12). By making methods and processes explicit, decision problems become more manageable, but more importantly, a certain degree of consistency across decisions is ensured.

Table 2.1 shows the NICE 'reference case' for a decision problem that concerns the assessment of a set of competing health technologies. It defines (to some extent) the following: the technologies that should be considered as competing decision options; the set of costs and outcomes that should be considered relevant; the type of economic evaluation (though the term 'cost-effectiveness analysis' in the table means cost-utility analysis as described in section §2.8.3); the sources of evidence for various components of $\boldsymbol{\theta}$; the unit of measurement for health effects (the QALY) and the discount rate for costs and effects. Not included in the table, but implicit in the way in which NICE makes decisions, is the assumption that the utility associated with a decision option is equal to its expected net monetary benefit.

Table 2.1: The NICE Reference Case

Element of health technology assessment	**Reference case**
Defining the decision problem	The scope developed by the Institute
Comparator	Therapies routinely used in the NHS, including technologies regarded as current best practice
Perspective on costs	NHS and PSS
Perspective on outcomes	All health effects on individuals
Type of economic evaluation	Cost-effectiveness analysis
Synthesis of evidence on outcomes	Based on a systematic review
Measure of health effects	QALYs
Source of data for measurement of HRQL	Reported directly by patients and/or carers
Source of preference data for valuation of changes in HRQL	Representative sample of the public
Discount rate	An annual rate of 3.5% on both costs and health effects
Equity weighting	An additional QALY has the same weight regardless of the other characteristics of the individuals receiving the health benefit

HRQL, health-related quality of life; NHS, National Health Service; PSS, personal social services; QALYs, quality-adjusted life years.
From NICE (2008) Guide to the methods of technology appraisal, p30.

By adopting the NICE reference case, the decision maker is restricting her ability to properly compute her posterior beliefs about the 'true' utility for each decision option. In fact, the reference case restricts the ability of the analysis to properly estimate the uncertain future costs and health effects. Not all costs will fall within the perspective of the NHS or social services, and 'health' cannot be captured perfectly by the QALY. We are in some senses trading off bias against uncertainty. An analysis that counts all relevant costs and measures health in all its facets may lead to a better specification of posterior beliefs about the net benefit of each decision option, but will result in larger uncertainties. We will not discuss this problem further, other than to note that we will assume the constraints of the NICE reference case when presenting the two case studies.

Finally, we note that NICE play a role in determining the implied value of health, in the sense that they set a 'threshold' value for a QALY. If an intervention, programme or service can generate QALYs at a cost per QALY that is less than this threshold value, then the intervention, programme or service is usually adopted. The assumption is that when any new intervention is funded, the activity that is displaced (due to the total budget constraint) was generating QALYs at a cost per QALY *greater* than the threshold. The threshold therefore represents an estimate of the 'shadow price' of the budget constraint, rather than a 'willingness to pay' for one unit of health (Claxton et al., 2010; McCabe et al., 2008; Culyer et al., 2007). Current NICE guidance is that the threshold should be set at between £20,000 and £30,000 (NICE, 2008).

2.16 Conclusion

In this chapter we have reviewed some basic decision theory that underpins health economic evaluation. We have identified the role of computer models in the process of economic evaluation as allowing the decision maker to derive the uncertainty distribution of relevant outcomes conditional on evidence. Various common types of economic model were shown to be special cases of a general model that integrates outcomes of interest over a population and over time. In order to compute the integration a series of states is defined in such a way to allow the resulting input parameters to be conditioned on evidence. The choice of states, along with the structure of the model that determines the evolution of the state vector with respect to time, is an important determinant of model 'correctness'.

The next chapter reviews the literature on managing uncertainty computer models. What can we learn from the wider literature that will help us to manage uncertainty in cost-effectiveness models?

Chapter 3

Managing Uncertainty in Computer Models

3.1 Introduction

We saw in chapter 2 that health economic evaluation methods typically make use of some kind of computer model in order to make predictions about uncertain future costs and health effects. In this chapter we broaden our perspective and review the management of uncertainty in computer models in general. Computer models are built for many different purposes and have a variety of forms, but typically an individual model can be represented as a (known or unknown) mathematical function, $\boldsymbol{y} = f(\boldsymbol{x})$, that is implemented in computer code.

We recognise that we are uncertain about the quantity we are trying to predict (indeed that is why we build a predictive model), but that after building the model we are almost always *still* uncertain about the true value of the target quantity. We discuss three important sources of this uncertainty: uncertainty about the 'true' values of the model inputs, uncertainty about the model output when the model is expensive (slow) to evaluate, and uncertainty about the 'true' structure of the model. We define 'true' input and 'true' structure later in the chapter.

There is an extensive literature on the treatment of uncertainty in computer models, particularly regarding the uncertainty in the inputs. More recently, methods for managing 'code' uncertainty (i.e. uncertainty due to an computationally expensive model) have been established, most notably those based on the construction of a statistical 'emulator' for an expensive 'simulator'. Finally we note the emerging literature on the treatment of computer model structural uncertainty in the field of health economics, including some applications of statistical approaches to model uncertainty.

3.2 Computer models

3.2.1 Definitions

We can think of a computer model as a representation of some process by a series of instructions that is understood and executed by computer. Computer models are used widely in the physical and social sciences for the purposes of exploring and understanding phenomenon, as well as for prediction. Importantly, models usually attempt to represent a complex process that may or may not be fully understood in a form that is less complex and/or better understood. This is reflected in the following definitions in Bayarri et al. (2009):

> "Computer models are computational representations of complex real-world processes ... often based on numerical implementations of mathematical models developed to approximately describe the real-world process, or on efforts at direct simulation of the process"

and Bunge (1967) (quoted in Devooght, 1998):

> "a model is a reduced and parsimonious representation of a physical, chemical, or biological system in a mathematical, abstract, numerical, or experimental form."

McKay and Morrison (1997) point out that a mathematical model is a formalisation of *assumptions*:

> "a formal statement of assumptions about a relationship between known quantities x and unknown quantities y. The structure of a model defines how characteristics of y are determined from those of x. Structure, in this sense, is a mathematical or computational algorithm."

Santner et al. (2003) regard computer models as simulators for physical processes and as such generate data just as physical processes do. We can therefore conduct *computer experiments* in order to observe the output of a computer model. This necessitates thinking about (statistical) *design* and *analysis* in a similar way to that which we would for a natural experiment. Where models are explicitly built to simulate a real world process in this way, the term 'simulator' is sometimes used. We will use this term when discussing the statistical emulation of computationally expensive models in section §3.4.2.

Kotiadis and Robinson (2008) describe a series of stages between the *world*, where the decision problem resides, and the computer model (figure 3.1). Firstly, the authors postulate the existence, within the mind of the modeller, of a 'system description' of the real world processes of interest. This is fairly loosely defined, but represents the modeller's understanding of all of the relevant real world processes. Given the system description, the modeller then defines a 'conceptual model': an abstraction and simplification of the system description that describes the 'objectives, inputs, outputs, content, assumptions and simplifications of the model.' (Robinson, 2008). Lastly, a 'computer model' is built (i.e. code is written) to represent the conceptual model.

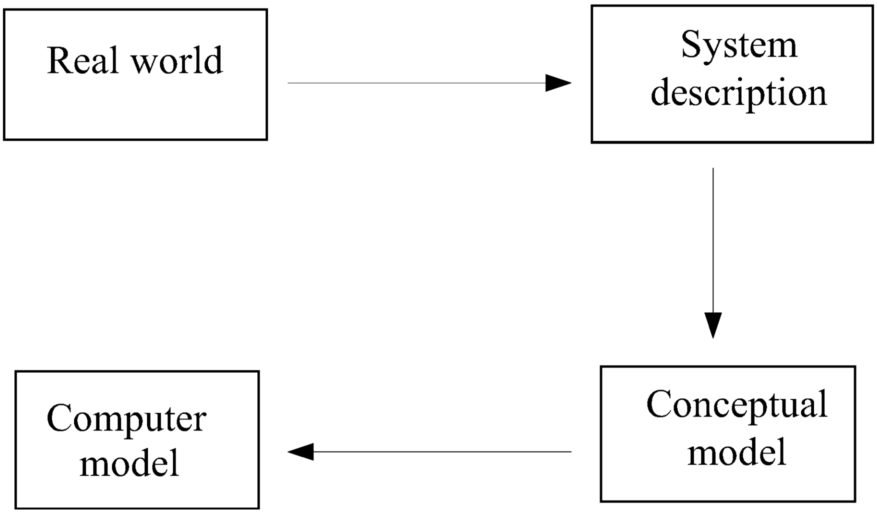


Figure 3.1: 'Artifacts of conceptual modelling' from Kotiadis and Robinson (2008).

This framework is helpful because it forces us to make a distinction between what we know, but choose not to model, and what we do not know. We can not say anything useful about the world beyond our understanding that comprises our system description. When we come to think about model error we can use this distinction to partition the model error into a portion that represents error that we can make some judgements about (the difference between the output of the model I build and the best model I might be able to imagine), and a portion that represents error that we can make no useful judgements about (the 'unknown unknown') that exists in the world that is outside of my system description.

Our understanding is that our 'best' model above represents Goldstein and Rougier (2009)'s *reified* simulator, the simulator that we do not expect to be able to build, but which reflects *all* our useful judgements about the system. Once we have conceived of our 'best' model we can, by definition, incorporate no further useful judgements about the system. Thus, the residual error between the prediction and reality is probabilistically independent (in the Bayesian subjective sense) of the built model and the best model and their corresponding inputs.

3.2.2 Saltelli's classification of models

Models vary in their purpose, as well as in their structure and underpinning assumptions about the world. Saltelli et al. (2008) categorise models on two dimensions: *diagnostic* versus *prognostic*, and *data-driven* versus *law-driven*.

The first of these dimensions, diagnostic versus prognostic, concerns the aim of the model. Diagnostic models are used to understand a law or phenomenon, whereas prognostic models are used for prediction. We may build a diagnostic model for the purposes of exploring the emergent properties of a complex system, or to play a series of 'what-if' games, or to better understand the range of behaviour of some system when parameters are pushed to their limits. The primary purpose is understanding the system, rather than to predict a quantity. Prognostic models on the other hand are primarily built to predict the value of an unknown quantity for the purposes of some decision or action. The distinction is of course blurred, and many models are useful for both purposes.

Saltelli et al. (2008)'s second dimension concerns the broad distinction between two underlying principles that drive the construction of models. Data-driven models begin with data and attempt to describe in functional form, usually as parsimoniously as possible, the (unknown) underlying process that led to the generation of the data. This contrasts with law-driven models that begin with an understanding of the laws, or 'building blocks', that drive the processes within some system, and explore the behaviour of a more complex process based on some aggregation of the building blocks. In the social sciences the law-driven nature of a model is sometimes encapsulated in the description of the model as a *logic* or *causal* model, i.e. a model as a sequence of events or states linked by causal pathways. This distinction between law-driven and data-driven models has also been described as 'mechanistic vs empirical' (e.g. in the context of a pharmacokinetic model in Nestorov et al., 1999) and in the econometrics literature in terms of 'highly structured vs reduced form' (Cameron and Trivedi, 2005).

Statistical models typically fall within the data-driven category, whereas health economic evaluation models typically fall within the law-driven category. Indeed, health economic models are built because of a *lack* of data on long term costs and health consequences. The aim of the model is to predict costs and consequences under the competing decision options, and the 'laws' that are used to structure the model are the plausible causal links between the various decision options and the costs and health consequences. The law-driven nature of the cost-effectiveness model has important implications for our choice of technique for managing structural uncertainty, as we discuss later.

Many other categorisations of computer models are possible, and even within Saltelli's classification, some models may have aspects of both diagnostic and prognostic, data-driven and law-driven. Climate models are a good example. One reason for the construction of these models is in order for us to make predictions about future global temperature under various greenhouse gas emission scenarios, and in this sense the models are clearly predictive. However, the models also inform our understanding of the highly complex climate system and are therefore also 'diagnostic'. Climate models are law-driven in that they are constructed with reference to well understood physical and chemical laws, but they also have data-driven aspects, usually through some form of calibration. The calibration against data is necessary to allow for gaps in our knowledge of the underlying physics, or our unwillingness (say, for practical reasons) to fully specify a model structure that reflects a highly complex physical system.

3.2.3 Notation

We denote the 'true' unknown values of the vector of quantities that we wish to predict as $\boldsymbol{Z}$. We represent our predictive computer model by the function $\boldsymbol{y} = f(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector of inputs, and $\boldsymbol{y}$ a vectors of outputs in the same units as $\boldsymbol{Z}$. Note that there may be no closed form expression for $f(\cdot)$, for example if the model consists of a set of partial differential equations, or if the model is of an individual level simulation form.

We may be uncertain about some or all the components of $\boldsymbol{x}$ and write the 'true' unknown values of the inputs as $\boldsymbol{X}$. If we are uncertain about the inputs, then this induces uncertainty in the outputs and we write the vector of outputs (conditional on the model f) as $\boldsymbol{Y} = f(\boldsymbol{X})$. We may have additional uncertainty about $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ if the model $f(\cdot)$ is stochastic, or if the model is computationally expensive such that we can only evaluate $f(\cdot)$ a small number of times.

The 'true' value of a component of $\boldsymbol{Z}$ or an input $\boldsymbol{X}$ is not easy to define, and can have a somewhat different meaning depending on whether the quantity has a real physical meaning, or whether it is in the model for the purposes of 'tuning' or 'calibration' (we will discuss the distinction below when we discuss types of inputs). For a quantity that relates to a real physical counterpart the 'true' value is that which would be observed in a perfect study (e.g. for a drug effect, the effect observed in a perfect randomised controlled trial with an infinite number of patients). For a 'tuning' parameter, the meaning of 'true' is less clear, but Kennedy and O'Hagan (2001) define it as that which leads to the best fit to observations against which the model is calibrated. They go on to suggest that in some circumstances inputs that have real physical meanings should be allowed to deviate from their real values (even when known) if this produces a better model fit.

Taking the three components $\boldsymbol{X}$, $\boldsymbol{Y}$ and $f(\cdot)$ in turn, we make the following observations.

3.2.4 Types of inputs

It is often helpful to distinguish different types of input variable in a computer model. We follow here the classification given in Santner et al. (2003). Firstly, there are variables that are set by the owner or user of the model. These variables represent quantities with values that can be chosen at will and are termed 'control' variables. So, for example, in the health economic context such variables would usually include the time horizon of the model, the discount rate and the decision maker's willingness to pay for a unit of health effect. Such variables are not considered uncertain. Our prediction of $\boldsymbol{Z}$ is conditional on the choice of the values of the control variables.

Secondly, a model will usually have inputs that represent real world quantities that are estimated from data, and therefore uncertain. These are termed 'environmental' variables by Santner et al. (2003), but in the context of health economic evaluation are usually called simply 'parameters'. Examples would be the relative risk of some health outcome given some intervention 'A' versus some control treatment 'B', or the number of QALYs associated with living in some health state for some period of time. Within a Bayesian framework we specify a joint probability distribution to reflect our judgements about the uncertain environmental variables.

Santner et al. (2003) recognise a third type of input called a 'model' variable or alternatively a 'tuning parameter'. These variables represent quantities that do not necessarily have a direct real world counterpart, but are used in a model to describe relationships between real world quantities. Unobservable rate constants in a pharmacokinetic model would fall into this class. Tuning parameters are often estimated through a process of *calibration*, in which variables that are functions of the tuning parameters (including perhaps the model output itself) are fitted to observed data.

We distinguish a further type of input that is not strictly within our control (and therefore not a 'control' input), but which we do not necessarily consider as uncertain. Covariates in a individual level simulation model would fall into this class. We do not usually consider the value of a covariate as uncertain, but often use a probability distribution to describe the *variability* of the covariate in our population of interest, something that is outside of our control (this variability is sometimes termed *heterogeneity*). One of the purposes of the individual level simulation is to integrate outcomes with respect to the variability described by this distribution (an integration that may not be possible analytically). Of course, the hyperparameters of the (joint) distribution of the covariates may be considered uncertain, and therefore themselves have an associated (joint) distribution.

3.2.5 Types of outputs

In the health economic context, the output of a cost-effectiveness model is usually a low dimensional vector that consists of a cost and health output under each decision $d = 1, \ldots, D$, for small D. In other contexts, the model output may be either a scalar quantity, or at the other extreme of very high dimension. High dimensional outputs typically occur when the model output relates to a time series, a spatially indexed quantity, or a quantity that is indexed both spatially and temporally. For example, a weather forecasting model may seek to predict hourly temperatures for several days into the future on a 4 km square grid over some area large area of a country or continent.

3.2.6 The form of $f(\cdot)$

We have represented our computer model by the function $f(\cdot)$ that takes inputs $\boldsymbol{x}$ and results in outputs $\boldsymbol{y}$. A computer model is a computer code implementation of some underlying model that has a certain mathematical or logical form. We do not seek to define 'form' rigorously, other than to describe certain broad types of model that are commonly in use.

Firstly, a model can be either deterministic, where each choice of inputs $\boldsymbol{X} = \boldsymbol{x}_i$ results in a unique $\boldsymbol{Y} = \boldsymbol{y}_i$, or stochastic, where input $\boldsymbol{X} = \boldsymbol{x}_i$ induces a distribution $p_i(\boldsymbol{Y})$ in the output. As we noted in chapter 2, there are examples of both deterministic and stochastic models in use in health economic evaluation. 'Cohort based' models that predict population level expected costs and health outcomes tend to be deterministic, whereas models that explicitly include individual level covariates tend to be simulation based and stochastic. In a stochastic model the component of the model output $\boldsymbol{Y}$ that represents population level outcomes is obtained by simulating a large number of individuals, thereby integrating out the covariates that vary across the population. As the number of simulations increases this component of the output $\boldsymbol{Y}$ converges to some value. We can think of these models as being deterministic for the expectation with respect to the individual level variability.

Another important distinction between different model forms is whether or not there are relationships between variables within the model that are defined through differential equations. Differential equation models typically require the use of numerical solvers and lead to approximate results. The choice of solver method and solver resolution will determine the accuracy of the numerical estimation, with greater accuracy usually coming at the expense of greater computational burden. Apart from in the modelling of decisions that relate to interventions for infectious diseases, differential equation type models are relatively rare in health economic evaluation.

Lastly, we make the distinction between models that are 'black boxes' and those that are 'white boxes'. By this we mean to distinguish between models that have a structure that is clearly comprehensible in its entirety (a white box model), and one that is of sufficient complexity that this kind of understanding is impossible (a black box model). Clearly, there is a continuum here rather than a dichotomy. When considering structural uncertainty a white box model implemented in a simple spreadsheet package may allow considerably more scope for 'getting inside' the structure and thinking about how the model differs from reality, than a black box model that is implemented as 100,000 lines of C code.

3.3 Uncertainty

Uncertainty is described by de Finetti (1974) as 'the extent of our own knowledge and ignorance'. Uncertainty exists within the individual, rather than being a property of the world. I own my uncertainty, along with any statements that I make that characterises this uncertainty. There is no 'objective' uncertainty: your uncertainty about some event can be quite different to my uncertainty about the same event. The use of probability as a measure of personal uncertainty goes back to the work of Savage (1954); DeGroot (1970) and de Finetti (1974). They argued strongly for a *subjectivist* interpretation of probability as the single coherent measure of uncertainty. Whether or not there really is a single 'objective' description of uncertainty about an unknown quantity is contentious and has underpinned the long running debate between subjective Bayesian's and the rest of the statistics fraternity (see Gelman, 2008, and associated comments for a flavour of this debate).

Sometimes the distinction is made between *aleatory* and *epistemic* uncertainty. Aleatory uncertainty describes uncertainty that arises from the inherent randomness in a non-determined system, so for example, the toss of a coin generates aleatory uncertainty. Epistemic uncertainty on the other hand describes uncertainty due to lack of knowledge. Once the coin is tossed and has landed, I have epistemic uncertainty about whether that particular coin toss has resulted in heads or tails until I look to see which way up it lies. Although this distinction between aleatory and epistemic uncertainty is sometimes useful, there are philosophical arguments about whether this distinction really exists. If we knew everything, then would there be any randomness? If randomness is merely a description of the behaviour of systems that are not fully determined (i.e. systems for which we do not have complete knowledge) then all uncertainty is surely epistemic (O'Hagan and Oakley, 2004; Nilsen and Aven, 2003).

3.3.1 Uncertainty in the context of computer models

Imagine that an individual wishes to base a decision on some quantity $\boldsymbol{Z}$. We could elicit the individual's judgements about $\boldsymbol{Z}$, and hence specify a distribution $p(\boldsymbol{Z})$. This judgement reflects the individual's 'prior' knowledge about the world, and is entirely coherent, but such probability statements may be of limited use in making the decision. The individual may wish to base her judgements about $\boldsymbol{Z}$, and therefore their decision, on *evidence*. She wishes to quantify $p(\boldsymbol{Z}|\boldsymbol{D})$, the distribution that describes her judgements about the target quantity given some observations on the world, $\boldsymbol{D}$. It is the quantification of this 'posterior' probability that motivates the building of a model. We need some method for linking $\boldsymbol{D}$ to $\boldsymbol{Z}$.

In the context of health economic evaluation our decision maker wishes to determine $\arg\max_d E\{U(\boldsymbol{Z}_d)\}$, where $\boldsymbol{Z}_d$ is the vector of uncertain outputs (costs and health outputs) associated with decision option $d \in \mathcal{D}$. The decision maker's prior beliefs about $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_D)$ is represented by the joint distribution $p(\boldsymbol{Z})$.

The decision maker could choose to make the decision based on her prior beliefs about $\boldsymbol{Z}$. This may potentially lead to two problems. Firstly, if $p(\boldsymbol{Z})$ is such that there is considerable uncertainty as to which decision is optimum, the *expected opportunity loss* could be large.

The expected opportunity loss of choosing decision option d' is defined as

$$EOL = E\{\max_d U(\boldsymbol{Z}_d)\} - E\{U(\boldsymbol{Z}_{d'})\}. \tag{3.1}$$

The first term is the expectation of the random variable $\max_d U(\boldsymbol{Z}_d)$, which is the utility of our optimum decision if we knew the true value of the vector $\boldsymbol{Z}$. The second term is the utility of the decision d'. If our optimal decision under current uncertainty is $d^* = \arg\max_d E\{U(\boldsymbol{Z}_d)\}$ then the expected opportunity loss of choosing this optimal decision based only on prior knowledge is

$$\begin{aligned} EOL &= E\{\max_d U(\boldsymbol{Z}_d)\} - E\{U(\boldsymbol{Z}_{d^*})\}, && (3.2)\\ &= E\{\max_d U(\boldsymbol{Z}_d)\} - \max_d E\{U(\boldsymbol{Z}_d)\}, && (3.3) \end{aligned}$$

which is also the expected value of perfect information (Raiffa, 1968). We will discuss value of information more fully in section §3.4.1.

Secondly, a decision based on the decision maker's prior $p(\boldsymbol{Z})$ may be all well and good for the decision maker herself, if she alone must live with the consequences. However, for decisions that effect others, it will probably be necessary for the decision maker to justify the choice of $p(\boldsymbol{Z})$. She therefore pays a modeller to evaluate the distribution of $\boldsymbol{Z}$, conditional on some evidence, some observations of the world $\boldsymbol{D}$. The primary motivation for building the model then is to link $\boldsymbol{Z}$ to $\boldsymbol{D}$ in order to evaluate $p(\boldsymbol{Z}|\boldsymbol{D})$.

However, in section §3.2.3 we defined the model as $\boldsymbol{Y} = f(\boldsymbol{X})$ not $\boldsymbol{Z} = f(\boldsymbol{D})$, so we need to describe the relationship between the quantity we wish to predict, $\boldsymbol{Z}$, the model, f, and the evidence, $\boldsymbol{D}$. We write

$$p(\boldsymbol{Z}|\boldsymbol{D}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}) p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{D}) p(\boldsymbol{X}|\boldsymbol{D}) d\boldsymbol{Y} d\boldsymbol{X}, \tag{3.4}$$

where we have factorised $p(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{D})$ into the three terms $p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$, $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{D})$ and $p(\boldsymbol{X}|\boldsymbol{D})$.

3.3.2 Sources of uncertainty

The three terms in (3.4) capture three different sources of uncertainty. The term $p(\boldsymbol{X}|\boldsymbol{D})$ is the specification of beliefs about the model inputs $\boldsymbol{X}$ given the observations $\boldsymbol{D}$. The uncertainty in $\boldsymbol{X}$ is called 'input uncertainty', or in the health economic evaluation context, 'parameter uncertainty'.

The term $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{D})$ represents uncertainty about $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ and $\boldsymbol{D}$, though we usually consider that $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{D}|\boldsymbol{X}$ and therefore that $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{D}) = p(\boldsymbol{Y}|\boldsymbol{X})$. If $f(\cdot)$ is deterministic and computationally cheap to evaluate then $\boldsymbol{Y}$ is known immediately that $\boldsymbol{X}$ is known, and $p(\boldsymbol{Y}|\boldsymbol{X})$ reduces to $\delta(\boldsymbol{Y} - f(\boldsymbol{X}))$ where $\delta(\cdot)$ is the Dirac delta function. If $f(\cdot)$ is computationally expensive such that we can evaluate it only a small number of times $\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)\}$, then we have uncertainty about $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ except at $\boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. This source of uncertainty is often termed 'code' uncertainty. These two sources of uncertainty, input and code, are termed *internal uncertainties* by Goldstein (2011).

Finally, the term $p(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{D})$ represents our beliefs about the target quantity $\boldsymbol{Z}$ given the model output $\boldsymbol{Y}$, inputs $\boldsymbol{X}$ and data $\boldsymbol{D}$, though again we usually consider that $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{D}|\boldsymbol{X},\boldsymbol{Y}$ and therefore $p(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{D}) = p(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X})$. If we imagine that we could build a model that directly predicts $\boldsymbol{Z}$, i.e. $\boldsymbol{Z} = f^*(\boldsymbol{X}^*)$ then we can interpret $p(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X})$ as in some way representing uncertainty about the model $f^*(\cdot)$ and on this basis call it *structural* uncertainty. Alternatively, if we write $\boldsymbol{Z} = f(\boldsymbol{X}) + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is the *discrepancy* between the model output and reality, we can consider the uncertainty in $\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X}$ as arising from uncertainty about the unknown *structural error* $\boldsymbol{\delta}$. This source of uncertainty is termed *external uncertainty* by Goldstein (2011).

In the context of health economic evaluation Briggs (2000) describes a fourth source of uncertainty which is termed 'methodological'. This describes uncertainty about *how* to perform the economic analysis and includes choices about the perspective (what outcomes should be included?), the time horizon (what time frame is relevant for this decision?) and the role of discounting (should discounting be applied, and if so, at what rate?). In our view, methodological 'uncertainty' is not really uncertainty, but describes a set of choices that need to be made *a priori* by the decision maker. These are *control variables* in the categorisation in section §3.2.4 rather than uncertain environmental variables or structural choices. The implementation of modelling guidelines, through for example the use of the NICE reference case (NICE, 2008), is an attempt to reduce variation in choices of methodology between decision models.

3.4 Managing uncertainty in computer models

In this section we discuss the 'management' of uncertainty, which we take to include the following: uncertainty quantification, uncertainty propagation, uncertainty analysis and sensitivity analysis. Under the broad heading of 'management of uncertainty' we draw a distinction between those methods that seek to *quantify* uncertainty, and those that seek to reveal the *sensitivity* of a result or prediction or decision to changes in the value of an uncertain input, or differences in assumptions about the structure of the model. We will refer to the former as *uncertainty analysis* and the latter as *sensitivity analysis.* A good introduction to the management of uncertainty in cost-effectiveness analysis is Claxton (2008).

In an uncertainty analysis of a computer model output we wish to quantify the uncertainty in $\boldsymbol{Y} = f(\boldsymbol{X})$ induced by uncertainty about $\boldsymbol{X}$. Unless we believe our model to be perfect, we may also wish to quantify uncertainty about our target quantity of interest, $\boldsymbol{Z}$, given our uncertainty about both $\boldsymbol{X}$ and $f(\cdot)$.

In a sensitivity analysis, we investigate how changes in the model inputs or structure lead to changes in the model output, or to any decision that is made based on the model output. Within the literature on sensitivity analysis as it relates to inputs, a distinction is made between 'local' and 'global' sensitivity analysis (Saltelli et al., 2008). Local sensitivity analysis quantifies the change in the model output due to small perturbations in a model input around some fixed point. Global sensitivity analysis seeks to understand the relationship between changes in the model output and changes in an input across its whole range.

3.4.1 Managing input parameter uncertainty

In this section we discuss the management of uncertainty in the output of the model due to uncertainty in the model inputs. We assume that the model itself is both deterministic and computationally cheap to run, so that given $\boldsymbol{X} = \boldsymbol{x}_i$, we will immediately know $\boldsymbol{Y} = \boldsymbol{y}_i = f(\boldsymbol{x}_i)$

Uncertainty analysis methods

We are uncertain about some or all of the inputs to our model, and represent our beliefs about the inputs via a joint probability distribution $p(\boldsymbol{X})$. Uncertainty in $\boldsymbol{X}$ induces uncertainty in the model output $\boldsymbol{Y} = f(\boldsymbol{X})$, with the joint distribution $p(\boldsymbol{Y})$ representing judgements about $\boldsymbol{Y}$.

The typical approach to deriving $p(\boldsymbol{Y})$ is via Monte Carlo sampling. We draw samples $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ from $p(\boldsymbol{X})$, and then evaluate $\{\boldsymbol{y}_1 = f(\boldsymbol{x_1}), \ldots, \boldsymbol{y}_n = f(\boldsymbol{x_n})\}$, treating this as a sample from $p(\boldsymbol{Y})$. In the health economics literature this is called 'probabilistic sensitivity analysis' (PSA) (Doubilet et al., 1985), and is considered a standard method in the assessment of health technologies (Claxton et al., 2005). As noted in section §2.3, even if we are just interested in the expectation $E(\boldsymbol{Y}) = E_{\boldsymbol{X}}\{f(\boldsymbol{X})\}$ we will need to compute the integral $\int_{\mathcal{X}} f(\boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X}$ if $f(\cdot)$ is non-linear in $\boldsymbol{X}$, or multilinear in $f(\cdot)$ with correlated components of $\boldsymbol{X}$ (Griffin et al., 2006). In the absence of a closed form solution to the integral, the mean of the Monte Carlo 'PSA' sample provides an estimate of this expectation.

Simple Monte Carlo analysis is straightforward if the model is computationally cheap to evaluate. We can obtain any degree of accuracy for any summary statistic based on the sample from $p(\boldsymbol{Y})$ simply by ensuring an adequate number of samples from the input space. If this is not practical due to either the cost of sampling from $p(\boldsymbol{X})$, or more likely the cost of evaluating $f(\cdot)$, then alternative methods must be used.

McKay et al. (1979) show that greater efficiency (i.e. a reduction in the variance of the Monte Carlo estimator) can be gained by employing either a stratified or Latin hypercube scheme when sampling from $p(\boldsymbol{X})$. The gain in efficiency depends on the choice of strata, the form of $p(\boldsymbol{X})$ and the nature of the model. Stein (1987) shows that the gain in efficiency for Latin hypercube sampling depends on the degree of additivity in the model, with efficiency increasing with degree of additivity. Thus, we can in most circumstances gain the same level of accuracy as we would under simple Monte Carlo sampling with fewer model evaluations if we employ stratified or Latin hypercube sampling. If the limitation on the number of model evaluations due to computational cost means that we cannot gain sufficient precision even with these sampling methods then we may have to consider building an *emulator* for the computer model. We discuss this approach in section §3.4.2.

Note that these Monte Carlo based methods (simple, stratified and Latin hypercube sampling) all quantify uncertainty in the model output, $\boldsymbol{Y}$, due to uncertainty in the model input, and *not* our uncertainty about the target quantity $\boldsymbol{Z}$. Alternatively, if we denote the unknown true model that links $\boldsymbol{X}$ to $\boldsymbol{Z}$ as $f^*(\cdot)$ then the sample $f(\boldsymbol{x_1}), \ldots, f(\boldsymbol{x_n})$ represents a sample from the distribution of $\boldsymbol{Z}|f^* = f$. To properly represent uncertainty about $\boldsymbol{Z}$ we must also consider uncertainty in the model structure. However, quantifying uncertainty in model structure is hard since it requires judgements about a model's ability to faithfully represent a complex real life decision problem. We return to this point in section §3.4.3.

Sensitivity analysis methods

Sensitivity analysis seeks to investigate the effects of changes to the inputs on the output of the model, or on the decision made based on the output of the model. 'Local' sensitivity analysis is concerned with determining the effect of small perturbations in some input of interest, usually around some central value. This contrasts with 'global' sensitivity analysis, which is concerned with determining the behaviour of the model output over the whole range of the input of interest. See (Saltelli et al., 2008) for an extensive discussion of model input sensitivity measures.

The simplest form of global sensitivity analysis is to select an single input of interest, and vary it over some plausible range while keeping all other inputs fixed. This 'one way' sensitivity analysis is easy to perform, but may lead to misleading results if inputs are correlated, or if the model is non-additive (Saltelli et al., 2008).

In most cases, a more sophisticated analysis that accounts for correlated inputs and/or non-additive model structure will be required. For a model with a scalar output, $Y = f(\boldsymbol{X})$, the *variance based* sensitivity analysis approach defines the 'main effect' of the input of interest, X_i, on Y as

$$Z_i(X_i) = E(Y|X_i) - E(Y), \tag{3.5}$$

and the 'main effect index' as the variance of the main effect normalised by the total variance of Y,

$$\frac{\text{var}_{X_i}\{Z(X_i)\}}{\text{var}(Y)} = \frac{\text{var}_{X_i}\{E(Y|X_i)\}}{\text{var}(Y)}. \tag{3.6}$$

Given the identity $\text{var}_{X_i}\{E(Y|X_i)\} = \text{var}(Y) - E_{X_i}\{\text{var}(Y|X_i)\}$ the main effect index can be seen to be the expected proportional reduction in variance of Y on learning X_i.

A related quantity is the 'total effect index'

$$\frac{\text{var}(Y) - \text{var}_{\boldsymbol{X}_{-i}}\{E_{X_i}(Y|\boldsymbol{X}_{-i})\}}{\text{var}(Y)} = \frac{E_{\boldsymbol{X}_{-i}}\{\text{var}_{X_i}(Y|\boldsymbol{X}_{-i})\}}{\text{var}(Y)}, \tag{3.7}$$

where $\boldsymbol{X}_{-i}$ is the vector of all inputs *except* X_i. The total effect index is a measure of the overall effect of the input X_i, including any interactions. It is the expected variance (as a proportion of the total variance) that is left when all inputs *except* X_i are fixed. In general, the main effect index is useful in determining the effect of learning a single input, whereas the total effect index is useful in determining non-influential inputs.

The decision theoretic approach to determining sensitivity asks the more general question, what effect does changing X_i have on the *decision* that will be made given the model output, rather than on the model output itself? It is quite possible for an input X_i to be highly influential on the output Y, but to have no influence on the decision.

If we imagine that we will choose a decision option, d from some set of possible options $\mathcal{D}$, based on the value of the model output $\boldsymbol{Y}$ (note that $\boldsymbol{Y}$ is no longer constrained to be scalar as it was above). We have a utility function $U(d, \boldsymbol{Y})$ for each decision d given the model output $\boldsymbol{Y}$, and wish to choose the optimal decision, defined as that which maximises expected utility,

$$d^* = \arg\max_{d\in\mathcal{D}} E\{U(d, \boldsymbol{Y})\} = \arg\max_{d\in\mathcal{D}} E[U\{d, f(\boldsymbol{X})\}]. \tag{3.8}$$

We call the maximised expected utility, $\max_{d\in\mathcal{D}} E\{U(d, \boldsymbol{Y})\}$, the *baseline* utility.

Within this framework, we can determine the *value* (in units of utility) of learning some set of components of the inputs $\boldsymbol{X}_i$. The maximised expected utility given $\boldsymbol{X}_i$ is

$$\max_{d\in\mathcal{D}} E_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}\{U(d, \boldsymbol{Y})\}, \tag{3.9}$$

where $\boldsymbol{X}_{-i}$ is the set of inputs not in $\boldsymbol{X}_i$. The maximised expected utility (3.9) itself is uncertain because $\boldsymbol{X}_i$ is uncertain, and has expectation

$$E_{\boldsymbol{X}_i}[\max_{d\in\mathcal{D}} E_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}\{U(d,\boldsymbol{Y})\}]. \tag{3.10}$$

The expected gain in utility on learning $\boldsymbol{X}_i$ is the difference between (3.10) and the utility at baseline,

$$E_{\boldsymbol{X}_i}[\max_{d\in\mathcal{D}} E_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}\{U(d,\boldsymbol{Y})\}] - \max_{d\in\mathcal{D}} E\{U(d,\boldsymbol{Y})\}. \tag{3.11}$$

This quantity is called the *expected value of perfect information* (EVPI) for $\boldsymbol{X}_i$ (or sometimes the ‘partial’ EVPI, to reflect that we are conditioning on only a subset of the inputs).

We can similarly derive expressions for the expected value of collecting a sample of data, D to inform the estimate of some subset of inputs $\boldsymbol{X}_i$. This is known as the expected value of sample information (EVSI).

It can be shown that the variance based measure is a special case of the EVPI where the decision problem is to estimate Y, and where utility is negative squared error, $U(Y) = -\{Y - d\}^2$ (i.e. quadratic loss). Utility is maximised (loss is minimised) at $d = E(Y)$, so under these conditions the EVPI (3.11) is

$$\begin{aligned}
EVPI &= E_{X_i}[\max_{d\in\mathcal{D}} E_{\boldsymbol{X}_{-i}|X_i}\{U(Y|X_i)\}] - \max_{d\in\mathcal{D}} E\{U(Y)\}, && (3.12)\\
&= E_{X_i}[\max_{d\in\mathcal{D}} -E_{\boldsymbol{X}_{-i}|X_i}\{Y|X_i - d\}^2] - \max_{d\in\mathcal{D}}[-E\{Y - d\}^2], && (3.13)\\
&= E_{X_i}[-E_{\boldsymbol{X}_{-i}|X_i}\{Y|X_i - E(Y|X_i)\}^2] + E\{Y - E(Y)\}^2, && (3.14)\\
&= E_{X_i}\{-\mathrm{var}_{\boldsymbol{X}_{-i}|X_i}(Y|X_i)\} + \mathrm{var}(Y), && (3.15)\\
&= \mathrm{var}_{X_i}\{E_{\boldsymbol{X}_{-i}|X_i}(Y|X_i)\}, && (3.16)
\end{aligned}$$

which is the numerator in (3.6).

Dealing with correlation in the inputs

Calculating the expected value of perfect information (or the main effect index as a special case) requires the computation of the conditional expectation $E_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}(\boldsymbol{Y}|\boldsymbol{X}_i)$ (or alternatively the conditional variance $\text{var}_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}(\boldsymbol{Y}|\boldsymbol{X}_i)$). This is problematic if there is no closed form solution to the expectation and no easy (i.e. computationally cheap) way to sample from the conditional distribution of $\boldsymbol{X}_{-i}|\boldsymbol{X}_i$. Clearly, if $\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{X}_{-i}$ then $E_{\boldsymbol{X}_{-i}|\boldsymbol{X}_i}(\boldsymbol{Y}|\boldsymbol{X}_i)$ reduces to $E_{\boldsymbol{X}_{-i}}(\boldsymbol{Y}|\boldsymbol{X}_i)$, and we only require to sample from the marginal joint distribution of $\boldsymbol{X}_{-i}$. It should always be possible to use a numerical method such as MCMC to sample from the conditional distribution of $\boldsymbol{X}_{-i}|\boldsymbol{X}_i$, but this is likely to be computationally demanding.

A number of solutions to this problem have been proposed. Firstly, Jacques et al. (2006) propose a simple solution whereby the vector of n inputs, $\boldsymbol{X}$, is partitioned into p sub vectors, i.e.

$$\begin{aligned} \boldsymbol{X} &= (X_1, \ldots, X_n) & (3.17)\\ &= (\underbrace{X_1, \ldots, X_{k_1}}_{\boldsymbol{U}_1}, \underbrace{X_{k_1+1}, \ldots, X_{k_1+k_2}}_{\boldsymbol{U}_2}, \ldots, \\ &\quad \underbrace{X_{k_1+k_2+,\ldots,+k_{p-1}+1}, \ldots, X_{k_1+k_2+,\ldots,+k_p}}_{\boldsymbol{U}_p}), & (3.18) \end{aligned}$$

where $k_1 + k_2 +, \ldots, k_p = n$. The partitioning is such that $\boldsymbol{U}_i \perp\!\!\!\perp \boldsymbol{U}_j \ \forall i \neq j$. Sensitivity measures are then calculated for each subset of inputs $\boldsymbol{U}_i,\ i = 1, \ldots, p$, noting that the conditional distribution $E_{\boldsymbol{U}_{-i}|\boldsymbol{U}_i}(\boldsymbol{Y}|\boldsymbol{U}_i) = E_{\boldsymbol{U}_{-i}}(\boldsymbol{Y}|\boldsymbol{U}_i)$. This is a reasonable approach if $\boldsymbol{X}$ can be partitioned in this manner, and if the sensitivity of $\boldsymbol{Y}$ to some subset of inputs $\boldsymbol{U}_i$ is of interest. If not, then this approach fails.

Da Veiga et al. (2009) suggest that for scalar Y and scalar X_i that the conditional moments, $E_{\boldsymbol{X}_{-i}|X_i}(Y|X_i)$ and $\text{var}_{\boldsymbol{X}_{-i}|X_i}(Y|X_i)$ are estimated using a local polynomial regression approach based. Firstly, given a single sample set, Y is regressed on X_i for X_i in the neighbourhood of some point x, assuming the p-th order polynomial regression equation

$$Y_i = \sum_{j=0}^{p} \beta_j (X_i - x)^j + \varepsilon_i. \tag{3.19}$$

The error term ε_i is estimated by a second local polynomial regression, this time a regression on X_i of the squared residuals from the first level of fit. The Da Veiga et al. (2009) method has some similarities to the solution we propose to the problem of correlated inputs which we describe in chapter 8. In our method we effectively regress Y_i on X_i for X_i in the neighbourhood of some x, but assume a simple mean only linear regression $Y_i = \beta_0 + \varepsilon_i$ with $E(\varepsilon_i) = 0$.

In theory the method presented by Da Veiga et al. (2009) is extendible to multidimensional $\boldsymbol{X}_i$, but will suffer from the 'curse of dimensionality'. If n samples are required for some specified precision in the case of a scalar X_i, then for the same precision with $\boldsymbol{X}_i$ of dimensionality d we will require a sample size of order n^d. This exponential growth in the number of samples required severely limits this kind of analysis to problems in which the vector of inputs of interest, $\boldsymbol{X}_i$ is of low dimension.

Bias modelling

So far we have implicitly assumed that the distribution $p(\boldsymbol{X}|\boldsymbol{D})$ represents the decision maker's judgements about the inputs, conditional on some observations of the world (or at least, that the decision maker is happy to accept $p(\boldsymbol{X}|\boldsymbol{D})$ as specified by the authors of the various studies who collected the data and wrote the papers, or perhaps as specified by the modeller who trawled the primary research literature). If for simplicity we take a single input X_i we find that in reality the decision maker may have good reasons for making quite different judgements about $X_i|\boldsymbol{D}_i$, than are specified in the $p(X_i|\boldsymbol{D}_i)$ implied in the published results of the paper(s) reporting $\boldsymbol{D}_i$. There are two reasons for this. Firstly, the decision maker may have concerns about the *internal* validity of the study, and may wish to correct certain important internal biases that may have arisen due to inadequacies in either study design or analysis. Secondly, the decision maker may believe that the circumstances of the study are sufficiently different to those of the decision problem that the resulting *external* biases must also be corrected.

Bias adjustment has received some attention in the literature, and we note the contributions of Eddy et al. (1990, 1992) (the confidence profile method), Spiegelhalter and Best (2003), Greenland (2005) and Turner et al. (2009). In particular, Greenland (2005) and Turner et al. (2009) discuss the important role of prior knowledge in specifying bias terms, given that these terms are typically unidentifiable from the data. The structural discrepancy modelling approach that we propose in chapters 5 and 6 of the thesis draws on these ideas of bias modelling, but instead of making judgements about errors in the inputs, we are concerned with making judgements about errors further 'downstream' within the structure of the model.

3.4.2 Managing code uncertainty

In this section we discuss the management of uncertainty when the computer model is expensive to run. A computationally expensive model adds an extra layer of uncertainty: uncertainty about $\boldsymbol{Y}$ for all values of $\boldsymbol{X} = \boldsymbol{x}$ where we do not have the time or resources to evaluate $\boldsymbol{y} = f(\boldsymbol{x})$.

Increasing the efficiency of the computation

One approach to overcoming the limitations imposed by a computationally expensive model is of course to use a faster computer. In particular, the type of problem we face, that of evaluating a model repeatedly for a large number of inputs sets lends itself to 'embarrassingly parallel' computation. An 'embarrassingly parallel' computation is one that can be easily divided into a number of independent tasks, each of which can be performed on a separate processor (Foster, 1995). In this form of parallel computation there is little or no communication required between processors. Functions to implement embarrassingly parallel computation within R (R Development Core Team, 2011) are readily available, for example within the `snow` and `snowfall` packages.

Alternatively, if the computer model is implemented in high level code, particularly if it is written in an *interpreted* language like R, then a significant gain in speed may be obtained by rewriting the model in a low level compiled language such as C.

Emulators for computationally expensive simulators

In this subsection we refer to our computationally expensive model, $f(\cdot)$, as the *simulator*. Due to the complexity of the simulator, and because we can only run it a small number of times, $\boldsymbol{Y} = f(\boldsymbol{X})$ is unknown for all input value sets except those at which we have evaluated it. One solution in this situation is to build a statistical model for the simulator. This model, which is called an *emulator* for the simulator, is a full joint probability specification of $f(\cdot)$ that allows us not only to determine $f(\boldsymbol{x})$ for any $\boldsymbol{x}$, but also allows us to quantify our uncertainty about $f(\boldsymbol{x})$ due to only having run the simulator a limited number of times (O'Hagan and Oakley, 2004; Oakley, 2011).

We denote the limited number of simulator runs that we are able to obtain as $T = \{\boldsymbol{y}_1 = f(\boldsymbol{x}_1), \ldots, \boldsymbol{y}_N = f(\boldsymbol{x}_N)\}$. We use these as 'training data' to construct an estimate, $\hat{f}(\cdot)$, of the unknown function $f(\cdot)$. An emulator should have the following properties: firstly, it should result in $\hat{f}(\boldsymbol{x}_i) = \boldsymbol{y}_i$ for the 'design' points $\boldsymbol{x}_i \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$; secondly, the mean value for $\hat{f}(\boldsymbol{x})$ should be a plausible interpolation (or extrapolation) of the training data; and thirdly, the distribution around the mean of $\hat{f}(\boldsymbol{x})$ should be a reasonable expression of uncertainty about $f(\boldsymbol{x})$ (Bastos and O'Hagan, 2009). The second and third of these criteria can be checked by comparing additional runs of the simulator with the predicted values, or alternatively by constructing the emulator from only a subset of available simulator runs, and using the remainder for validation.

An established method for modelling an uncertain function $f(\cdot)$ is via the specification of a *Gaussian process*. A Gaussian process (GP) is a stochastic process in which any finite set of samples has a multivariate normal distribution (Grimmett and Stirzaker, 2001). The GP is specified by a *mean function* $m(\cdot)$

and a *covariance function* $c(\cdot,\cdot)$, and we write

$$f(\cdot) \sim GP\{m(\cdot), c(\cdot,\cdot)\}. \tag{3.20}$$

Given a set of inputs, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, for which we have not evaluated the simulator, our uncertainty about the corresponding outputs, $\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)\}$ is described by a multivariate normal distribution with some mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{V}$.

The mean function is typically defined as having the following linear form

$$m(\boldsymbol{x}) = h(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta}, \tag{3.21}$$

where $h(\cdot)$ is a vector of regressor functions called 'basis' functions. The simplest case is $h(\boldsymbol{x}) = 1$, which implies that $m(\boldsymbol{x}) = \beta$ then represents an unknown overall mean for the simulator output. Or, we could define $h(\boldsymbol{x})^{\mathrm{T}} = (1, \boldsymbol{x})$, leading to $m(\boldsymbol{x}) = \beta_1 + \beta_2 x_1 + \ldots + \beta_{1+p} x_p$ where p is the dimensionality of the input vector $\boldsymbol{x}$. This corresponds to a belief that the simulator output has a linear trend with respect to all the inputs. Higher order polynomial forms for $m(\boldsymbol{x})$ can be specified through writing $h(\boldsymbol{x})^{\mathrm{T}} = (1, \boldsymbol{x}, \boldsymbol{x}^2)$, $h(\boldsymbol{x})^{\mathrm{T}} = (1, \boldsymbol{x}, \boldsymbol{x}^2, \boldsymbol{x}^3)$ and so on.

As with the mean function, there are many possible choices for the form of the covariance function, with the most common being the 'Gaussian' form

$$\sigma^2 c(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\{-(\boldsymbol{x} - \boldsymbol{x}')^{\mathrm{T}} \boldsymbol{C} (\boldsymbol{x} - \boldsymbol{x}')\}. \tag{3.22}$$

Here, $\boldsymbol{C}$ is a diagonal matrix, with diagonal elements $\{\omega_1^{-2}, \ldots, \omega_p^{-2}\}$. The hyperparameter $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_p\}$ is known as the *correlation length* and controls the smoothness of the resulting Gaussian process. Prior beliefs about the smoothness of the unknown function $f(\cdot)$ are specified through this hyperparameter.

Analytic solutions to the posterior Gaussian process conditional on training data are shown in Kennedy and O'Hagan (2001) and Oakley and O'Hagan (2002) for certain classes of prior distribution for $\boldsymbol{\beta}$ and σ^2, and a plug-in estimate for $\boldsymbol{\omega}$. A fully Bayesian posterior specification allowing for uncertainty in $\boldsymbol{\omega}$ is obtainable via MCMC, though this is computationally demanding (Neal, 1999).

Given the posterior Gaussian process Oakley and O'Hagan (2004) show how various measures of input uncertainty and sensitivity can be calculated either analytically, or using Monte Carlo methods where the number of runs required of the emulator is much smaller than would have been required of the simulator for the same level of accuracy.

The use of emulation is not widespread in health economic evaluation, probably reflecting the relative simplicity of the models that are typically employed. Notable examples of Gaussian process emulation for cost-effectiveness studies are Tappenden et al. (2004), Stevenson et al. (2004) and Rojnik and Naversnik (2008).

Specifying beliefs about the simulator via a Gaussian process is not the only route to building an emulator. An alternative approach, which only requires the expression of beliefs about the first two moments of the uncertain simulator output, is to build a *Bayes linear* emulator (Craig et al., 2001; Goldstein and Rougier, 2006, 2009; Goldstein, 2011). The Bayes linear approach has a somewhat different underpinning philosophy from the fully Bayesian approach in that prior belief statements are made as *expectations* rather than as *probabilities*. Treating expectation as the primitive quantity avoids the need to make a full joint probabilistic specification over the unknowns, which can be an extremely daunting task. Updating the expectation of some random quantities B (about which we specify prior second order beliefs) given some observations D generates an *adjusted* expectation (in the Bayes linear terminology). For each element B_i in B the adjusted expectation, written $E_D[B_i]$ is the linear combination $\mathbf{a}_i^T D$ that minimises $E[(B_i - \mathbf{a}_i^T D)^2]$. This leads to expressions for the adjusted expectation for B,

$$E_D[B] = E[B] + \text{cov}[B, D]\text{var}[D]^{-1}(D - E[D]), \tag{3.23}$$

and adjusted variance

$$\text{var}_D[B] = \text{var}[B] - \text{cov}[B, D]\text{var}[D]^{-1}\text{cov}[D, B]. \tag{3.24}$$

The adjusted expectation, $E_D[B]$, is an approximation to the conditional expectation $E(B|D)$ that would be obtained under a full Bayesian prior-to-posterior analysis. The adjusted variance $\text{var}_D[B]$ provides an upper bound for the conditional variance $\text{var}(B|D)$. The approximations are exact in the case where the joint probability distribution of B and D is multivariate normal.

3.4.3 Managing structural uncertainty

We now discuss the management of uncertainty about the target quantity $\boldsymbol{Z}$ due to our uncertainty about the 'true' model. A law driven model can be thought of as a representation of judgements about the relationship between the model inputs and the model outputs. If we are uncertain what this 'true' structural relationship is, then even if we were to run the model at its true inputs, there would be an uncertain 'structural error' in the model prediction. We denote this uncertain error 'structural uncertainty'. Note that we use the term 'true' value of the input to mean that which we would estimate in some perfect study with infinite sample size, and 'true' structural relationship to mean a (possibly non-unique) functional relationship that would result in the correct output given any set of 'true' values of the inputs.

Unless we are able to build a model that is true in the sense above we should expect an uncertain structural error in the model prediction. What are our judgements about this error? Given that we are likely to be uncertain about some or all of the model inputs is the uncertainty about the model structure important? Is the imperfect model good enough for the decision that we will base on the model output? If not, which bit of the model is inadequate? In the words of Box (1976)

> "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad".

Quantifying structural uncertainty is, however, difficult since it involves making judgements about a model's ability to faithfully represent a (possibly highly complex) real life system.

We note two important distinctions that are made in the treatment of model uncertainty. Firstly, is the focus of the uncertainty analysis the correctness of the model, or the discrepancy between the model output and reality? Historically, the statistical treatment of model uncertainty has located the uncertainty in the *model* itself, explicitly seeking to determine the adequacy of the model (usually within the context of a set of competing models). This contrasts with the treatment of the problem in the computer models literature, where the uncertain *model error* is primarily of interest. A key concept in this latter treatment is that of *discrepancy*: the difference between the model run at its 'best' or 'true' input, and the true value of the output quantity (Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2009).

The second distinction we make is whether or not there are observations on the target quantity $\boldsymbol{Z}$. For a statistical model the answer is yes, since for statistical models the data that are modelled *are* the observations on $\boldsymbol{Z}$. These observations are the beginning point for a statistical model, and as we have noted before, such models are therefore termed *data-driven*. However, for law-driven models we may be in a position where there are no observations on the target quantity. Certainly, in the context of health economic evaluation, at the point of making a decision we have no direct observations of the costs and health effects for each of our decision options. However, this lack of data is not the case in all situations in which law-driven models are used. In physical systems modelling for example, the target quantity is often partitioned as $\boldsymbol{Z} = \{\boldsymbol{Z_o}, \boldsymbol{Z_u}\}$, where we have (noisy) observations $\boldsymbol{z}$ on $\boldsymbol{Z_o}$, but no observations on $\boldsymbol{Z_u}$. We may have historic observations and wish to predict future values (forecasting), or we may have observations at a set of points in space and wish to predict values at locations in between (interpolation). Of course, the problem of defining the 'correct' model structure arises whether or not we have observations on $\boldsymbol{Z}$, rather, the distinction is important since it determines whether we can use the likelihood (or similar measure) of the observations on $\boldsymbol{Z}$ to update beliefs about the *whole* model structure (we will elaborate on this when we discuss model averaging).

Alternative scenario analysis

A basic form of structural sensitivity analysis is to explore the sensitivity of the model prediction to underlying assumptions in a 'what if' scenario analysis in which sets of alternative assumptions are modelled (see Bojke et al. (2009) for a review of the methods that are currently used to manage health economic evaluation model uncertainty, and Kim et al. (2010) for a specific example of modelling alternative scenarios). However, this process cannot in any formal sense quantify the sensitivity of the results to the assumptions, and nor can it quantify any resulting prediction uncertainty. If a decision maker is presented with the results of a number of different models, with each model having a different set of structural assumptions, it is not clear how she should proceed with her decision. What should her posterior beliefs about $\boldsymbol{Z}$ be?

Model averaging

Model uncertainty has been addressed from a statistical perspective (e.g. Draper, 1995; Kass and Raftery, 1995). Here, a key concept is that of *model averaging* in which the predictions or probability statements of a number of plausible models are averaged, with weights based either on some measure of model adequacy, or some measure of the probability that the statistical model is 'correct'. Bernardo and Smith (1994) and Kadane and Lazar (2004) offer a Bayesian decision theoretic treatment of model averaging in the general statistical context, and Jackson et al. (2009, 2010) and Bojke et al. (2009) illustrate its application to health economic evaluation.

Suppose we have a set of plausible models $\{M_i, i \in I\}$, with $M_i = \{f_i(\boldsymbol{X}), p_i(\boldsymbol{X})\}$. We draw on Bernardo and Smith (1994)'s notion of describing a set of possible models as $\mathcal{M}-$*closed* or $\mathcal{M}-$*open*. A set of models, $\{M_i, i \in I\}$, is described as $\mathcal{M}-$*closed* if we believe that one of the models in $\{M_i, i \in I\}$ is 'true', but we do not know which. Conversely, a set of models is described as $\mathcal{M}-$ *open* if we do not believe that one of the models in $\{M_i, i \in I\}$ is correct. In model averaging, we predict $\boldsymbol{Z}$ using a weighted mean value of the individual model outputs. The weighting process could simply consist of choosing the model from the set that we believe is ‘best’ while discarding the rest, effectively placing all the weight on a single model. Or, we may want to more formally assess our beliefs about how likely the different models are, and weight the outputs by these probabilities.

If we have data, $\boldsymbol{D}$, (that have not been used to inform the inputs $\boldsymbol{X}$) and can calculate some measure of the adequacy of the model, given $\boldsymbol{D}$, then we can weight the model outputs by (some function of) the adequacy measure. If we believe our set of models is $\mathcal{M}-\textit{closed}$ then, within a Bayesian framework, we can specify prior model probabilities, $p(M_i)$, and calculate the posterior probabilities given $\boldsymbol{D}$ via

$$p(M_i|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|M_i)p(M_i)}{\sum_{i\in I} p(\boldsymbol{D}|M_i)p(M_i)}, \tag{3.25}$$

leading to a weighted mean output

$$p(\boldsymbol{Z}|\boldsymbol{D}) = \sum_{i\in I} p(\boldsymbol{Z}|M_i, \boldsymbol{D})p(M_i|\boldsymbol{D}). \tag{3.26}$$

Where there are two or more conflicting, defensible models, then such an approach has obvious benefits, in comparison with using one model only and ignoring the others. However, model averaging approaches are unlikely to be sufficient for fully quantifying uncertainty about $\boldsymbol{Z}$, and do have some practical limitations.

The obvious shortcoming is that we will not usually believe any of the models are correct: we do not believe that $f_i(\boldsymbol{X}) = \boldsymbol{Z}$ for any i. Jackson et al. (2010) acknowledge this, and instead of using (3.25) to obtain weights for use in (3.26) adopt an $\mathcal{M}-\textit{open}$ view and derive a weight $p(M_i|\boldsymbol{D})$ based on the probability that model M_i “gives the best predictions on a replicated data set, among the models being compared”. However, even with this weighting scheme, why should we believe that a weighted average of the outputs from an $\mathcal{M}-\textit{open}$ set of models (i.e. a set of models that are all wrong) represents our posterior beliefs about $\boldsymbol{Z}$?

Another important limitation is the form of the available data, $\boldsymbol{D}$. Model averaging approaches involve evaluation of a likelihood function for each model:

$$p(\boldsymbol{D}|M_i) = \int_{\mathcal{X}} p_i\{\boldsymbol{D}|f_i(\boldsymbol{X})\}p_i(\boldsymbol{X})d\boldsymbol{X}. \tag{3.27}$$

However, in many applications, health economic evaluation included, we do not have data $\boldsymbol{D}$. There are no observations on the model output $f_i(\boldsymbol{X})$, as this would have been the reason for building the model in the first place. We may, however, have some relevant data, $\boldsymbol{D}^*$. For example, in the health economics context, $f_i(\boldsymbol{X})$ may correspond to mean costs and benefits over a twenty year period, and $\boldsymbol{D}^*$ may be observations of the treatment efficacy in a clinical trial over a two year period. If we imagine that f_i contains a 'sub-function', g_i, that describes efficacy at two years, then we would have

$$p(\boldsymbol{D}^*|M_i) = \int_{\mathcal{X}} p\{\boldsymbol{D}^*|g_i(\boldsymbol{X})\}p(\boldsymbol{X})d\boldsymbol{X} \tag{3.28}$$

which we could plug into (3.25) to get (3.26).

This is helpful, but not sufficient. All those elements that differ between models and that are downstream of the sub-function $g_i(\cdot)$ (and which therefore may lead to different predictions of $\boldsymbol{Z}$) remain untested. These methods clearly have a role in guiding structural choices for parts of the model where intermediate outputs can be fitted to data, but they cannot guide choices about the whole model structure since we do not observe future costs and health effects under each of our competing decision options.

Finally, we may wish to consider competing models in the absence of any data that would inform the choice between them. In this latter case we are left with just our prior model probabilities $p(M_i)$. If we were to adopt an $\mathcal{M} - closed$ view then our posterior beliefs about $\boldsymbol{Z}$ would be

$$p(\boldsymbol{Z}) = \sum_{i \in I} p(\boldsymbol{Z}|M_i)p(M_i). \tag{3.29}$$

Learning about $p(M_i)$ might now be considered to be an expert elicitation problem, the extreme example being that of a modeller choosing a single model because they believe it to be 'best'. Elicitation of prior model probabilities is not, however, a trivial problem. For example, how would an expert decide how much probability to place on two competing Markov models, one with three health states, and one with four?

More likely is the situation in which we have a set of plausible models, none of which we believe to be 'true', *and* we have no observations to guide model selection. In this situation even the notion of the prior model probability is problematic, since by assuming that our model set is $\mathcal{M} - open$ we have already effectively stated that $p(M_i) = 0 \; \forall i$.

Discrepancy based approaches

A fundamentally different approach to quantifying structural uncertainty is instead to represent our uncertainty about model structure through our judgements about the *discrepancy* between the model output and the 'true' quantities we wish to make statements about. Rather than consider some measure of the 'correctness' of our model, we instead make judgements about the structural error that arises from its imperfection. Important papers demonstrating the approach include Kennedy and O'Hagan (2001); Craig et al. (2001); Higdon et al. (2005) and Goldstein and Rougier (2009).

In the model discrepancy approach to structural uncertainty we focus on the discrepancy, $\boldsymbol{\delta}$, between the output of a model evaluated at the 'true' inputs, and the true target value,

$$\boldsymbol{Z} = f(\boldsymbol{X}) + \boldsymbol{\delta}. \tag{3.30}$$

The discrepancy term, $\boldsymbol{\delta}$, quantifies the *structural error*: the difference between the output of the model evaluated at its true inputs and the true target quantity. Instead of specifying model weights, the key task is now to usefully quantify our beliefs about the discrepancy via $p(\boldsymbol{X}, \boldsymbol{\delta})$. We are explicitly recognising in equation (3.30) that our model may be deficient, but note that when we speak about model deficiency we are not concerned with mistakes, 'slips', 'lapses' or other errors of implementation (for a discussion on this topic see Chilcott et al., 2010b). Rather, we are concerned with deficiencies arising as a result of the gap between our model of reality, and reality itself. Obtaining a joint distribution that reflects our beliefs about inputs and discrepancies, $p(\boldsymbol{X}, \boldsymbol{\delta})$, allows us then to fully quantify our uncertainty in the target quantity due to both uncertain inputs and uncertain structure. This approach has the important advantage that only a single model need be built, though methods for making inferences about discrepancy in the context of multiple models have also been explored (Goldstein and Rougier, 2009).

In the situation where we have partial, noisy, observations $\boldsymbol{z}$ on $\boldsymbol{Z}$, Kennedy and O'Hagan (2001) propose a method for fully accounting for the uncertainty in $\boldsymbol{Z}$, given $\boldsymbol{z}$ and uncertain inputs $\boldsymbol{X}$. They begin by specifying a Gaussian process for the model function $f(\cdot)$, and then specify a second Gaussian process for the model discrepancy $\boldsymbol{\delta}$. Prior beliefs about the model and model error are incorporated via the hyperprior terms in the Gaussian process specifications, and observations (both $\boldsymbol{z}$, and the output from the 'simulator' $f(\cdot)$ training runs) are then used to update beliefs within a Bayesian framework.

Goldstein and Rougier address the same problem, but from the Bayes linear perspective, for the case when there is a single simulator (Goldstein and Rougier, 2006) and when there are multiple simulators (Goldstein and Rougier, 2004, 2009).

The difficulty in health economics with any of the discrepancy approaches described above is the lack of observations on the target quantity $\boldsymbol{Z}$. We do not directly measure the costs and health consequences of sets of competing decisions, and calibration of the model output ($\boldsymbol{Y}$) against data (observations on $\boldsymbol{Z}$) is therefore usually not possible. In this case we are left with expert judgement as the means by which to make statements about the model discrepancy $\boldsymbol{\delta}$. In theory, the modeller at least should be a able to make some judgement about the model error, if only a very crude one.

Imagine that a modeller builds a model to predict the incremental net benefit of recommending drug A versus the current best alternative (*incremental* net benefit here is the difference between the net benefits of the two decision options). The model output suggests the expected incremental net benefit, given uncertainty in all the inputs, is $E(Y) = £500$ (per person treated).

When asked to estimate the model error (i.e. the difference between the true incremental net benefit, Z, and the model output Y) it is likely that the modeller will at least be able to specify a distribution that is less vague than figure 3.2a.

They might believe that their model will underestimate the true value and specify something like figure 3.2b. Or perhaps the modeller has reasonable confidence in the model, with figure 3.2c representing their beliefs. Of course, in reality, the implied distribution on model error in most cases is figure 3.2d. We think that in general, it is unlikely that a crude evaluation of the model error will be particularly useful or robust to criticism.

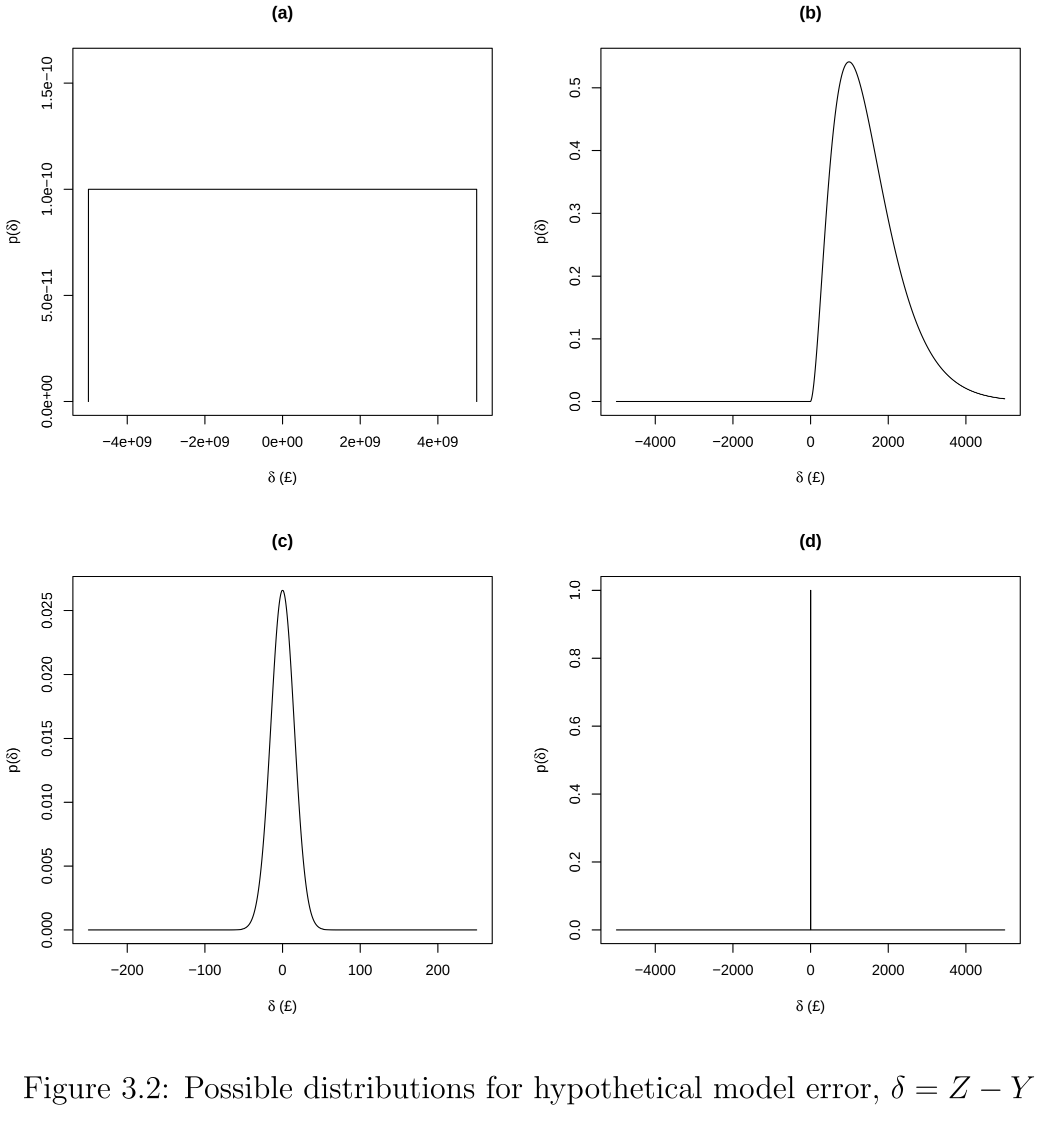


Figure 3.2: Possible distributions for hypothetical model error, $\delta = Z - Y$

Zio and Apostolakis (1996) suggested a possible solution to this problem when a set of plausible models is available, one of which is considered by experts to be the 'best'. The case study the authors present concerns a set of six plausible predictive models for ground water flow in the context of a radioactive waste repository safety analysis. In the discrepancy analysis it is assumed that the structural error for the best model is normally distributed with expectation equal to the mean of the differences in output between this best model and the other five models in the set (termed 'residuals'), and variance equal to the sample variance of the residuals. In a second analysis each model residual was weighted by the probability that the corresponding model provided an 'appropriate description of the ground water phenomenon for the objective of the analysis'. This does beg the question: how good does a model have to be to be 'appropriate' for some analysis?

Zio and Apostolakis (1996)'s empirical approach to specifying the discrepancy is attractive in that it directly incorporates information from all six models, but it excludes any prior information that we may have, say, about a source of error that is common to the whole set of models. What if we were faced with six models that were similar in structure, with similar outputs, but were all very wrong?

3.5 Constructing a single framework

We are uncertain about the (comparative) net benefits of range of decision options. Our beliefs about these unknown net benefits are informed by 'data' that we model statistically, but also by our ideas about the causal relationships ('laws') that underlie the order that we observe in the world. We have a range of techniques (parameter uncertainty analysis, bias modelling, emulation for expensive simulators, model averaging, discrepancy analysis) that can help us make statements about the unknown net benefits that are consistent with our knowledge of the world, and that properly reflect our true uncertainty. None of the approaches for managing uncertainty that we have discussed are mutually exclusive, and all can coexist within a single modelling framework. More specifically, the techniques that relate to ‘structural’ uncertainty, model averaging and model discrepancy, can comfortably sit together. Both rely on ‘augmenting’ an existing model or a set of existing models to construct a larger augmented model or ‘meta-model’[1] that has additional parameters (model weights in the case of model averaging, and discrepancy terms in the case of discrepancy analysis). Indeed, if we consider our problem as determining $p(\boldsymbol{Z}|\boldsymbol{D})$ rather than $p(\boldsymbol{Z}|\boldsymbol{X})$ then we might envisage a single ‘augmented’ model within which are located bias modelling parameters *and* discrepancy terms *and* sub-model weights.

3.6 Conclusion

In this chapter we have reviewed the management of uncertainty in computer models. We have determined the sources of our uncertainty about the target quantity given the model we have built, and have considered some of the methods available for managing these uncertainties. We have argued that the problem of model uncertainty in health economic evaluation is particularly difficult because we do not observe net benefits. We cannot assess the ‘fit’ of our model output to data. This makes both likelihood based methods (model averaging) and model calibration based methods (e.g. as proposed by Kennedy and O’Hagan, 2001) challenging. Given an absence of data it is always possible for a Bayesian to turn to expert judgement, but even this may be of little use when making judgements about the net benefits of some set of decision options conditional on the output of a health economic evaluation model.

In chapters 5 and 6 we return to the problem of specifying model discrepancy in the absence of observations on the model output. We will suggest that instead of making judgements about the discrepancy at the level of the model output, that we ‘open up’ the model and make judgements about model error at the level of the revealed intermediate parameters *within* the model.

[1] The term ‘meta-model’ is also used to denote a statistical emulator for a computationally expensive simulator, so to avoid confusion we will use the term ‘augmented model’ instead

Chapter 4

Elicitation

4.1 Introduction

This chapter is about elicitation, the process of assisting an expert to express his or her knowledge about an unknown quantity in a probabilistic form. We saw in the last chapter that, in the context of health economic model uncertainty, we are likely to be in a position of having to make statements about model error in the absence of data. Hence, we briefly review some of the relevant theory that underpins elicitation, describe methods that have been established to conduct the process, and discuss examples of the use of elicitation in health services and medical research.

We end the chapter with a discussion of some of the challenges that we are likely to face when using elicitation to derive distributions in the context of model structural uncertainty.

4.2 What is elicitation?

Elicitation is the process of extracting expert knowledge about an uncertain quantity (or quantities), and representing this knowledge as a (joint) probability distribution for the unknown(s) (O'Hagan et al., 2006). We recognise that in many decision problems experts have relevant information that either augments the information contained within some data, or provides helpful statements about unknowns in the absence of data. Elicited information has a particular role in forming the prior distribution for the purposes of doing Bayesian inference. Health economic evaluation is inherently Bayesian because we wish to make probabilistic statements about unknown quantities (costs and health effects) that represent our beliefs about those quantities for the purposes of making some decision. Elicitation fits naturally within this framework as a tool for formalising the process of incorporating judgements into the decision making process (Stevens and O'Hagan, 2002).

Elicitation has been used in a wide range of applications including engineering reliability (Bedford et al., 2006) accident risk analysis for nuclear power plants (Cooke and Goossens, 2000), reliability of nuclear weapons (Wilson et al., 2011), water industry planning (Garthwaite and O'Hagan, 2000), agricultural land management (Orton et al., 2011), climate science (Rougier, 2007; Dijkstra and Dixon, 2010), ecology (Kuhnert et al., 2010), including future polar bear populations under climate change (O'Neill et al., 2008), seismic hazard (Klügel, 2008) and the probability that unexploded ordnance will explode (MacDonald et al., 2008). Quantifying very small risks for events that cause considerable levels of public concern is a particularly fraught area for policy makers and elicitation has been used here to assess the health impact of chemicals in food (van der Voet et al., 2009), and the health risk of nano particles (Kandlikar et al., 2007). We note some of the many other applications in health care and medical research in section §4.4.

4.3 How does elicitation work?

There has been a considerable volume of research directed at establishing robust methods for elicitation. Some important papers are O'Hagan (1998); Garthwaite et al. (2005) and the recent review by Johnson et al. (2010). A comprehensive text on elicitation is O'Hagan et al. (2006). Software to assist the elicitation process and to allow the fitting of probability distributions is available[1] along with accompanying tutorials for univariate (Oakley, 2010) and multivariate cases

[1] http://www.tonyohagan.co.uk/shelf/

(Daneshkhah and Oakley, 2010).

The basic idea of elicitation is as follows. There are four 'actors': the *decision maker*, the *expert*, the *statistician* and the *facilitator* (although it is common for the last two roles to be played by the same person). We introduced the decision maker in chapter 2. This is the individual who requires the probability distribution for the purposes of making some decision. The expert is the individual who is deemed to have useful subject matter knowledge, and can therefore provide meaningful statements about the uncertain quantities in question (she[2] needs not be an 'expert' in any more formal way than this). The elicitation process involves elements of training in the basic ideas of probability, and the validation of results, both of which require a 'statistician' (again, not necessarily in a formal sense). Finally the facilitator is the expert in the application of the elicitation process who manages the dialogue with the expert.

Sometimes the term 'analyst' is used to describe the combined statistician and facilitator role taken by a single individual, and in some circumstances the expert, statistician and facilitator are the same person. In this latter case we refer to the resulting process as self-elicitation. We note here that determining one's own probabilities is not easy, vagueness being the major obstacle (Savage, 1971). This may be relevant later when we discuss the use of elicitation by modellers with the aim of improving their own models.

The elicitation process involves a number of steps. First the uncertain quantities are identified. This sounds trivial, but it is extremely important that the expert is asked to provide information about the quantity or quantities that really are required for the inference or decision. Next the expert is identified and recruited. The elicitation session itself begins with a careful explanation of the process and training in the basic ideas of probability. After a process of *calibration* (which we discuss below) the expert is asked to make a series of statements about the unknown quantity or quantities that reveal aspects of her underlying subjective distribution. The facilitator or statistician fits a probability distribution to these summaries. Through an iterative process of feedback and checking

[2] We adopt the convention of referring to the expert as 'she' and the facilitator as 'he'.

the elicited distribution is refined until the expert is confident that it reflects her judgements.

In the following sections we briefly review some specific issues involved in the elicitation process.

4.3.1 Probabilities versus distributions

Firstly, we make a distinction between the elicitation of distributions versus probabilities. An expert may be asked to make a statement about some quantity, for example the relative risk of death given drug A versus drug B, and here the use of a probability distribution to represent the expert's beliefs about the relative risk is clear. However, if the expert is asked to make a statement concerning a single *probability* ("what is the probability that it will rain tomorrow?") then it is sometimes helpful to consider the probability as representing a long run frequency or proportion, about which the expert can express beliefs via a probability distribution.

4.3.2 Calibration

It is helpful to have some measure of the 'quality' of an expert's ability to make probabilistic statements about unknowns. We may wish, for example, to weight each of a group of experts according to how accurate they are in judging uncertain quantities. Calibration refers to the process of comparing an expert's statements with reality. The expert is asked to make a series of judgements about unrelated quantities and then for each judgement a probability of some event is extracted (e.g $P(a < \theta < b)$, where θ is the unknown quantity of interest). This results in a set of probabilities. For the set of judgements concerning events in which we have determined some event probability p we then compare p with the actual relative frequency of those events. If the proportion of events is equal to p then the expert is perfectly calibrated. Of note is the common finding that experts are *overconfident* (for a set of events assigned probability 90%, only 60% occur), and/or that they exhibit *over-extremity* (on low frequency events they place too low a probability, and on high frequency events they place too high a probability).

4.3.3 Specific methods for eliciting responses

There are a number of different methods for eliciting judgements that require the expert to make statements about different, but related quantities. We will not discuss this in detail, other than to note that since we are deriving a series of statements of the form $P(a < \theta < b)$, methods can focus on judging P given a and b (fixed interval methods), or focus on judging b (or a) given P and b (or a) (variable interval methods).

4.3.4 Fitting a distribution

Once a series of statements of the form $P(a < \theta < b)$ for different a and b are made then a probability distribution can be 'fitted' which reflects those statements. In theory an infinite number of statements of the form $P(a < \theta < b)$ will need to be made to properly determine the distribution of a continuous variable θ, and this is one of the motivations for the Bayes linear approach in which only the first two moments are specified (Goldstein, 1992). However, in practice, we can usually assume that beliefs are represented by smooth (usually uni-modal) distributions, allowing a fully Bayesian approach.

If we choose to use a parametric distribution to represent the beliefs, then we will be guided in choice of distribution by the nature of the uncertain quantity, whether unrestricted over the whole real line, strictly non-negative or positive, or bounded. Respectively we may choose for example, a normal, a gamma, and a beta distribution to represent the experts probability judgements. Once the distribution form is chosen, it is 'fitted' to the probability statements obtained from the expert. The fit will be exact if the expert's probability statements uniquely determine a single distribution. If we have elicited more probability statements than required to determine a unique distribution, then the fit can be based on a method such as least squares (this is called overfitting, which we discuss below).

4.3.5 Obtaining a 'good' elicited distribution: feedback and overfitting

It is important to check with the expert that the distribution that is fitted following the initial elicitation does really represent her judgements. This can be done by reporting back to the expert some of the implications of the fitted distribution, a process known as *feedback*. For example, if a $N(5, 1)$ distribution is fitted to reflect judgements about some unknown quantity θ, then the expert could be informed that this implies that there is approximately only a 5% probability that the value of θ lies outside the interval $(3, 7)$. The expert then has a chance to review her initial statements if the implied distribution does not fit her beliefs about θ.

Overfitting is a somewhat different approach to obtaining a 'good' representation of the expert's beliefs. Here, the expert is asked to make more statements that are necessary to fit the chosen form of parametric distribution. This (usually) then results in a series of statements that are incompatible with a single distribution. Fitting is then an optimisation problem whereby a single distribution is chosen that best fits the range of statements made by the expert (e.g. via minimising a sum of squared differences, or sum of absolute differences). The resulting distribution is based on a greater number of pieces of information obtained from the expert than in the simple approach, and as such might be expected to be a better representation of her beliefs. Overfitting also allows the checking of 'residuals', the differences between the expert's probability statements and those implied by the fitted distribution. The presence of large residuals may imply either that the distribution form does not reflect the expert's beliefs very well, or that there are inconsistencies in the expert's statements. In either case, further discussion with the expert is necessary to ensure an adequate elicitation of beliefs.

4.3.6 Multivariate distributions

Elicitation of beliefs about independent quantities is hard. Elicitation of beliefs about dependent quantities as even harder. Imagine we wish to learn the values of two quantities: the efficacy of drug A relative to placebo (as a relative risk), and the proportion of patients taking the drug who will experience adverse effects (again compared to placebo). We have reason to believe that there is a certain action of the drug that causes both a component of the therapeutic effect and an adverse effect. For the purposes of the elicitation the two uncertain quantities are therefore not independent, and we must represent beliefs by a multivariate distribution of some kind.

Immediately we envisage two problems. Firstly, how do we elicit beliefs about dependent quantities, and secondly, how do we choose a distribution that reflects these beliefs? Important issues include the choice whether to elicit joint summaries or conditional summaries, and whether or not it is helpful to elicit second moments or correlation coefficients directly (this can be difficult, see Kadane and Wolfson, 1998; Garthwaite et al., 2005).

There has been significant work done to formalise methods for elicitation of the parameters of a multivariate normal distribution (e.g. Garthwaite and Al-Awadhi, 2001), and for the parameters of regression models (both linear and GLM, see e.g. Kadane et al., 1980; Bedrick et al., 1996), but the research question remains very much open. Moala and O'Hagan (2010) propose a general non-parametric approach in which the expert's unknown multivariate density function is modelled using a vague Gaussian process prior. The expert is asked to provide certain (mainly marginal, but with some joint) summary quantities (probabilities or quantiles, or perhaps means, etc). These are treated as data and a Bayesian update performed to derive a posterior multivariate distribution that represents the expert's beliefs. In chapter 6 we describe a similar Gaussian process based approach for representing beliefs about the model error in a Markov model.

4.3.7 Uncertainty about uncertainty? Imprecision and the need for sensitivity analysis

An expert can not precisely state her subjective probability for some event. It is not reasonable to expect that she can meaningfully make a statement that her probability is 0.1 rather than 0.11, or 0.1002342 say. Elicitation is not precise.

And, even if the expert were able to make such precise statements, there would still be the problem of fitting the *unique* distribution that properly represents her beliefs across the whole range of the uncertain quantity in question. For an expert to specify her 'true' distribution for a continuous quantity would require to be able to specify perfectly an infinite number of probabilities. This is not feasible, so, we have epistemic uncertainty about the expert's beliefs.

In order to avoid the infinite regress of specifying uncertainty about uncertainty, O'Hagan and Oakley (2004) suggest that we consider the epistemic uncertainty about the expert's distribution as being owned by the facilitator (or 'analyst'). The facilitator expresses his own prior beliefs about the expert's density function, which are then updated by the expert's beliefs in a formal Bayesian analysis (see Oakley and O'Hagan, 2007). This approach implies that it is the analyst's posterior uncertainty that is of interest, and that the expert is treated as a source of (noisy) data. Of course, in reality, it is the decision maker's distribution that will inform the decision. The decision maker may accept the distribution that results from the elicitation exercise and adopt it as their own if they have very weak prior knowledge (or perhaps more likely, they accept the elicited distribution if it does not conflict with the distribution they, in some informal sense, elicit from themselves). This approach of treating the expert's elicited summaries as data does not entirely solve the imprecision problem since the facilitator's (or decision maker's) prior distribution is only an imprecise representation of their own uncertainty. The difference is that, in the case of the decision maker, they own the problem, and therefore choose in some sense to live with the consequences of the imprecise nature of their own probability statements.

A more informal approach to the problem of elicitation imprecision is to take a range of distributions that may reflect the expert's beliefs and determine the sensitivity of the resulting inference or decision to changes in the uncertainty specification. If the choice does not materially affect the outcome, then this is reassuring.

4.4 Elicitation in health services and medical research

Formally elicited expert knowledge has been used in a diverse range of settings within health services and medical research. We note the following practical applications of elicitation. In drug trials it has been used to inform patient allocation in a clinical trial (Kadane, 1994); sample size (Oremus et al., 2002); dosage schedules and dose response in early phase trials (Morita, 2011; Zohar et al., 2011); and cancer survival in a phase 3 trial (Hiance et al., 2009). Elicitation has also been used in the meta analysis of trial results to augment missing data (White et al., 2008), and to 'bias adjust' studies (Turner et al., 2009).

Within the field of epidemiology, elicitation has been used to make judgements about the outcomes of patients for whom follow-up data are incomplete (Shardell et al., 2008; Paddock and Ebener, 2009), as well as in the context of studies that assess the health benefits of air quality control (Kinney et al., 2010). In the occupational health setting elicitation has been used to assess the probability of nickel exposure in the workplace (Ramachandran et al., 2003), and in health services management, the risk of clinical untoward events in a hospital pharmacy department (Cagliano et al., 2011).

4.4.1 Elicitation in cost-effectiveness studies

There are very few reports in the literature of the use of elicitation to inform parameters specifically for the purposes of a cost-effectiveness analysis[3] despite Stevens and O'Hagan (2002)'s call for its adoption. Five years later Leal et al. (2007) noted the lack of the use of formal elicitation methods in health economic evaluation studies, and developed a practical computer based elicitation tool with the aim of introducing theoretically sound methods into practice. The tool was

[3] We note that in the context of health economic evaluation the term elicitation is commonly used to describe the process of obtaining health state preferences (e.g. Ryan et al., 2001). This is not elicitation in the sense in which we use the term to mean the assessment of subjective uncertainty, but rather describes the collection of data on individuals' preferences *without* any assessment of their uncertainty.

tested in an application in which six experts made judgements about a set of parameters required for a cost-effectiveness model for the diagnosis of hypertrophic cardiomyopathy. The invited experts reported that although the elicitation task was difficult, the process was straightforward.

A 'real life' application of elicitation is described in Girling et al. (2007) where informative priors were obtained from a group of experts for parameters that represented perioperative mortality and median survival in patients who were fitted with a left ventricular assist device. The device was deemed unlikely to be cost effective at £30,000/QALY, largely due to its high cost of £60,000. Even if the device were to be given away free there was still substantial decision uncertainty, driven primarily by the uncertainty in the survival benefits of the device. The elicited priors were therefore important in determining the expected value of perfect information, which was reported as being greater than the cost of an randomised controlled trial under certain assumptions. The authors made the following (rather tongue in cheek) observation about their results:

> "The subjective nature of the cost-effectiveness probabilities means that healthcare providers may view them with little more than academic interest..."

This probably reflects a reality that properly elicited subjective information is likely to be treated with scepticism in the health care allocation decision process, even though the NICE health technology assessment methods guide explicitly invites the use of formal elicitation methods (NICE, 2009).

Bojke et al. (2010) elicited progression rates for psoriatic arthritis in patients on treatment with anti-tumour necrosis factor and after treatment failure in order to populate a cost-effectiveness model. They found that the results (expressed as incremental cost-effectiveness ratios) were sensitive to the method used to derive a single distribution given multiple experts. This highlights the problem, inherent in all group elicitation exercises, of how to derive a single distribution from multiple experts.

We will not discuss multiple expert elicitation at length, other than to note that methods for combining opinion fall into two broad categories: *mathemat-*

ical or *behavioural*. Mathematical methods construct a single distribution $p(\theta)$ from the experts densities $\{p_1(\theta), \ldots, p_n(\theta)\}$, for example by linear pooling where $p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta)$ for some set of weights $w_1, \ldots, w_n$ (O'Hagan et al., 2006). Or the density may be derived using formal Bayesian methods (known in this context as *supra-Bayesian* methods), an approach first proposed by Morris (1974). Behavioural aggregation approaches attempt to elicit a single distribution from the group of experts, who may or may not use a formal consensus generating procedure (e.g. Delphi, or Nominal Group Technique).

Recently, Soares et al. (2011) described a comprehensive elicitation study to inform the parameters for a cost-effectiveness study of pressure ulcer treatments. Notably, this ambitious project showed that it is feasible to elicit from front line health care workers the range of quantities necessary to parameterise a three state Markov model with time varying transition probabilities. Twenty three nurses took part in the exercise and distributions were pooled (with equal weighting), despite highly discordant judgements about some quantities. Again, the problem of how to derive a single distribution from multiple experts was not easy to resolve.

Finally, two further cost-effectiveness studies of note are Stevenson et al. (2008), in which parameters concerning vCJD epidemiology and surgical instrument decontamination were elicited in order to populate a patient level simulation model, and Stevenson et al. (2009), in which beliefs about the long term efficacy of an osteoporosis treatment were elicited. In this latter study, the purpose of the economic model was explicitly to guide a decision on whether a randomised controlled trial to establish the long term efficacy was cost-effective, rather than guiding the drug adoption decision itself.

4.4.2 Elicitation and model uncertainty

Health economic model uncertainty may not be resolvable using data for the reasons we have explained in chapter 3. Elicitation therefore has the potential to play an important role in managing this source of uncertainty, but at present there are very few published descriptions of its use in this context. In Bojke et al. (2009)'s paper on the characterisation of structural uncertainty the authors suggest that expert elicitation could be used to derive distributions for model weights for the purposes of averaging, or alternatively to provide distributions for parameters that control choices between competing sub-models within a larger model that encompasses the smaller models as special cases.

This approach was tested in the Bojke et al. (2010) paper described above. The uncertainty in the psoriatic arthritis study was described as 'structural', reflecting the approach of replacing a set of models that differ with respect to a set of 'structual' assumptions with a single model that contains uncertain parameters that 'index' the assumptions. The newly introduced parameters are then the subject of an elicitation exercise, and the value of learning them can be established using standard methods (see also the related paper, Jackson et al., 2011).

The use of elicitation to inform model weights was tested by Negrín and Vázquez-Polo (2008) in a cost-effectiveness study of anti-retroviral treatment regimens for HIV. The competing models were identical except for the parameterisation of two linear regression equations that related costs and health effects to a series of patient level covariates. Experts were asked to judge, for each covariate, the probability that it should be included in the model. Models were then averaged over this (joint) distribution. It is not clear from the paper how easy the experts found this task. Eliciting beliefs about the probability that a covariate should be included in a regression is quite removed from asking an expert to make a statement about an observable quantity.

4.4.3 Some other challenges

In the clinical encounter, health care professionals are used to sequentially updating prior beliefs (knowledge of the background prevalence of disease in the presenting patient) with data in the form of the patient history, then data in the form of examination findings and finally data in the form of test results. Decision making under uncertainty and the ideas underlying Bayesian prior to posterior analysis are therefore not alien to practitioners and researchers in the field (see for example O'Connor and Sox, 1991). Quite properly there are reservations about the use of a prior distribution derived from an expert with a conflict of interest (it is not appropriate to ask a drug company sponsor for their prior beliefs about the efficacy of their new drug), but this should not prevent the use of elicitation *per se* (as argued by Stevens and O'Hagan, 2002). However, despite the comfort practitioners may have with Bayesian decision making, there are still widespread objections to the use of subjective information in health related research, as summed up by the following from pharmaceutical statistician Senn (2008):

> "... the gloomy conclusion to which I am drawn on reading de Finetti (1974) is that ultimately the Bayesian theory is destructive of any form of public statistics."

A recent review of the use of elicitation by Johnson et al. (2010) found that although there were now a reasonable number of studies (33 at that point) in the health and medical literature, there had been very little evaluation of the methods that had been employed. Given the hostility to the explicit use of subjective information in medical decision making (as encapsulated in Senn's statement above) we would do well not to underestimate the importance of ensuring the quality of the elicitation process.

4.5 Conclusion

In this chapter we have reviewed the process of eliciting expert judgements about uncertain quantities in probabilistic form. We have reviewed some of the applications of elicitation in health services and medical research as well as specifically in health economic evaluation. We have seen that formal elicitation methods, though well described in the methodological literature, are not yet established in routine research practice. Of concern is the general lack of evaluation of the process where elicitation has been used. This makes it somewhat difficult to determine at present how robust the process of elicitation is when used in every day research practice.

A number of studies reported experts as describing the task of making judgements about unknown quantities as difficult, even if the elicitation process was straightforward. This difficulty is likely to increase with the complexity of the elicitation exercise, and particularly when there are many uncertain quantities with a complicated dependency structure. This makes the application of formal elicitation methods to the problem of making judgements about computer model error potentially daunting.

In the next chapter we introduce a method for incorporating judgements about model discrepancy into the analysis for a decision problem. We illustrate the method in two case studies in which model discrepancy distributions were derived in a rather informal process of 'self-elicitation'. This elicitation approach was deemed sufficient for the purposes of this initial 'proof of concept' piece of work, but it is not clear whether informal self-elicitation would be sufficient in a real application. We suspect though that any serious consideration of potential model error will add value to the modelling process. How we might elicit such judgements in a real application is an open question, and we return to this point in chapter 8.

Chapter 5

Case Study 1 - Managing Structural Uncertainty in a Decision Tree Model

5.1 Introduction

In this chapter[1] we propose a method for quantifying model error in a health economic model, and demonstrate the application of the method in a simple decision tree cost-effectiveness model. We have seen in chapter 3 that making judgements about model error is difficult, particularly in the context of health economic evaluation where there are no observations on the model output against which to calibrate the model. However, one advantage that health economic modellers do perhaps have over other modellers is the relative simplicity of their models. Health economic models tend to be 'white box' models that can be 'opened up' and examined, unlike the highly complex models used, say, in climate science. Does this ability to easily comprehend the internal structure of the model allow us to make more informed statements about the model error than by just considering the model output?

We propose a method for quantifying model discrepancy based on decompos-

[1] The content of this chapter is published as Strong M., Oakley J.E. and Chilcott, J. (2012). Managing structural uncertainty in health economic decision models: a discrepancy approach. *Journal of the Royal Statistical Society, Series C (Applied Statistics).* In press.

ing the cost-effectiveness model into a series of sub-functions, and considering potential error at the *sub-function* level, rather than at the model output level. We then use a variance based sensitivity analysis to locate the important sources of discrepancy within the model in order to guide model refinement. The resulting improved model is judged to contain less structural error, and the distribution on the model output better reflects our true uncertainty about the costs and effects of the intervention.

In section §5.2 we introduce the model that forms the basis for our case study. The model is a modified version of a cost-effectiveness model that was published by the National Institute of Health and Clinical Excellence (NICE) and used to inform a decision concerning interventions to promote physical activity in a sedentary population (NICE, 2006). We finish the section by reporting the 'base case' results without any assessment of structural uncertainty. In section §5.3 we propose a method for managing structural uncertainty, and then in section §5.4 describe the application of the method to the case study model. We report results of the discrepancy analysis in section §5.5.

5.2 The base case model: a physical activity intervention cost-effectiveness model

Our simplified version of the NICE physical activity intervention cost-effectiveness model aims to predict the incremental net benefit of two competing decision options: exercise on prescription (e.g. from a general medical practitioner) to promote physical activity (the 'intervention', $d = 2$), and a 'do nothing' scenario ('no intervention', $d = 1$). We assume that the intervention impacts on health by reducing the risks of three diseases: coronary heart disease (CHD), stroke and diabetes and we wish to include in the model health effects that relate to these three diseases. We are interested in costs that accrue as a result of the treatment of the three diseases, as well as those that relate to the intervention itself. The net benefit under decision option $d \in (1, 2)$ is

$$NB_d = \lambda Q_d - C_d \tag{5.1}$$

where Q_d is the population mean per person health effect measured in Quality Adjusted Life Years (QALYs), C_d is the population mean per person cost, and λ is the NICE 'threshold' monetary value of one QALY (as discussed in section §2.15). Our target quantity is the incremental net benefit of decision 2 over decision 1, measured in monetary units. This is defined as

$$Z = NB_2 - NB_1. \tag{5.2}$$

The decision maker's utility for decision option 2 over option 1 is taken to be equal to Z, the incremental net benefit, and we assume the constraints of the NICE Reference Case (section §2.15).

5.2.1 Description of 'base case' model - no assessment of structural uncertainty

The model is a simple static cohort model which can be viewed as a decision tree (figure 5.1). The left-most node represents the two decision options, $d = 1$, no intervention, and $d = 2$, the exercise prescription intervention. The first 'chance' node represents the proportion of the population who undertake new exercise under each decision option, with the second node representing the proportion of the population who maintain exercise, conditional on starting new exercise. The third node represents the proportion of the population who experience eight mutually exclusive health states conditional on each of the three outcomes from the first two nodes: exercise that is maintained, exercise that is not maintained, and no exercise (sedentary lifestyle).

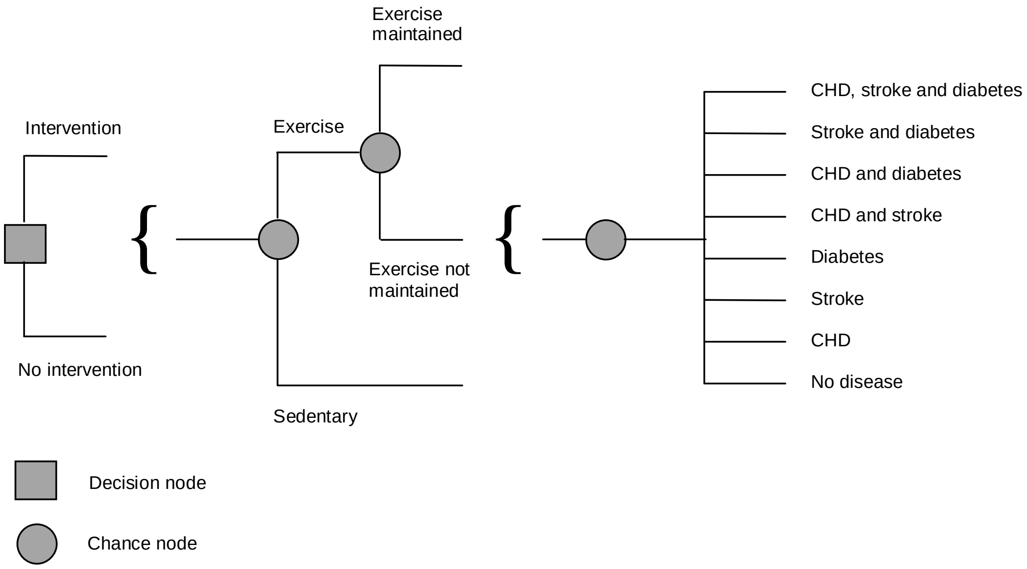


Figure 5.1: The model expressed as a decision tree

The structure of the model represents our beliefs about the causal links between the intervention and exercise, and exercise and health outcomes. There are no data available that relate to the model outputs; we have not observed costs and health outcomes for control and treatment groups on the exercise intervention. However, separate data sources are available regarding the effectiveness of the intervention in promoting exercise, and the risks of the various disease outcomes for active versus sedentary patients, and the availability of such data has guided the choice of model structure.

In our model each comorbid health state (e.g. the state of CHD *and* stroke) is treated as having a single onset point in time. Individuals do not progress, say, from the disease free state, to CHD and then to CHD plus stroke as they might do in reality. This is clearly unrealistic and is a consequence of the choice to use a very simple decision tree structure. Modelling sequential events is possible using a decision tree structure, but the number of terminal tree branches quickly becomes very large in all but the simplest of problems (Sonnenberg and Beck, 1993). A Markov or discrete event model structure would be more suited to addressing our decision problem (see Karnon (2003) for a comparison of these methods), but we have chosen to retain the important features of the structure of the model published by NICE, upon which our case study is based (NICE, 2006).

We denote the set of eight health states, *disease free, CHD alone, stroke alone, diabetes alone, CHD and stroke, CHD and diabetes, stroke and diabetes, CHD and stroke and diabetes* as $\mathcal{H} = \{h_i, i = 1, \ldots, 8\}$, where i indexes the set in the order given above. Each of the eight health states $h_i \in \mathcal{H}$, under each decision option d, has a cost c_{id} (measured in £), a health effect (measured in Quality Adjusted Life Years) q_{id}, and a probability of occurrence π_{id} (as approximated by the relative frequency with which this health state occurs within a large cohort). Total costs and total health effects for decision d are obtained by summing over health states, i.e. $C_d = \sum_{i=1} n^8 c_{id}\pi_{id}$ and $Q_d = \sum_{i=1}^{8} q_{id}\pi_{id}$. Given these, the model predicted incremental net benefit, Y is

$$Y = \lambda(Q_2 - Q_1) - (C_2 - C_1) = \lambda \Delta Q - \Delta C. \tag{5.3}$$

To relate the notation we have introduced here back to that of chapter 2, we note that the relevant outcomes for health state i under decision d are the costs, c_{id} and health effects q_{id}, i.e. that $\boldsymbol{o}_{i_d} = \{c_{id}, q_{id}\}$.

The costs c_{id}, health effects q_{id}, and health state probabilities π_{id} are not themselves input parameters in the model, but instead are functions of input parameters. There are 24 uncertain and three fixed input parameters that relate to the costs, quality of life and epidemiology of CHD, stroke and diabetes, and the effectiveness of the intervention in increasing physical activity. These inputs are denoted $\boldsymbol{X} = (X_1, \ldots, X_{27})$, and uncertainty is represented via the joint distribution $p(\boldsymbol{X})$. Finally, we denote the deterministic function that links the model inputs to the model output as f, i.e. $Y = f(\boldsymbol{X})$, and call this the *base case model.*

5.2.2 Input distributions

Our case study, although for illustrative purposes only, is based on a published model (NICE, 2006). However, the original modelling document did not contain enough information to allow us to derive distributions for all uncertain inputs. Where information was insufficient we derived plausible distributions through an informal elicitation process. The input distributions are described in tables 5.1

and 5.2. Starred entries denote those inputs for which distributions were derived through informal elicitation.

Where we had only point estimates for inputs we considered the following as plausible: costs were Gamma distributed with variance equal to the mean; counts of cases were Poisson distributed; the correlation between the proportion of the sedentary cohort who began new exercise in the intervention group and the corresponding proportion in the non-intervention group was 0.5; the correlation between the proportion of the new exercise cohort who maintained exercise in the intervention group and the corresponding proportion in the non-intervention group was 0.9; the standard deviations for the mean age of onset of disease and mean ages of death from the three diseases were all 2 years.

The input parameters that relate to the effectiveness of the intervention in increasing physical activity were estimated from a randomised controlled trial (Lamb et al., 2002), cited in the modelling document (NICE, 2006). The observed relative effectiveness would therefore be expected to be a reasonable estimate of the 'true' relative effectiveness in a similar population to that recruited to the Lamb et al. (2002) trial. In contrast, the inputs that relate to the risks of disease conditional on exercise status were estimated from observational studies, and those estimates are therefore more prone to bias and confounding. For the purposes of our analysis, however, we assume that the distributions in table 5.1 represent our best judgements about the inputs, given the limitations of the studies from which they were obtained. Clearly, there is a large medical statistics literature that relates to the problem of estimating treatment effects in trials, and similarly a large epidemiology literature that relates to the problem of estimating associations between risk factors and disease in observational studies.

Table 5.1: Uncertain inputs and their distributions

Input	Label	Description	Distribution	Hyperparameters	
X_1^*	c_0	Intervention cost (£)	Gamma	shape=100; scale=1	
X_2^*	t_{chd}	Total NHS costs (2005) for CHD (£)	Gamma	sh=3.677×10^9; sc=1	
X_3^*	t_{str}	Total NHS costs (2005) for stroke (£)	Gamma	sh=2.872×10^9; sc=1	
X_4^*	t_{dm}	Total NHS costs (2005) for diabetes (£)	Gamma	sh=5.314×10^9; sc=1	
X_5^*	n_{chd}	Number of UK cases of CHD	Poisson	$\mu = 2.60 \times 10^6$	
X_6^*	n_{str}	Number of UK cases of stroke	Poisson	$\mu = 1.40 \times 10^6$	
X_7^*	n_{dm}	Number of UK cases of diabetes	Poisson	$\mu = 1.53 \times 10^6$	
X_8	$q_{chd}^{(dec)}$	Discounted decremental health effect for CHD (QALYs)	Normal	$\mu = 6.71$; $\sigma = 0.048$	
X_9	$q_{str}^{(dec)}$	Discounted decremental health effect for stroke (QALYs)	Normal	$\mu = 10.23$; $\sigma = 0.048$	
X_{10}	$q_{dm}^{(dec)}$	Discounted decremental health effect for DM (QALYs)	Normal	$\mu = 2.08$; $\sigma = 0.048$	
X_{11}^*	$p_1^{(ex)}$	Proportion of sedentary cohort who begin new exercise in non-intervention group	MVN	$\mu = 0.246$; $\sigma = 0.038$	$\rho = 0.5$
X_{12}^*	$p_2^{(ex)}$	Proportion of sedentary cohort who begin new exercise in intervention group		$\mu = 0.294$; $\sigma = 0.040$	
X_{13}^*	$p_1^{(mnt)}$	Proportion of new exercise cohort who maintain exercise in non-intervention group	MVN	$\mu = 0.5$; $\sigma = 0.1$	$\rho = 0.9$
X_{14}^*	$p_2^{(mnt)}$	Proportion of new exercise cohort who maintain exercise in intervention group		$\mu = 0.5$; $\sigma = 0.1$	
X_{15}	$r_{chd}^{(sed)}$	Risk of CHD in a sedentary group	Beta	$\alpha = 80$; $\beta = 385$	
X_{16}	$r_{str}^{(sed)}$	Risk of stroke in a sedentary group	Beta	$\alpha = 226$; $\beta = 4072$	
X_{17}	$r_{dm}^{(sed)}$	Risk of diabetes in a sedentary group	Beta	$\alpha = 346$; $\beta = 3344$	
X_{18}	RR_{chd}	Relative risk of CHD in active vs sedentary pop	Lognormal	$\mu = 0.666$; $\sigma = 0.130$	
X_{19}	RR_{str}	Relative risk of stroke in active vs sedentary pop	Lognormal	$\mu = 0.720$; $\sigma = 0.343$	
X_{20}	RR_{dm}	Relative risk of diabetes in active vs sedentary pop	Lognormal	$\mu = 0.710$; $\sigma = 0.123$	
X_{21}^*	$age^{(onst)}$	Average age of onset of disease (same for all diseases)	Normal	$\mu = 57.5$; $\sigma = 2$	
X_{22}^*	$age_{chd}^{(dth)}$	Average age of death for CHD (years)	Normal	$\mu = 71$; $\sigma = 2$	
X_{23}^*	$age_{str}^{(dth)}$	Average age of death for stroke (years)	Normal	$\mu = 59$; $\sigma = 2$	
X_{24}^*	$age_{dm}^{(dth)}$	Average age of death for diabetes (years)	Normal	$\mu = 61$; $\sigma = 2$	

Table 5.2: Fixed inputs

Input	Label	Description	Value
X_{25}	$age^{(int)}$	Average age of cohort at time of intervention (years)	50
X_{26}	θ	Discount rate (per year)	0.035
X_{27}	λ	NICE threshold for value of 1 QALY (£/QALY)	20,000

5.2.3 Base case model results

The model function (which we describe in detail in section §5.4) was implemented in R (R Development Core Team, 2010). We sampled the input space and ran the model 100,000 times. The mean of the model output, Y, at λ=£20,000/QALY was £247 and the 95% credible interval was (-£315, £1002). The probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at λ =£20,000 was 0.77. Results for the base case model are shown graphically in figure 5.2

Figure 5.2 shows the cost-effectiveness plane (with 100 Monte Carlo samples). The sloped line shows the NICE threshold of £20,000 per QALY. To aid clarity figure 5.2b is a contour plot representation of the cost-effectiveness plane, showing the 95^{th} percentile of an empirical kernel density estimate of the joint distribution of costs and effects. Figure 5.2c shows the cost-effectiveness acceptability curve (i.e. a plot of $P(\text{INB} > 0)$ against λ) for values of λ from £0/QALY to £40,000/QALY. Finally, figure 5.2d shows the kernel density estimate for Y, the base case model estimate of the incremental net benefit at λ =£20,000.

The expected population mean incremental net benefit of £247 implies that the intervention will accrue costs and health effects that have a positive net value of £247 per person treated (assuming that a QALY is valued at £20,000). The probabilistic sensitivity analysis implies that, at λ=£20,000/QALY, a choice to recommend the intervention would have a probability of 0.77 of being better than the choice not to recommend.

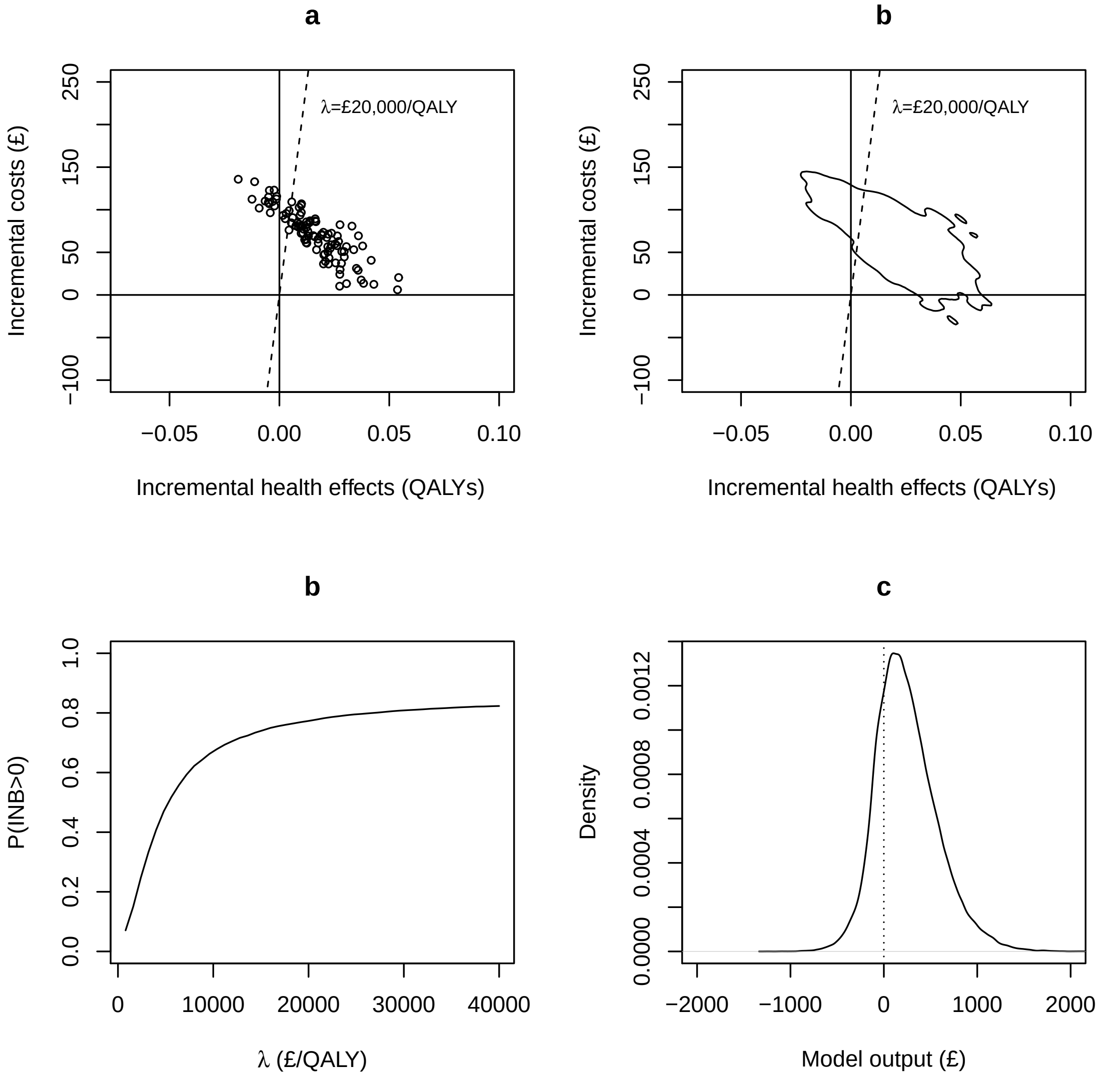


Figure 5.2: Base case model output shown as (a) cost-effectiveness plane (b) cost-effectiveness plane contour plot (c) cost-effectiveness acceptability curve (d) incremental net benefit empirical density.

5.3 Managing uncertainty due to structure: a discrepancy approach

Given a model, written as a function f, with (uncertain) inputs $\boldsymbol{X}$, we link the model output $Y = f(\boldsymbol{X})$ to the true, but unknown value of the target quantity we wish to predict, Z via

$$Z = f(\boldsymbol{X}) + \delta_z, \tag{5.4}$$

The discrepancy term, δ_z, quantifies the *structural error*: the difference between the output of the model evaluated at its true inputs and the true target quantity.

For the decision maker to base their decision on the model output, the model must have credibility. The model must be judged good enough to support the decision being made. The primary goal of our analysis is therefore to provide a means for quantifying judgements about structural error and specifically to determine the relative importance of structural compared to input uncertainty in addressing the decision problem. If uncertainty about structural error is large then we may wish to review the model structure. Conversely, if we can demonstrate that the uncertainty about structural error is small in comparison to that due to input uncertainty, then we have a stronger claim to have built a credible model.

In building the base case model we made a series of assumptions, for example we assumed that occurrences of CHD, stroke and diabetes are independent at the level of the individual and therefore that disease risks act multiplicatively. Such assumptions drive the structural choices that we make when formulating a model, and incorrect assumptions will lead to structural error. We must therefore focus our attention on the assumptions within a model if we are to assess its adequacy and properly quantify our uncertainty about the target quantity.

We therefore propose a method for deriving a distribution for the model discrepancy, δ_z, as defined in equation (5.4). In contrast to the model averaging approach (chapter 3) we do not attempt to make assessments about the adequacy of the model structure in relation to alternative structures; we instead assess how large an error might be due to the structure of the model at hand.

5.3.1 Discrepancy between model output and reality

Making meaningful judgements about the model discrepancy will be difficult, though it should always be possible to make a crude evaluation of a plausible range of orders of magnitude of δ_z, for example by asking questions like ‘could the true incremental net benefit of decision 1 over decision 2 be a billion pounds greater than that predicted by the model, or a million pounds greater, or only a hundred pounds greater?’. However, it may be easier to make judgements about

δ_z indirectly. If we consider f in more detail we may be able to determine where in the model structural errors are likely to be located, and what their consequences might be. We therefore propose making judgements about discrepancy at the *sub-function* level.

5.3.2 Discrepancy at the 'sub-function' level

Any model f, except the most trivial, can be decomposed into a series of sub-functions that link the model inputs to the model output. To illustrate, figure 5.3a shows a hypothetical model with ten inputs, $Y = f(X_1, \ldots, X_{10})$, that aims to predict a quantity Z. The model has been decomposed into a series of sub-functions, for example $Y_1 = f_1(X_1, X_2, X_3)$ and $Y_2 = f_2(X_4, X_5)$, revealing a set of six 'intermediate' parameters $Y_1, \ldots, Y_6$ that have 'true' values $Z_1, \ldots, Z_6$.

For each sub-function, we ask the question 'would this sub-function, if evaluated at the true values of its inputs, result in the true value of the sub-function output?'. If not then we recognise potential structural error and introduce an uncertain discrepancy term, δ_j, either on the additive scale, i.e. $Y_j = f_j(\cdot) + \delta_j$, or multiplicative scale, i.e. $\log(Y_j) = \log\{f_j(\cdot)\} + \log(\delta_j)$. The idea is that, because each sub-function represents a much simpler process than the full model f, making judgements about discrepancy in f_j will be easier than making judgements about discrepancy in f.

Repeating the process for all sub-functions in the model will leave us with a series of n discrepancy terms, which we denote $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. Note that for some sub-functions we will judge there is no structural error, usually when an intermediate parameter is by definition equal to the sub-function that generates it.

Returning to our hypothetical model, we judge there to be structural error in three of the sub-functions, and therefore introduce three discrepancy terms to correct the error. The three terms are introduced on the additive scale giving: $Z_1 = Y_1 + \delta_1$, $Z_5 = Y_5 + \delta_2$ and $Z_6 = Y_6 + \delta_3$. Figure 5.3b shows the incorporation of the three uncertain discrepancy terms.

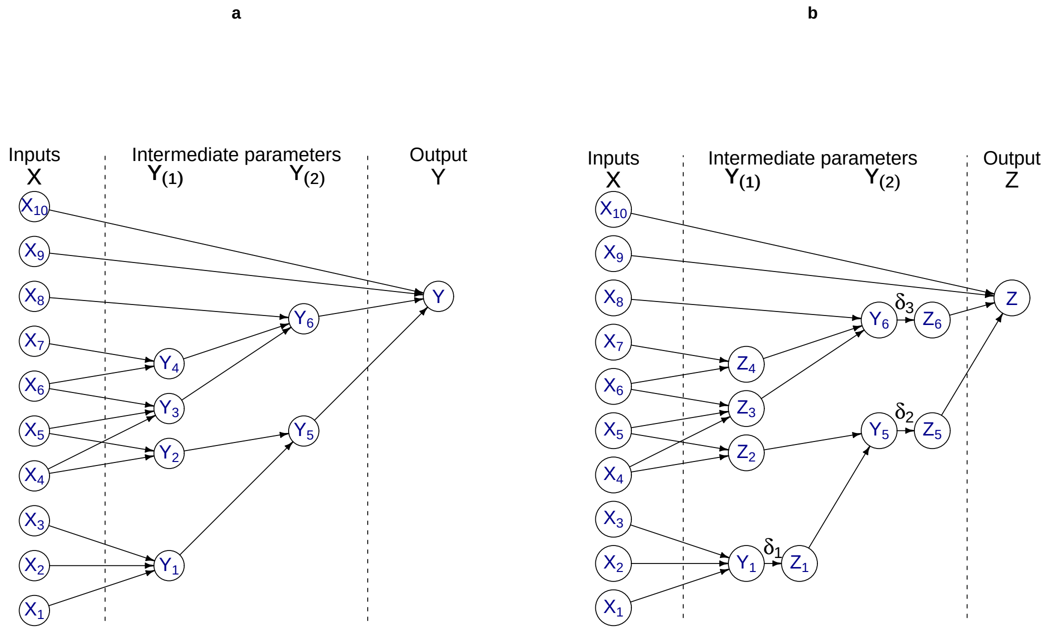


Figure 5.3: (a) Hypothetical model with ten inputs and one output, decomposed to reveal six intermediate parameters. (b) We suppose that there is structural error in the sub-functions that result in Y_1, Y_5 and Y_6. Three discrepancy terms are added to correct the error, i.e. $Z_1 = Y_1 + \delta_1$, $Z_5 = Y_5 + \delta_2$ and $Z_6 = Y_6 + \delta_3$.

There will usually not be a unique decomposition of the model f into a series of sub-functions that links the model inputs $\boldsymbol{X}$ to the model output Y. However, some decompositions will be more useful than others for assessing discrepancy. Following the advice that it is preferable to elicit beliefs about observable quantities (Kadane and Wolfson, 1998; O'Hagan et al., 2006), we search for decompositions where both inputs and outputs of the sub-functions are observable.

Once we have introduced discrepancy terms at the locations within the model where we judge there is potential structural error, we must make judgements about the discrepancies via the specification of the joint probability distribution $p(\boldsymbol{X}, \boldsymbol{\delta})$. We assume in our case study that discrepancies are independent of inputs, such that we can factorise the joint density $p(\boldsymbol{X}, \boldsymbol{\delta})$=$p(\boldsymbol{X})p(\boldsymbol{\delta})$. This independence assumption does not need to hold for the discrepancy method to be valid, but specification of $p(\boldsymbol{\delta})$ independent of $p(\boldsymbol{X})$ will clearly be easier than specifying $p(\boldsymbol{X}, \boldsymbol{\delta})$ where there is dependence.

Even if we are able to assume that discrepancies are independent of inputs, we will need to think quite hard about the correlation structure in the discrepancies themselves. For example, in figure 5.3b the discrepancy term δ_1 is 'proximal' to δ_2. We are assuming that we have correctly specified δ_1 when we specify δ_2. We are also (implicitly here) assuming that the joint distribution on the inputs, $p(\boldsymbol{X})$, is properly specified. This may have required introducing discrepancy components even more 'proximally' at the level of the model inputs. This provides a nice link to Turner et al. (2009)'s work on bias modelling.

In specifying $p(\boldsymbol{\delta})$ we begin by considering the mean and variance for each discrepancy term δ_j, $j = 1, \dots, n$. In our case study we make judgements about the sizes of the discrepancies relative to the mean values of the corresponding intermediate parameters, and set variances such that $\sqrt{\text{var}(\delta_j)} = v_j|E(Y_j)|$, with v_j chosen to reflect our judgements. Determining plausible values for v_j may not be a trivial task, a point to which we return later. We treat each δ_j as independent of all other uncertain quantities, unless there are constraints that prevent this (a constraint would arise, for example, in relation to a set of population proportion parameters that must sum to one) or unless there are good reasons to assume strong correlation between terms. Finally we select appropriate distributions with the specified means and variances.

Propagating the uncertainty we have specified for $\boldsymbol{\delta}$ through the model, along with the uncertainty in the inputs, $\boldsymbol{X}$, allows us to check that the uncertainty in Z that our specification of $p(\boldsymbol{\delta})$ implies is plausible. If this is not the case then we must rethink our choice of distributions for the components of $\boldsymbol{\delta}$, most easily through altering our choices for v_j.

The sub-function discrepancy approach has two important consequences. Firstly, if we can adequately make judgements about all the discrepancy terms in the model (there may be many) then we will derive $p(\delta_z)$ and hence be able to make statements about our uncertainty about the incremental net benefit that incorporates beliefs about both inputs and structure. Perhaps more usefully though, we can use sensitivity analysis techniques to investigate the relative importance of the different structural errors, allowing us improve the parts of the model where this is most needed. If, after repeating the sensitivity analysis in our improved model, we find that discrepancies now have a lesser impact on the output uncertainty, then we have in an important sense built a more robust model structure.

5.4 Applying the sub-function discrepancy method to our physical activity model

We now return to our base case physical activity model, and beginning at the net benefit equation (5.3), work 'backwards' through the model, assessing potential structural error at each sub-function.

5.4.1 Assessment of sub-function generating the output parameter Y

The model output, Y predicts the incremental net benefit,

$$Y = \lambda(Q_2 - Q_1) - (C_2 - C_1) = \lambda\Delta Q - \Delta C. \tag{5.5}$$

Evaluation of equation (5.5) at the true values of ΔQ and ΔC would, by definition, result in the true value of the incremental net benefit, Z, so there is no structural error at this point in the model, and therefore no discrepancy term. We take as given that the two decision options are mutually exclusive and exhaustive, but clearly, if this were not to be so, then this would be a (possibly very important) source of structural error.

5.4.2 Assessment of sub-function generating the intermediate parameter ΔQ

The incremental health effect of the intervention over the non-intervention, ΔQ is

$$\Delta Q = \sum_{i=1}^{8} \pi_{i2} q_{i2} - \sum_{i=1}^{8} \pi_{i1} q_{i1}, \tag{5.6}$$

where π_{id} and q_{id} are the probabilities and discounted health effects in QALYs respectively for health state i under decision d. Future health effects (and future costs) are discounted to reflect time preference whereby higher value is placed on benefits that occur in the near future than on those occurring in the distant future. See Krahn and Gafni (1993) for a discussion of the role of discounting in health economic evaluation.

Health effects for each state i are assumed to be equal under each decision d, i.e. that $q_{i1} = q_{i2} = q_i$. The total health effects are

$$\begin{aligned}
\Delta Q &= Q_2 - Q_1 && (5.7)\\
&= \sum_{i=1}^{8} \pi_{i2} q_i - \sum_{i=1}^{8} \pi_{i1} q_i && (5.8)\\
&= \sum_{i=1}^{8} (\pi_{i2} - \pi_{i1}) q_i && (5.9)\\
&= \sum_{i=1}^{8} (\pi_{i2} - \pi_{i1})(q_i - q_1) && (5.10)\\
&= \sum_{i=1}^{8} (\pi_{i2} - \pi_{i1}) q_i^{(dec)}, && (5.11)
\end{aligned}$$

where the final term is a re-expression in terms of the *decremental* health effect, $q_i^{(dec)}$ relative to the disease free state $i = 1$.

We ask the question, ‘given the true values of π_{id} and q_i, does (5.7) result in the true value of ΔQ?’. Because we imagine that the intervention could have an impact on a number of diseases other than CHD, stroke and diabetes we recognise potential structural error and introduce an uncertain additive discrepancy term,

$\delta_{\Delta Q}$ into (5.7), which becomes

$$\Delta Q = \sum_{i=1}^{8} (\pi_{i2} - \pi_{i1}) q_i^{(dec)} + \delta_{\Delta Q}. \tag{5.12}$$

Since exercise can result in poor health outcomes as well as good outcomes, for example through musculo-skeletal injuries or accidents, we specify a mean of zero for $\delta_{\Delta Q}$. We could assume a non-zero mean for $\delta_{\Delta Q}$ if we felt that increased exercise was likely to be on balance beneficial. This will have the effect of shifting the mean of the model output unless the sub-function related to the discrepancy is entirely unimportant.

We judge that $\delta_{\Delta Q}$ is unlikely to be more than ±10% of ΔQ, and we represent our beliefs about $\delta_{\Delta Q}$ using a normal distribution with a standard deviation equal to 5% of the mean of ΔQ, i.e. $\delta_{\Delta Q} \sim \mathrm{N}[0, \{0.05 \times E(\Delta Q)\}^2]$.

5.4.3 Assessment of sub-function generating the intermediate parameter ΔC

The incremental cost of the intervention over the non-intervention, ΔC is

$$\Delta C = \sum_{i=1}^{8} \pi_{i2} c_{i2} - \sum_{i=1}^{8} \pi_{i1} c_{i1}, \tag{5.13}$$

where π_{id} and c_{id} are the probabilities and discounted costs respectively that are associated with health state i under decision d.

Costs, not including the cost of the intervention itself c_0, are assumed to be equal across decision arms, i.e. that $c_{i2} = c_{i1} + c_0$, and therefore that

$$\begin{aligned} \Delta C &= \sum_{i=1}^{8} \pi_{i2}(c_{i1} + c_0) - \sum_{i=1}^{8} \pi_{i1} c_{i1} & (5.14) \\ &= c_0 + \sum_{i=1}^{8} (\pi_{i2} - \pi_{i1}) c_{i1}, & (5.15) \end{aligned}$$

where c_0 is a model input.

As above, there may be costs that relate to diseases other than CHD, stroke and diabetes that are not included in ΔC and we therefore introduce an additive discrepancy term, $\delta_{\Delta C}$, and specify that $\delta_{\Delta C} \sim \mathrm{N}[0, \{0.05 \times E(\Delta C)\}^2]$.

5.4.4 Assessment of sub-function generating the intermediate parameters c_{i1}

The intermediate parameters c_{i1} represent the discounted cost associated with the eight health states under decision 1. In the base case model the costs for the eight states are derived from the costs associated with the three individual diseases, with costs for comorbid states assumed to be the sum of the costs for the constituent diseases, i.e.

$$\begin{aligned} c_{1,1} &= c_{well}, && (5.16)\\ c_{2,1} &= c_{chd}, && (5.17)\\ c_{3,1} &= c_{str}, && (5.18)\\ c_{4,1} &= c_{dm}, && (5.19)\\ c_{5,1} &= c_{chd} + c_{str}, && (5.20)\\ c_{6,1} &= c_{chd} + c_{dm}, && (5.21)\\ c_{7,1} &= c_{str} + c_{dm}, && (5.22)\\ c_{8,1} &= c_{chd} + c_{str} + c_{dm}. && (5.23) \end{aligned}$$

Costs may not be additive in this way, so we introduce additive discrepancy terms, δ_{c_j}, for the intermediate parameters that relate to the comorbid states, c_{i1} $i = 5, \ldots, 8$ (equations (5.20) to (5.23)).

We judge that comorbid state costs could be higher or lower than the sum of the constituent costs, so we assumed a mean of zero for each discrepancy term, δ_{c_i}, $i = 5, \ldots, 8$. We represent beliefs about δ_{c_i} via $\delta_{c_i} \sim \mathrm{N}[0, \{0.05 \times E(c_{i1})\}^2]$, $i = 5, \ldots, 8$.

5.4.5 Assessment of sub-function generating the intermediate parameters c_{chd}, c_{str} and c_{dm}

The discounted costs for CHD, stroke and diabetes are

$$c_k = c_k^* \times \alpha_k, \tag{5.24}$$

where k indexes the set $\{CHD, stroke, diabetes\}$. Costs (other than the cost of the intervention) are assumed to occur at some time in the future, and are discounted at 3.5% per year. The parameters c_k^* represent undiscounted costs, and α_k, are the discounting factors for the length of time between the intervention and the occurrence of the relevant health outcomes.

Given true values for c_k^* and α_k equation (5.24) will result in a true value for c_k, and there is no structural error at this point.

5.4.6 Assessment of sub-function generating the intermediate parameters c_{chd}^*, c_{str}^* and c_{dm}^*

The undiscounted mean per-person lifetime costs for CHD, stroke and diabetes are

$$c_k^* = \frac{t_k}{n_k}\left(age_k^{(dth)} - age_k^{(onst)}\right), \tag{5.25}$$

where k indexes the set $\{CHD,\ stroke,\ diabetes\}$, and where t_k are total annual NHS costs for disease k, and where n_k are UK prevalent cases of disease k for the same year. The parameters t_k, n_k, $age_k^{(dth)}$ and $age^{(onst)}$ are model inputs.

Mean per person undiscounted costs are calculated as the mean per person annual NHS cost multiplied by the mean length of time in the disease state. If the per person per year cost of disease is dependent on the length of time the individual spends in the disease state (e.g. if costs are greater near to the end of life), then c_{chd}^*, c_{str}^* and c_{dm}^* as calculated will not equal the mean per person per year costs. To properly calculate the mean we need to know the joint distribution of the costs and length of time in the disease state. To account for the difference we introduce discrepancy terms $\delta_{c_k^*}$.

We judge that disease costs could in reality be higher or lower than the modelled costs as a result of the structural error, so we assume a mean of zero for each discrepancy term, $\delta_{c_k^*}$. We represent beliefs about $\delta_{c_k^*}$ via $\delta_{c_k^*} \sim \mathrm{N}[0, \{0.05 \times E(c_k^*)\}^2]$.

5.4.7 Assessment of sub-function generating the intermediate parameters α_{chd}, α_{str} and α_{dm}

The discounting factors for CHD, stroke and diabetes are

$$\alpha_k = (1+\theta)^{-l_k}\,, \tag{5.26}$$

where l_k is the mean length of life remaining at the time of intervention for disease $k \in \{CHD,\ stroke,\ diabetes\}$, and θ is the per-year discount rate for both costs and health effects. The mean length of life remaining, l_k, is given by

$$l_k = \frac{1}{2}\left(age_k^{(onst)} + age_k^{(dth)}\right) - age^{(int)}, \tag{5.27}$$

where $age_k^{(onst)}$ is the mean age of onset of disease k, $age_k^{(dth)}$ is the mean age of death from disease k and $age^{(int)}$ is the mean age of the cohort at the time of the intervention. The parameters θ, $age_k^{(dth)}$, $age_k^{(onst)}$ and $age^{(int)}$ are model inputs.

In the base case model we assume that the costs of each disease will be realised at a time midway between the average age of disease onset, and the average age of death from that disease. This is not necessarily true and we introduce additive discrepancy terms δ_{α_k}.

Discount factors must lie in $(0, 1]$, and so discrepancies must lie in $(-\alpha_k, 1-\alpha_k]$. To satisfy this constraint we assume that $\alpha_k + \delta_{\alpha_k}$ follows a beta distribution. We have no reason to believe that the true values of the discount rates will be higher or lower than the modelled values, so we assume that δ_{α_k} has mean zero for all k. As above, we assume that the standard deviation is 5% of the mean value of the intermediate parameter, i.e. that $\sqrt{\mathrm{var}(\delta_{\alpha_k})} = 0.05E(\alpha_k)$.

The more general Dirichlet distribution specification of uncertainty is required for other discrepancy terms in the model, so for brevity we treat $\alpha_k + \delta_{\alpha_k}$ and $1 - (\alpha_k + \delta_{\alpha_k})$ as 'sum-to-one' parameters and the beta distribution as a special case of the Dirichlet distribution. We describe in section §5.4.12 how we chose Dirichlet distribution hyperparameters to satisfy these requirements.

5.4.8 Assessment of sub-function generating the intermediate parameters $q_i^{(dec)}$

The intermediate parameters $q_i^{(dec)}$ represent the discounted decremental health effects (in QALYs) associated with the eight health states. In the base case model these terms are derived from the discounted decremental health effects associated with the three individual diseases, with decremental effects for comorbid states assumed to be the sum of the decremental effects for the constituent diseases, i.e.

$$\begin{aligned}
q_1^{(dec)} &= 0, && (5.28)\\
q_2^{(dec)} &= q_{chd} - q_{well} = q_{chd}^{(dec)}, && (5.29)\\
q_3^{(dec)} &= q_{str} - q_{well} = q_{str}^{(dec)}, && (5.30)\\
q_4^{(dec)} &= q_{dm} - q_{well} = q_{dm}^{(dec)}, && (5.31)\\
q_5^{(dec)} &= (q_{chd} - q_{well}) + (q_{str} - q_{well}) = q_{chd}^{(dec)} + q_{str}^{(dec)}, && (5.32)\\
q_6^{(dec)} &= (q_{chd} - q_{well}) + (q_{dm} - q_{well}) = q_{chd}^{(dec)} + q_{dm}^{(dec)}, && (5.33)\\
q_7^{(dec)} &= (q_{str} - q_{well}) + (q_{dm} - q_{well}) = q_{str}^{(dec)} + q_{dm}^{(dec)}, && (5.34)\\
q_8^{(dec)} &= (q_{chd} - q_{well}) + (q_{str} - q_{well}) + (q_{dm} - q_{well}) \\
&= q_{chd}^{(dec)} + q_{str}^{(dec)} + q_{dm}^{(dec)}, && (5.35)
\end{aligned}$$

where the parameters $q_{chd}^{(dec)}$, $q_{str}^{(dec)}$ and $q_{dm}^{(dec)}$ are model inputs. Decremental health effects may not be additive in this way, so we introduce discrepancy terms δ_{q_i} for the comorbid health states $i = 5, \ldots, 8$ (equations (5.32) to (5.35)).

We judge that comorbid state decremental health effects could be higher or lower than the sum of the constituent terms, so assume a mean of zero for each discrepancy term, δ_{q_i}, $i = 5, \ldots, 8$. We represent beliefs about δ_{q_i} via $\delta_{q_i} \sim \mathrm{N}[0, \{0.05 \times E(q_i)\}^2]$, $i = 5, \ldots, 8$.

5.4.9 Assessment of sub-function generating the intermediate parameters π_{id}

The proportions of the population who are expected to experience each disease state $i = 1, \ldots, 8$ under decision options $d = 1, 2$ are

$$\pi_{id} = p_d^{(ex)}\, p_d^{(mnt)}\, r_i^{(ex)} + p_d^{(ex)} \left(1 - p_d^{(mnt)}\right) r_i^{(sed)} + \left(1 - p_d^{(ex)}\right) r_i^{(sed)}, \qquad (5.36)$$

where $r_i^{(ex)}$ and $r_i^{(sed)}$ are the risks of disease state i in those who exercise and in those who are sedentary, respectively. The probability of new exercise under decision option d is $p_d^{(ex)}$, and the probability of maintenance of exercise is $p_d^{(mnt)}$. The parameters $p_d^{(ex)}$ and $p_d^{(mnt)}$ are model inputs.

Parameters defining health state probabilities lie in $[0, 1]$, and must sum to one over i, so discrepancies must lie in $[-\pi_{id}, 1 - \pi_{id}]$, and must sum to zero over i. To satisfy this constraint we assume a Dirichlet distribution for $\pi_{id} + \delta_{\pi_{id}}$.

We have no reason to believe that the true values of the health state probabilities would be higher or lower than the modelled values, so we assume that $E(\delta_{\pi_{id}}) = 0,\ \forall i, d$. We assume that the standard deviation was 5% of the mean value of the intermediate parameter, i.e.

$$\frac{1}{8} \sum_{i=1}^{8} \frac{\sqrt{\mathrm{var}(\delta_{\pi_{id}})}}{E(\pi_{id})} = 0.05. \qquad (5.37)$$

See section §5.4.12 for details of the calculation of the Dirichlet hyperparameters that satisfy these requirements.

5.4.10 Assessment of sub-function generating the intermediate parameters $r_i^{(ex)}$ and $r_i^{(sed)}$

The parameters $r_i^{(ex)}$ and $r_i^{(sed)}$ represent the risks of health state i in a population that exercises and in a sedentary population, respectively. In the base case model we assume that occurrences of CHD, stroke and diabetes are independent, and therefore that the $r_{chd}^{(ex)}$, $r_{str}^{(ex)}$ and $r_{dm}^{(ex)}$ act multiplicatively to generate the $r_i^{(ex)}$

(and similarly multiplicatively in the sedentary population). So for example,

$$r_1^{(ex)} = (1 - r_{chd}^{(ex)})(1 - r_{str}^{(ex)})(1 - r_{dm}^{(ex)}). \tag{5.38}$$

We assume that occurrences of CHD, stroke and diabetes are independent, which may not be true, so we introduce additive discrepancy terms $\delta_{r_i^{(sed)}}$ and $\delta_{r_i^{(ex)}}$. Following the same argument as that in 5.4.9 we assume a Dirichlet distributions for $r_i^{(ex)} + \delta_{r_i^{(ex)}}$ and for $r_i^{(sed)} + \delta_{r_i^{(sed)}}$. We have no reason to believe that the true values of the disease risks would be higher or lower than the modelled values, so we assume that $E(\delta_{r_i^{(ex)}}) = E(\delta_{r_i^{(sed)}}) = 0$, $\forall i$. We assume that the standard deviations were 5% of the mean values of the intermediate parameters, i.e.

$$\frac{1}{8}\sum_{j=1}^{8} \frac{\sqrt{\text{var}\left(\delta_{r_i^{(ex)}}\right)}}{E\left(r_i^{(ex)}\right)} = \frac{1}{8}\sum_{i=1}^{8} \frac{\sqrt{\text{var}\left(\delta_{r_i^{(sed)}}\right)}}{E\left(r_i^{(sed)}\right)} = 0.05. \tag{5.39}$$

5.4.11 Assessment of sub-function generating the intermediate parameters $r_k^{(ex)}$

The parameters $r_k^{(ex)}$ where k indexes the set $\{CHD, stroke, diabetes\}$ represent the risks of CHD, stroke and diabetes in those who exercise. They are calculated by multiplying baseline risk by the relative risk of disease given exercise, i.e.

$$r_k^{(ex)} = r_k^{(sed)} \times RR_k, \tag{5.40}$$

where $r_k^{(sed)}$ and RR_k are model inputs.

Given true values for $r_k^{(sed)}$ and RR_k, and an assumption that the relative effect size is constant with respect to baseline risk, sub-function (5.40) will result in the true value of $r_k^{(ex)}$ by definition of a relative risk. There is therefore no structural error at this point. The assumption of constant effect size with respect to baseline is felt to be reasonable in this case, but if in another circumstance it was not then we would add a discrepancy term at this point.

5.4.12 Generating a sample from the distribution on the discrepancy relating to a sum-to-one parameter

We denote a sum-to-one intermediate parameter as $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, where $Y_j \in [0,1]\ \forall j$ and $\sum_{j=1}^{n} Y_j = 1$.

The true unknown value of the intermediate parameter is denoted $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ where $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{\delta_Y}$ and $\boldsymbol{\delta_Y} = (\delta_{Y_1}, \dots, \delta_{Y_n})$. The same constraints apply to $\boldsymbol{Z}$ as to $\boldsymbol{Y}$, i.e. $Z_j \in [0,1]\ \forall j$ and $\sum_{j=1}^{n} Z_j = 1$.

We state the following beliefs about $\boldsymbol{\delta_Y}$. Firstly, that $E(\delta_{Y_j}) = 0\ \forall j$, and secondly that the mean of the ratio of the standard deviation of the discrepancy to the expected value of the parameter is some constant v, i.e. that

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\mathrm{var}(\delta_{Y_j})}}{E(Y_j)} = v. \tag{5.41}$$

We generate a sample from $p(\boldsymbol{Z})$ as follows. Firstly, we sample $\{\boldsymbol{y}_s, s = 1, \dots, S\}$ from $p(\boldsymbol{Y})$. Conditional on $\boldsymbol{Y}$ we then generate a sample $\{\boldsymbol{z}_s, s = 1, \dots, S\}$ from $p(\boldsymbol{Z})$, where each $\boldsymbol{z}_s$ is a single draw from $p(\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}_s)$. The conditional distribution of $\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}_s$ is Dirichlet with hyperparameter vector $\gamma\boldsymbol{y}_s = (\gamma y_{1,s}, \dots, \gamma y_{n,s})$.

The expectation of δ_{Y_j} is

$$E(\delta_{Y_j}) = E(Z_j) - E(Y_j) = E_{Y_i}\{E_{Z_j}(Z_j|Y_j)\} - E(Y_j) = 0, \tag{5.42}$$

as required. The variance of δ_{Y_j} is

$$\begin{aligned}
\text{var}(\delta_{Y_j}) &= \text{var}(Z_j) + \text{var}(Y_j) - 2\text{Cov}(Z_j, Y_j) && (5.43)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}_{Y_j}\{E_{Z_j}(Z_j|Y_j)\} \\
&\quad +\text{var}(Y_j) - 2\text{cov}(Z_j, Y_j) && (5.44)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}_{Y_j}\{E_{Z_j}(Z_j|Y_j)\} \\
&\quad +\text{var}(Y_j) - 2\text{var}(Y_j) && (5.45)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} + \text{var}(Y_j) + \text{var}(Y_j) - 2\text{var}(Y_j) && (5.46)\\
&= E_{Y_j}\{\text{var}_{Z_j}(Z_j|Y_j)\} && (5.47)\\
&= E_{Y_j}\left\{\frac{(Y_j(1-Y_j)}{\gamma+1}\right\} && (5.48)\\
&= \frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1} - \frac{\text{var}(Y_j)}{\gamma+1} && (5.49)\\
&\simeq \frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1}. && (5.50)
\end{aligned}$$

The final step follows because $\text{var}(Y_j)$ is small relative to $E(Y_j)\{1 - E(Y_j)\}$ in this application. The Y_j $(j = 1, \ldots, n)$ are intermediate parameters in the case study model where they represent proportions of the population in certain states. Because they are derived from other uncertain quantities they usually do not have a standard distribution. However, if we assume that Y_j is approximately beta distributed with hyperparameters α_j and β_J, then the ratio of $E(Y_j)\{1 - E(Y_j)\}$ to $\text{var}(Y_j)$ is approximately equal to $\alpha_j + \beta_j + 1$. Unless the distribution of Y_j is very dispersed, $\alpha + \beta + 1$ will not be small and therefore $E(Y_j)\{1 - E(Y_j)\}$ will dominate $\text{var}(Y_j)$.

The hyperparameter γ is chosen such that the mean of the ratio of the standard deviation to the expected value of the parameter is v, i.e. so that

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\text{var}(\delta_{Y_j})}}{E(Y_j)} = \frac{1}{n}\sum_{j=1}^{n}\frac{\sqrt{\frac{E(Y_j)\{1-E(Y_j)\}}{\gamma+1}}}{E(Y_j)} = v. \qquad (5.51)$$

Approximating $E(Y_j)$ by the sample mean $\bar{y}_j$ and rearranging gives

$$\gamma = \frac{1}{v^2}\left\{\frac{1}{n}\sum_{j=1}^{n}\sqrt{\frac{1-\bar{y}_j}{\bar{y}_j}}\right\}^2 - 1. \tag{5.52}$$

5.5 Discrepancy analysis results

Following the discrepancy analysis in the case study model a total of 48 discrepancy terms were introduced. The addition of the discrepancy terms ‘corrects’ any structural error, and allows us now to write

$$Z = f^*(\boldsymbol{X}, \boldsymbol{\delta}), \tag{5.53}$$

where f^* takes the same functional form as f, but with the inclusion of the discrepancy terms as described in section §5.4.

5.5.1 Model output after inclusion of discrepancy terms

We sampled the input and discrepancy distributions and ran the model f^* 100,000 times. This resulted in a predicted mean incremental net benefit of £247, which is equal to the that predicted by the base case model. The 95% credible interval was -£886 to £1444, which is wider than that of the base case model, reflecting the recognition of our additional uncertainty about the true incremental net benefit due to possible model structural error.

Figure 5.4 shows the model results after the addition of the 48 discrepancy terms. We note the larger cloud of points on the cost-effectiveness plane (figures 5.4a and 5.4b), reflecting the additional uncertainty. The additional uncertainty has reduced the probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at $\lambda =$£20,000 to 0.66 (closer to the value of 0.5 that represents complete uncertainty), and flattened the cost-effectiveness acceptability curve towards the horizontal line at $P(\text{INB} > 0) = 0.5$ (figure 5.4c). The additional uncertainty is also reflected in the wider empirical distribution in figure 5.4d.

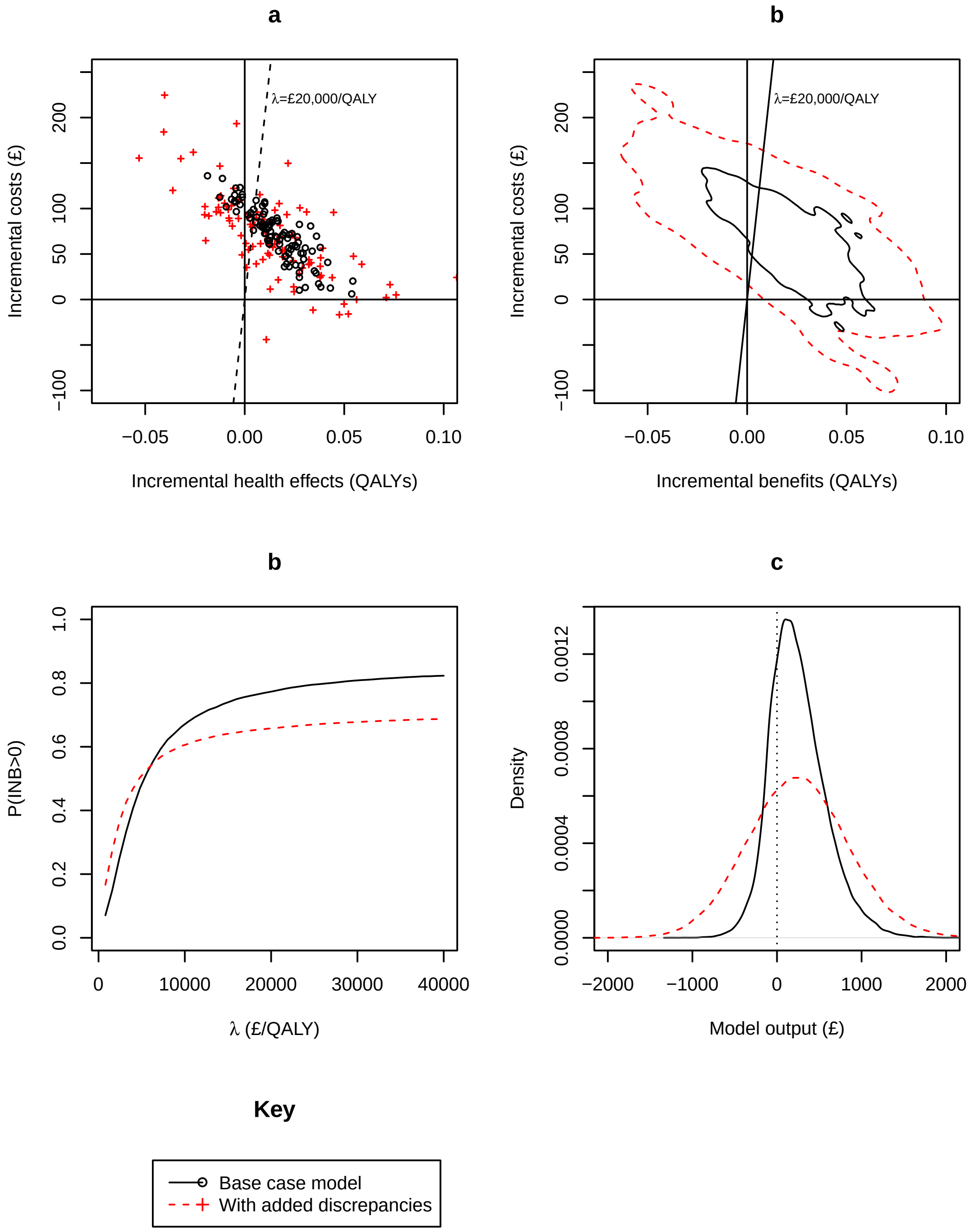


Figure 5.4: Results after the addition of discrepancy terms as (a) cost-effectiveness plane (b) cost-effectiveness plane contour plot (c) cost-effectiveness acceptability curve (d) incremental net benefit empirical density.

5.5.2 Determining important structural errors via variance based sensitivity analysis

Following our analysis of structural error we may then wish to make improvements to the model. It is unlikely that all the sub-model discrepancy terms are equally 'important', by which we mean that some terms may be located in parts of the model in which structural errors contribute very little to uncertainty about Z, the incremental net benefit. If we can identify the most important discrepancy terms, we can consider reducing structural errors through better modelling, perhaps by relaxing certain assumptions, or by including features that were omitted initially. Similarly, identifying unimportant discrepancy terms will tell us where it is *not* worth improving the model.

Note that any re-modelling following a sensitivity analysis may not reduce uncertainty about Z, for example if the improved model structure introduces new, uncertain parameters. In this situation we are effectively 'transferring' our uncertainty from structure to inputs. This may be helpful simply because input uncertainty is generally easier to manage, but in any case we believe that a formal consideration of the balance between uncertainty due to model structure and uncertainty due to model inputs is desirable.

We can identify a set of important discrepancy terms using sensitivity analysis techniques. Various methods exist (as discussed in chapter 3), but for the purposes of this case study we have chosen to use a variance based sensitivity analysis approach. In this approach the measure of importance for each discrepancy term, δ_j $j = 1, \ldots, n$, is defined as its 'main effect index',

$$\frac{\mathrm{var}_{\delta_j}\{E(Z|\delta_j)\}}{\mathrm{var}(Z)}. \tag{5.54}$$

Given the identity $\mathrm{var}(Z) = \mathrm{var}_{\delta_j}\{E(Z|\delta_j)\} + E_{\delta_j}\{\mathrm{var}(Z|\delta_j)\}$ the numerator of the main effect index gives the expected reduction in the variance of Z obtained by learning the value of δ_j.

The main effect index for uncorrelated discrepancy terms is straightforward to calculate using Monte Carlo methods. In this case $E(Z|\delta_j)$ can be approximated by

$$E(Z|\delta_j) \simeq \frac{1}{S}\sum_{s=1}^{S} f^*(\boldsymbol{x}_s, \boldsymbol{\delta}_{-j,s}, \delta_j), \tag{5.55}$$

where $\{(\boldsymbol{x}_s, \boldsymbol{\delta}_{-j,s}), s = 1, \ldots, S\}$ is a (large) sample from the distribution $p(\boldsymbol{X}, \boldsymbol{\delta}_{-j})$.

However, if δ_j is correlated with other discrepancy terms or inputs, then this method would require us to draw samples from the conditional distribution $p(\boldsymbol{X}, \boldsymbol{\delta}_{-j}|\delta_j)$. Such conditional distributions may not be known, so we approximated the conditional expectation using a novel method that we describe in chapter 7.

Following a variance based sensitivity analysis of the discrepancy terms in our model, eight of the terms appeared to be important, having main effects $> 5\%$. The pattern of importance suggests that re-expressing the sub-functions for the parameters π_{id} is key to reducing structural error (table 5.3).

Table 5.3: Main effect indices for discrepancy terms ($>$ **5%** in bold)

Discrepancy	Main effect	Discrepancy	Main effect	Discrepancy	Main effect
$\delta_{r_1^{(ex)}}$	0.002	$\delta_{\pi_{1,1}}$	**0.266**	$\delta_{c^*_{chd}}$	0.002
$\delta_{r_2^{(ex)}}$	0.002	$\delta_{\pi_{2,1}}$	**0.128**	$\delta_{c^*_{str}}$	0.002
$\delta_{r_3^{(ex)}}$	0.003	$\delta_{\pi_{3,1}}$	**0.076**	$\delta_{c^*_{dm}}$	0.001
$\delta_{r_4^{(ex)}}$	0.002	$\delta_{\pi_{4,1}}$	0.002	$\delta_{d_{chd}}$	0.002
$\delta_{r_5^{(ex)}}$	0.003	$\delta_{\pi_{5,1}}$	**0.054**	$\delta_{d_{str}}$	0.002
$\delta_{r_6^{(ex)}}$	0.002	$\delta_{\pi_{6,1}}$	0.025	$\delta_{d_{dm}}$	0.002
$\delta_{r_7^{(ex)}}$	0.003	$\delta_{\pi_{7,1}}$	0.014	δ_{q_5}	0.002
$\delta_{r_8^{(ex)}}$	0.004	$\delta_{\pi_{8,1}}$	0.010	δ_{q_6}	0.002
$\delta_{r_1^{(sed)}}$	0.002	$\delta_{\pi_{1,2}}$	**0.257**	δ_{q_7}	0.002
$\delta_{r_2^{(sed)}}$	0.002	$\delta_{\pi_{2,2}}$	**0.124**	δ_{q_8}	0.002
$\delta_{r_3^{(sed)}}$	0.002	$\delta_{\pi_{3,2}}$	**0.076**	δ_{c_5}	0.002
$\delta_{r_4^{(sed)}}$	0.002	$\delta_{\pi_{4,2}}$	0.002	δ_{c_6}	0.002
$\delta_{r_5^{(sed)}}$	0.002	$\delta_{\pi_{5,2}}$	0.049	δ_{c_7}	0.002
$\delta_{r_6^{(sed)}}$	0.002	$\delta_{\pi_{6,2}}$	0.025	δ_{c_8}	0.002
$\delta_{r_7^{(sed)}}$	0.002	$\delta_{\pi_{7,2}}$	0.013	δ_{Δ_q}	0.003
$\delta_{r_8^{(sed)}}$	0.002	$\delta_{\pi_{8,2}}$	0.008	δ_{Δ_c}	0.001

We noted in section §3.4.1 that the main effect index does not measure the sensitivity of the *decision* to changes in the inputs, and we will address this problem in the next case study by computing the expected value of perfect information rather than the main effect index. If, however, the focus of a modelling effort is to predict an unknown quantity rather than to support a decision, then the variance based approach is entirely reasonable. In this case, we would also compute the total effect index for each input in order to ensure that inputs that have a small main effect index are not influential due to interactions with other inputs. The total effect index for X_i is

$$\frac{\text{var}(Y) - \text{var}_{\boldsymbol{X}_{-i}}\{E_{X_i}(Y|\boldsymbol{X}_{-i})\}}{\text{var}(Y)} = \frac{E_{\boldsymbol{X}_{-i}}\{\text{var}_{X_i}(Y|\boldsymbol{X}_{-i})\}}{\text{var}(Y)}, \tag{5.56}$$

where $\boldsymbol{X}_{-i}$ is the vector of all inputs *except* X_i. As discussed in section §3.4.1 the total effect index measures the overall effect of the input X_i, including any interactions. It is the expected variance (as a proportion of the total variance) that is left when all inputs *except* X_i are fixed. In general, the main effect index is useful in determining the effect of learning a single input, whereas the total effect index is useful in determining non-influential inputs.

5.5.3 The relative importance of parameter to structural error uncertainty

We may also wish to understand the relative importance of the contributions of uncertainty about structural error and uncertainty about input parameters to the overall uncertainty in Z. We can measure this using the *structural parameter uncertainty ratio*, which we define as

$$\frac{\text{var}_{\boldsymbol{\delta}}\{E_{\boldsymbol{X}}(Z|\boldsymbol{\delta})\}}{\text{var}_{\boldsymbol{X}}\{E_{\boldsymbol{\delta}}(Z|\boldsymbol{X})\}}. \tag{5.57}$$

This is straightforward to calculate if $\boldsymbol{\delta}$ is independent of $\boldsymbol{X}$ since $E_{\boldsymbol{X}}(Z|\boldsymbol{\delta} = \boldsymbol{\delta}') = E_{\boldsymbol{X}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta})|\boldsymbol{\delta} = \boldsymbol{\delta}'\} = E_{\boldsymbol{X}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta}')\}$ and $E_{\boldsymbol{\delta}}(Z|\boldsymbol{X} = \boldsymbol{x}) = E_{\boldsymbol{\delta}}\{f^*(\boldsymbol{X}, \boldsymbol{\delta})|\boldsymbol{X} = \boldsymbol{x}\} = E_{\boldsymbol{\delta}}\{f^*(\boldsymbol{x}, \boldsymbol{\delta})\}$. If $\boldsymbol{\delta}$ and $\boldsymbol{X}$ are not independent calculating the conditional expectations is more difficult, though methods are available (for example via the specification of a Gaussian process emulator, Oakley and O'Hagan, 2004).

The structural parameter uncertainty ratio in our model is 2.0 indicating that, given our specification of discrepancy, learning the discrepancy terms would result in double the expected reduction in the variance of the output compared with the expected reduction in variance on learning the true values of all the input parameters.

5.5.4 Analysis of robustness to different choices of v_j

In our case study we set v_j (the ratio of the discrepancy standard deviation to the mean of the corresponding intermediate parameter) to 5% equally for all discrepancy terms, judging this to be an appropriate reflection of the likely range of structural error. The resulting additional uncertainty in the model output was plausible, and the variance based sensitivity analysis implied that there was important structural error in the sub-model that generates the health state probability parameters, π_{id} (section §5.4.9).

In order to test the robustness of our conclusion to minor variations in the specification of the discrepancies we altered values for v_j over a plausible range. We grouped the discrepancy terms into four sets: terms relating to cost parameters, terms relating to health effect parameters, terms relating to population proportion parameters, and terms relating to the discount factors. Within each set the values for v_j were either doubled, halved or maintained at 5%. Given three levels for v_j and four sets of discrepancy terms there are $3^4 = 81$ combinations of choices for v_j including our original specification of $v_j = 5\%$ for all j.

In all 81 cases a very similar pattern of main effect indexes to that reported in table 5.3 was observed, with the $\delta_{\pi_{id}}$ terms dominating, indicating robustness to choices of v_j over the range 2.5% to 10%.

5.5.5 Remodelling the sub-functions where there is important structural error

Variance based sensitivity analysis has identified $\delta_{\pi_{id}}$ to be important discrepancy terms, indicating that we have important structural error in the sub-model that generates the health state probability parameters, π_{id}.

In the base case model the proportion of people who begin and then maintain exercise is assumed constant over time. If we believe that there will be a decline in the proportion of people who exercise over time then we could re-structure the model sub-function to reflect this. We could, for example, assume an exponential decline, whereby the proportion exercising at each year in the future is equal to the proportion exercising in the previous year multiplied by some (uncertain) constant. If the risk of each disease state i decreased (increased for the well state) linearly from $r_i^{(sed)}$ to $r_i^{(ex)}$ with increasing time spent exercising (with a threshold achieved after, say, four years exercise), then we could write

$$\begin{aligned}\pi_{id} &= \left(1-p_d^{(ex)}\right)r_i^{(sed)} + p_d^{(ex)}\left(1-m_d\right)r_i^{(sed)} \\ &+ p_d^{(ex)}\left(m_d - m_d^2\right)\left(\frac{1}{4}r_i^{(ex)} + \frac{3}{4}r_i^{(sed)}\right) + p_d^{(ex)}\left(m_d^2 - m_d^3\right)\left(\frac{1}{2}r_i^{(ex)} + \frac{1}{2}r_i^{(sed)}\right) \\ &+ p_d^{(ex)}\left(m_d^3 - m_d^4\right)\left(\frac{3}{4}r_i^{(ex)} + \frac{1}{4}r_i^{(sed)}\right) + p_d^{(ex)}m_d^4 r_i^{(ex)}, \qquad (5.58)\end{aligned}$$

where m_d is the proportion of the population who exercised in year t who continue to exercise in year $t+1$, under decision d.

To complete the new model specification we need to specify distributions for m_1 and m_2. After informal discussion with an expert we specified m_1 and m_2 as jointly normally distributed with means of 0.5, variances of 0.01 and a correlation of 0.9.

5.5.6 Results following sub-function remodelling

The mean net benefit following remodelling was £71 (95% credible interval -£273 to £572), with the probability that the intervention is cost-effective, $P(\text{INB} > 0)$, at $\lambda =$£20,000 equal to 0.59. Figure 5.5 shows the results after remodelling. We see that there is now a smaller cloud a points on the cost-effectiveness plane, and that these are shifted towards the left and the line of no effect (at $\Delta Q = 0$). The cost-effectiveness acceptability curve (figure 5.5c) suggests that following remodelling we predict that the intervention has a lower probability of being cost-effective than predicted by the base case model at all values of λ. The leftwards shift of the incremental net benefit density towards zero supports this (figure 5.5d).

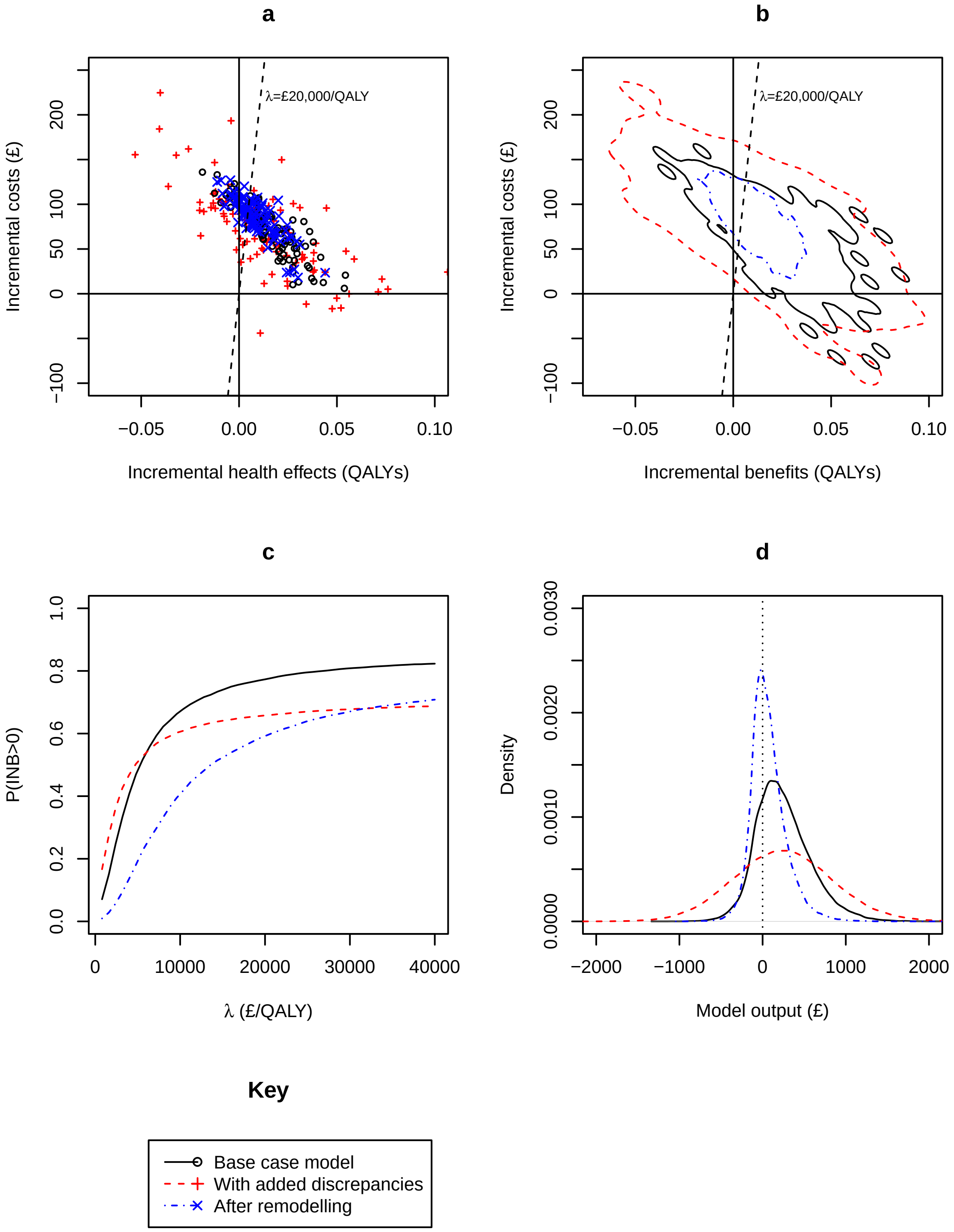


Figure 5.5: Results after remodelling as (a) cost-effectiveness plane (b) cost-effectiveness plane contour plot (c) cost-effectiveness acceptability curve (d) incremental net benefit empirical density.

By re-structuring the important sub-function in the model to better incorporate our beliefs about real-world processes, we find that the incremental net benefit distribution is shifted downwards. This is due to our judgement that a proportion of those who begin new exercise will cease exercising, and that instead of this drop being a single step change, the fall will be exponential over time. This results in a lower proportion of maintained exercise in both the intervention and non-intervention groups, and a lower absolute reduction in disease risk and smaller incremental benefit.

5.6 Conclusion

In this chapter we introduced a method for making judgements about model discrepancy at the level of the intermediate parameters within the model. We showed how it is possible to determine the subset of discrepancy terms that are important in driving model output uncertainty. This then allowed us to reconsider those parts of the model where the structure is uncertain, and where this uncertainty was an important contributor to output uncertainty. The resulting model better reflected our judgements about the underlying process. However, the method relies on the ability to meaningfully specify judgements about the model error at the sub-function level via $p(\boldsymbol{X}, \boldsymbol{\delta})$. This may not be easy. We return to this point in chapter 8.

In the next chapter we apply our method in a rather more complex model, a Markov model where the addition of time dependency considerably increases the number of discrepancy terms that must be specified. We recognise that the main effect index is not an adequate measure of importance for the discrepancy terms in a decision theoretic context and therefore compute the expected value of perfect information instead.

Chapter 6

Case Study 2 - Managing Structural Uncertainty in a Markov Model

6.1 Introduction

In this chapter we illustrate the application of the discrepancy method that we introduced in chapter 5 to another common type of health economic decision model, the Markov model. We imagine a scenario where we have built a relatively simple Markov model, but recognise that reality is more complex. We do not believe that even if we were to learn the 'true' values of the Markov transition probabilities and all other uncertain inputs in the model, that the predicted costs and health outcomes would equal their true values. We know that the model is a simplification, and we seek to answer the question 'is it good enough?'.

The Markov model that we use for this second case study predicts the costs and health effects of two competing treatments for HIV/AIDS. The time dependency that is present in the Markov model, but was absent from the decision tree model, presents us with the new challenge of specifying a joint distribution for a large number of discrepancy terms that are correlated in time, as well as within and between decision options. In order to parsimoniously specify the distribution on these terms we use a Gaussian process to represent our judgements about the model error.

In chapter 5 we used a variance-based sensitivity analysis to determine the relative importance of the discrepancy terms in driving the uncertainty in the output. In this chapter we adopt a decision theoretic approach and compute the expected value of learning the true values of the discrepancies. This avoids the need for the output of the model to be scalar, and more importantly tells us where uncertain model structure is likely to have an effect on the *decision*, rather than just on the model output. It is quite possible for a discrepancy term to be influential on the model output, but not to change the decision. If the value of learning the true values of the discrepancies is small compared with the expected value of learning the true values of inputs, then this offers reassurance that our current model is good enough for the decision at hand. We interpret the expected value of perfect information for the discrepancy terms as giving an upper bound on the *expected value of model improvement* (EVMI).

The chapter is organised as follows. In section §6.2 we introduce the case study: a simple Markov model designed to predict the costs and health effects of two competing treatment options for HIV/AIDS. In section §6.3 we apply the discrepancy analysis method in three scenarios that represent plausible sets of assumptions regarding the structural error. In section §6.4 we present results including the 'expected value of model improvement' (EVMI) in each scenario.

6.2 Case study model

In order to illustrate the method we introduce a case study that is based on a four state Markov model first described in Chancellor et al. (1997) and subsequently used for illustrative purposes in Drummond et al. (2005) and Briggs et al. (2006). The purpose of the model is to predict costs and health outcomes (in life years) under two decision options, zidovudine monotherapy versus zidovudine plus lamivudine combination therapy, in people with HIV. Allowable transitions between the four health states are shown in figure 6.1. The authors of the original paper chose time steps of 1 year and ran the model to a time horizon of 20 years.

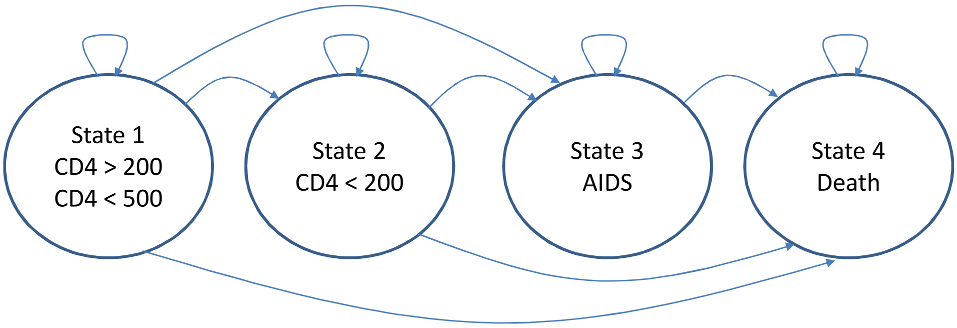


Figure 6.1: Structure of the case study Markov model

Note that this is no longer a credible model given the development of our understanding of the pathology of HIV/AIDS. Importantly, it is now understood that transitions can occur from worse to better states (see for example the models described in Miners et al., 2001; Simpson et al., 2004), transitions that are not possible in the Chancellor et al. (1997) model.

6.2.1 Notation

We index the monotherapy and combination therapy decision options $d = 1, 2$ respectively, the four mutually exclusive health states as $i = 1, \ldots, 4$, and the time steps in years as $t = 0, \ldots, 20$. If we imagine a cohort of people exposed to decision option d, we denote π_{djt} as the proportion of the cohort who are in health state i during time step t (alternatively, π_{djt} represents, under decision option d at time step t, the probability that a single individual exists in health state i versus the other states). We call $\boldsymbol{\pi}_{dt} = (\pi_{d1t}, \ldots, \pi_{d4t})'$ the *state vector* for decision option d at time step t, and note the constraint $\sum_{i=1}^{4} \pi_{djt} = 1 \; \forall d, t$.

We denote the costs and health effects accrued for health state i during time step t under decision d as c_{djt} and e_{djt} respectively. Costs and outcomes are time dependent to allow the discounting of costs and effects accrued in the future (Krahn and Gafni, 1993). We can therefore write costs and effects at time step t in terms of costs and effects at time zero via $c_{djt} = c_{di0}(1 + r_c)^{-t}$ and $e_{djt} = e_{di0}(1 + r_e)^{-t}$, where r_c and r_e are the per-year discount rate for costs and health effects. The health effect of interest for this decision problem is life years, so

$e_{di0} = 1$ for health states $i = 1, 2, 3$, and zero for the death state $i = 4$. We denote the vector of costs for all health states at time step t under decision d as $\boldsymbol{c}_{dt} = (c_{d1t}, \ldots, c_{d4t})'$, and the vector of health effects as $\boldsymbol{e}_{dt} = (e_{d1t}, \ldots, e_{d4t})'$.

6.2.2 The Markov model

The authors assumed a simple time-homogeneous Markov process (i.e. transition probabilities remain fixed for all time steps). Under this assumption the probability that an individual will move from health state x to health state y under decision d is given by p_{dxy}, and we note the constraints that $p_{dxy} \geq 0 \;\; \forall d, x, y$ and $\sum_{y=1}^{4} p_{dxy} = 1 \;\; \forall d, x$.

Transition from a worse health state to a better health state is considered impossible in this decision scenario. The transition matrix for the monotherapy ($d = 1$) option is therefore of the form,

$$\boldsymbol{M}_1 = \begin{pmatrix} p_{111} & p_{112} & p_{113} & p_{114} \\ 0 & p_{122} & p_{123} & p_{124} \\ 0 & 0 & p_{133} & p_{134} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.1}$$

where the lower diagonal elements are zero. Death is an absorbing state.

The matrix $\boldsymbol{M}_1$ is modified by the incorporation of a combination therapy treatment effect parameter, RR, to give the transition matrix for the combination therapy ($d = 2$) option,

$$\boldsymbol{M}_2 = \begin{pmatrix} 1 - RR(p_{112} + p_{113} + p_{114}) & RR \cdot p_{112} & RR \cdot p_{113} & RR \cdot p_{114} \\ 0 & 1 - RR(p_{123} + p_{124}) & RR \cdot p_{123} & RR \cdot p_{124} \\ 0 & 0 & 1 - RR \cdot p_{133} & RR \cdot p_{134} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6.2}$$

Given the transition matrix $\boldsymbol{M}_d$ and state vector $\boldsymbol{\pi}_{dt}$, we can generate $\boldsymbol{\pi}_{d,t+1}$ via the evolution equation

$$\boldsymbol{\pi}'_{d,t+1} = \boldsymbol{\pi}'_{dt}\boldsymbol{M}_d, \tag{6.3}$$

and we can therefore express $\boldsymbol{\pi}_{dt}$ in terms of the state vector at time step 0, i.e. $\boldsymbol{\pi}'_{dt} = \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t$, where $\boldsymbol{M}_d^t = \prod_{l=1}^{t} \boldsymbol{M}_d$.

If we value (in cost units) one unit of health outcome at λ, our final model for the net monetary benefit associated with decision option d is

$$NB_d = \lambda e_d^{tot} - c_d^{tot} = \lambda \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t \boldsymbol{e}_{dt} - \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}\boldsymbol{M}_d^t \boldsymbol{c}_{dt}. \tag{6.4}$$

Assuming that we are uncertain about some or all the inputs into the model, our optimum decision is that which maximises the expected net benefit.

6.2.3 Base case input parameter values

Transition probabilities, costs and the treatment effect parameter are all considered uncertain in the base case model, with distributions shown in tables 6.1 and 6.2.

Table 6.1: Transition probability distributions for $d = 1$

$(p_{111}, p_{112}, p_{113}, p_{114})$	$\sim$ Dirichlet (1251,350,115,14)
$(p_{121}, p_{122}, p_{123}, p_{124})$	$\sim$ Dirichlet (0,731,512,15)
$(p_{131}, p_{132}, p_{133}, p_{134})$	$\sim$ Dirichlet (0,0,1312,437)
$(p_{141}, p_{142}, p_{143}, p_{144})$	$= (0, 0, 0, 1)$

Table 6.2: Cost and relative risk distributions

Label	Description	Distribution	Mean	SD
cc_1	Undiscounted care costs of 1 time step in state 1 (£)	normal	2756	400
cc_2	Undiscounted care costs of 1 time step in state 2 (£)	normal	3052	437
cc_3	Undiscounted care costs of 1 time step in state 3 (£)	normal	9007	1449
RR	Treatment effect (combi vs monotherapy)	lognormal	log(0.509)	0.05

Drug treatment costs are considered fixed and known, as are discount rates (table 6.3). We assume that the combination therapy is effective throughout the whole of the modelled 20 year period, rather than just for the first year (this is presented as an alternative scenario rather than the base case in Chancellor et al., 1997).

Table 6.3: Fixed inputs

Label	Description	Value
c_Z	Zidovudine cost (£)	2278
c_L	Lamivudine cost (£)	2087
r_c	Discount rate for costs	3.5% per year
r_e	Discount rate for outcomes	3.5% per year

6.3 Discrepancy analysis

6.3.1 Incorporating judgements about model structural error into the Markov model

We believe that the transition of individuals through health states is not adequately described by the simple 'base case' time-homogeneous Markov model described above, and therefore expect there to be error in the predicted costs and health effects. We wish to quantify this structural error to determine whether we need to build a more complex model. In particular we wish to determine the expected value of improving the model. We restrict ourselves in this case study to considering only structural error that relates to the Markov model itself. In many applications a Markov model is part of a larger model that may also include, for example, a decision tree element where we may also judge there to be structural error.

We recognise that there are many potential sources of structural error in our base case model, given its simplicity, and the new knowledge that has accumulated in the 15 years between the development of the model and now. In our analysis we will explore three potential sources of error in order to illustrate our method. Clearly, the three scenarios that we present are in no way exhaustive.

We introduce a series of discrepancy terms, each of which represents the difference between the output of a sub function in the built model and the true value of that output quantity. Discrepancy terms are incorporated in the model at the level of the evolution of the health state vector, replacing equation (6.3) with

$$\boldsymbol{\pi}'_{dt} = \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_d + \boldsymbol{\delta}_{dt}, \tag{6.5}$$

where $\boldsymbol{\delta}_{dt}$ is a vector of discrepancy terms that quantifies the error in the state vector at time t for decision option d.

In the analysis for our case study we have found it more intuitive to think about discrepancies as applying to the transition matrix rather than to the state vector, writing $\boldsymbol{\delta}_{dt} = \boldsymbol{\pi}'_{d,t-1}\boldsymbol{\Delta}_{dt}$ and expressing judgements about the model error via $\boldsymbol{\Delta}_{dt}$, a matrix of discrepancy terms of the same dimensionality as $\boldsymbol{M}_d$. We re-express equation (6.5) as

$$\begin{aligned} \boldsymbol{\pi}'_{dt} &= \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_d + \boldsymbol{\delta}_{dt}, \\ &= \boldsymbol{\pi}'_{d,t-1}\boldsymbol{M}_{dt} + \boldsymbol{\pi}'_{d,t-1}\boldsymbol{\Delta}_{dt}, \\ &= \boldsymbol{\pi}'_{d,t-1}(\boldsymbol{M}_d + \boldsymbol{\Delta}_{dt}). \end{aligned} \tag{6.6}$$

The matrix $(\boldsymbol{M}_d+\boldsymbol{\Delta}_{dt})$ must obey the same constraints as $\boldsymbol{M}_d$, i.e. all elements must lie within the interval $[0, 1]$ and each row must sum to one. We can ensure this if each element of $\boldsymbol{\Delta}_{dt}$, δ_{dtxy}, is constrained to lie in the interval $[-p_{dxy}, 1 - p_{dxy}]$, and if each row of $\boldsymbol{\Delta}_{dt}$ sums to zero.

Given the transition probability matrices (equations 6.1 and 6.2), there are potentially six such unconstrained discrepancy terms per decision option per time step, and we denote these δ_{djt}, $j = 1, \ldots, 6$. The discrepancy matrix $\boldsymbol{\Delta}_{dt}$ is therefore

$$\boldsymbol{\Delta}_{dt} = \begin{pmatrix} -(\delta_{d1t} + \delta_{d2t} + \delta_{d3t}) & \delta_{d1t} & \delta_{d2t} & \delta_{d3t} \\ 0 & -(\delta_{d4t} + \delta_{d5t}) & \delta_{d4t} & \delta_{d5t} \\ 0 & 0 & -\delta_{d6t} & \delta_{d6t} \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{6.7}$$

We may judge that structural error relates only to a subset of the transitions in the model. Where we judge there to be no structural error the corresponding discrepancy term will be zero.

6.3.2 Scenario 1 - time dependent transition probabilities

In the first scenario of our case study we judge that there is an important time dependent relationship between age and the probability of death that is not captured in the simple time homogeneous model. We therefore introduce three discrepancy terms (per time step per decision), one for each transition from an alive state to the death state. Given the general expression for the discrepancy matrix in equation (6.7) we expect that δ_{djt} is non-zero for $j = 3, 5, 6$, and zero for $j = 1, 2, 4$. Given three discrepancy terms per decision option per time step there are $3 \times 2 \times 21 = 126$ discrepancy terms in total. Specifying judgements about the model discrepancy via the joint distribution of such a large number of terms clearly requires a parsimonious parametrisation that reflects the dependencies between discrepancy terms.

To illustrate our approach to this specification problem we consider the discrepancy term, δ_{d6t}, that describes the structural error in the built model with respect to the probability of transition from AIDS to death. We judge that the probability of this transition increases monotonically over time rather than being constant in the base case model, but we are unsure as to the exact nature of the relationship between the probability of death and time. This belief implies that the uncertain discrepancy term δ_{d6t} must also increase monotonically with respect to time. We judge that at $t = 0$ the probability of death may be approximately 20% lower than the constant value (0.250) in the built model, and at $t = 20$ may be approximately 20% higher, but we have considerable uncertainty. Figure 6.2 represents some plausible realisations of the discrepancy δ_{d6t} as a function of time for $d = 1$.

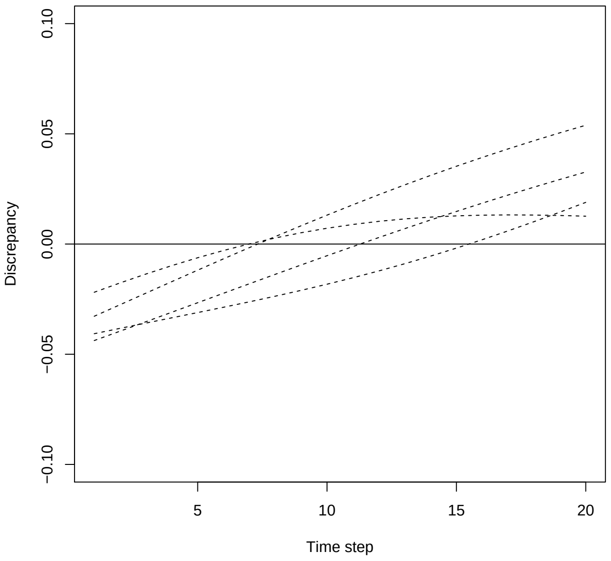


Figure 6.2: Four plausible realisations of the discrepancy term δ_{d6t} for $d = 1$ in scenario 1

6.3.3 Parametrising the discrepancy using a Gaussian process

We wish to find a convenient and parsimonious parametrisation for the joint distribution of the 126 discrepancy terms δ_{djt}, $d = 1, 2$, $j = 3, 5, 6$, and $t = 0, \ldots, 20$. We begin by noting that the reason for choosing a Markov model structure for our built model was to reflect a dynamic time dependent process, so it seems reasonable to consider discrepancy as a function of time step, i.e. $\delta_{djt} = f_{dj}(t)$. We then assume that the functions $f_{dj}(t)$ follow a Gaussian process, i.e. that $\{f_{1,1}(0), \ldots, f_{2,6}(20)\}$ has a multivariate normal distribution with mean function, $E\{f_{dj}(t)\} = m(d, j, t)$ and covariance function $\text{Cov}\{f_{dj}(t), f_{d^*j^*}(t^*)\} = c(d, j, t, d^*, j^*, t^*)$.

This highly flexible and parsimonious parametrisation of set of unknown functions allows us to specify not only our uncertainty about each δ_{djt}, but also the correlation structure of discrepancies through time, the correlations between the three non-zero discrepancy terms per decision, and the correlation between the discrepancy terms for the $d = 1, 2$ decisions for each transition j.

Gaussian processes are well understood and have many attractive properties, hence their widespread use in the modelling of unknown complex processes (their use is ubiquitous in the Managing Uncertainty in Complex Models project for example, see `www.mucm.ac.uk`). However, we must remember that the Gaussian process is only a model for the unknown discrepancy 'process', and its attractiveness may be illusory. Does the Gaussian process with the mean and covariance functions that we specify really represent our beliefs about the discrepancies? It may be very difficult to answer this.

We may judge that an alternative representation of our beliefs about the discrepancies is more appropriate. Other options for characterising the joint distribution of the discrepancies on the transition matrix would include the use of a copula (Possolo, 2010). A copula is a device for generating a multivariate distribution that has univariate margins with certain defined properties. So, in our case we might want to define a copula with a Beta distributed margin corresponding to each discrepancy. Our beliefs about the correlation between discrepancies would be specified by the hyperparameters of a multivariate normal distribution. The copula can be thought of as the function that links the univariate Beta marginals to the underlying multivariate normal distribution.

A second possibility, given the sum-to-one constraint on the rows of the transition matrix, is the specification of a series of Dirichlet distributions. However, the difficulty with this approach is the problem of describing correlations *between* rows and over time. This is not straightforward. It may be possible to combine these latter two approaches and specify a copula with Dirichlet marginals; an interesting area for further work.

Specifying the mean function

We specify the mean for each discrepancy, $E\{f_{dj}(t)\}$, as a function $m(d, j, t)$. For scenario 1 a linear form, $E\{f_{dj}(t)\} = m(d, j, t) = \beta_{0,dj} + \beta_{1,dj}t$, adequately reflects our judgements, but depending on the decision problem alternative choices might be higher order polynomial, $E\{f_{dj}(t)\} = m(d, j, t) = \beta_{0,dj}+, \ldots, +\beta_{n,dj}t^n$, exponential, $E\{f_{dj}(t)\} = m(d, j, t) = \beta_{0,dj} + \beta_{1,dj}\exp(\beta_{2,dj}t)$, or stepped, $E\{f_{dj}(t)\} = m(d, j, t) = \beta_{0,dj} + \beta_{1,dj}I(t > \beta_{2,dj})$. We placed normal distributions on the linear mean function parameters $\beta_{0,dj}$ and $\beta_{1,dj}$ with hyperparameters shown in table 6.4.

Table 6.4: Hyperparameters to specify GP mean function

Hyperparameter	Scenario 1	Scenario 2	Scenario 3	Transition
Intercept ($\beta_{0,dj}$)	Mean (sd) $\times 10^{-3}$	Mean (sd) $\times 10^{-3}$	Mean (sd) $\times 10^{-3}$	state x to y; decision
$\beta_{0,11}$	0 (0)	0 (0)	0 (0)	1 to 2; monotherapy
$\beta_{0,12}$	0 (0)	0 (0)	0 (0)	1 to 3; monotherapy
$\beta_{0,13}$	-1.0 (0)	0 (0)	0 (0)	1 to 4; monotherapy
$\beta_{0,14}$	0 (0)	0 (0)	0 (0)	2 to 3; monotherapy
$\beta_{0,15}$	-1.2 (0)	0 (0)	0 (0)	2 to 4; monotherapy
$\beta_{0,16}$	-25.0 (0)	0 (0)	0 (0)	3 to 4; monotherapy
$\beta_{0,21}$	0 (0)	0 (0)	0 (0)	1 to 2; combi therapy
$\beta_{0,22}$	0 (0)	0 (0)	0 (0)	1 to 3; combi therapy
$\beta_{0,23}$	-0.5 (0)	0 (0)	0 (0)	1 to 4; combi therapy
$\beta_{0,24}$	0 (0)	0 (0)	0 (0)	2 to 3; combi therapy
$\beta_{0,25}$	-0.61 (0)	0 (0)	0 (0)	2 to 4; combi therapy
$\beta_{0,26}$	-12.7 (0)	0 (0)	0 (0)	3 to 4; combi therapy
Slope ($\beta_{1,dj}$)	Mean (sd) $\times 10^{-4}$	Mean (sd) $\times 10^{-4}$	Mean (sd) $\times 10^{-4}$	state x to y; therapy
$\beta_{1,11}$	0 (0)	0 (0)	0 (0)	1 to 2; monotherapy
$\beta_{1,12}$	0 (0)	0 (0)	0 (0)	1 to 3; monotherapy
$\beta_{1,13}$	1.0 (0)	0 (0)	0 (0)	1 to 4; monotherapy
$\beta_{1,14}$	0 (0)	0 (0)	0 (0)	2 to 3; monotherapy
$\beta_{1,15}$	1.2 (0)	0 (0)	0 (0)	2 to 4; monotherapy
$\beta_{1,16}$	25.0 (0)	0 (0)	0 (0)	3 to 4; monotherapy
$\beta_{1,21}$	0 (0)	24.8 (13.78)	0 (0)	1 to 2; combi therapy
$\beta_{1,22}$	0 (0)	8.2 (4.57)	0 (0)	1 to 3; combi therapy
$\beta_{1,23}$	0.51 (0)	1.2 (0.68)	0 (0)	1 to 4; combi therapy
$\beta_{1,24}$	0 (0)	50.0 (27.8)	0 (0)	2 to 3; combi therapy
$\beta_{1,25}$	0.61 (0)	1.5 (0.82)	0 (0)	2 to 4; combi therapy
$\beta_{1,26}$	12.7 (0)	30.7 (17.0)	0 (0)	3 to 4; combi therapy

Specifying the covariance function

We make a number of simplifying assumptions when specifying the covariance function, but note that all of these assumptions may be relaxed at the cost of specifying a greater number of hyperparameters. We assume in scenario 1 that the variance of each discrepancy δ_{djt} remains constant for all t, requiring the specification of $2\times 3 = 6$ variances, which we denote σ^2_{dj}. We state beliefs about the within-decision, between-transition term correlation through a parameter $\phi_{j,j^*} = \text{cor}(\delta_{djt}, \delta_{dj^*t})$, assuming that this is constant over time t and across decisions. We state beliefs about the between-decision correlation through a parameter $\psi_{d,d^*} = \text{cor}(\delta_{djt}, \delta_{d^*jt})$, assuming that this is constant over time t and across transitions j.

Finally we state beliefs about the correlation of the discrepancies through time by defining a correlation function $\rho(\cdot,\cdot)$ that depends on the distance between time steps, assuming this holds for all d and j. For the purposes of scenario 1 we use the 'Gaussian form'

$$\rho(t,t^*) = \exp\left\{-\left(\frac{t-t^*}{\omega}\right)^2\right\}, \tag{6.8}$$

where ω is the correlation length. The correlation length determines the degree of correlation between discrepancy terms at any particular 'distance', where distance is the number of Markov time steps between the terms. See Neal (1999) for a discussion of alternatives to this simple Gaussian form of correlation function.

The overall covariance function is therefore

$$\begin{aligned}\mathrm{Cov}\{f_{dj}(t), f_{d^*j^*}(t^*)\} &= c(d,j,t,d^*,j^*,t^*),\\ &= \sigma_{dj}\sigma_{d^*j^*}\psi_{d,d^*}\phi_{j,j^*}\rho(t,t^*),\\ &= \sigma_{dj}\sigma_{d^*j^*}\psi_{d,d^*}\phi_{j,j^*}\exp\left\{-\left(\frac{t-t^*}{\omega}\right)^2\right\}. \end{aligned} \tag{6.9}$$

Finally we specify a correlation structure for the discrepancies as they evolve through time via the correlation function with parameter ω (equation 6.8). Values chosen are shown in table 6.5.

Table 6.5: Hyperparameters to specify GP covariance function

Variance hyperparameters (σ_{jd})	Scenario 1 ($\times 10^{-3}$)	Scenario 2 ($\times 10^{-3}$)	Scenario 3 ($\times 10^{-3}$)	Transition
σ_{11}	0	0	28.9	A to B monotherapy
σ_{12}	0	0	9.6	A to C monotherapy
σ_{13}	1.0	0	1.4	A to D monotherapy
σ_{14}	0	0	58.1	B to C monotherapy
σ_{15}	1.2	0	1.7	B to D monotherapy
σ_{16}	25.0	0	35.7	C to D monotherapy
σ_{21}	0	5.1	14.7	A to B combi therapy
σ_{22}	0	1.7	4.9	A to C combi therapy
σ_{23}	0.51	0.25	0.7	A to D combi therapy
σ_{24}	0	10.4	29.6	B to C combi therapy
σ_{25}	0.61	0.31	0.9	B to D combi therapy
σ_{26}	12.7	6.4	18.1	C to D combi therapy

Correlation hyperparameters	Scenario 1	Scenario 2	Scenario 3	Description
ϕ	0.8	0.9	0	Between discrepancy term correlation
ψ	0.9	0	0	Between decision correlation
ω	32	7	7	Correlation length parameter

Monotonicity constraint for $f_{dj}(t)$

We have chosen the Gaussian process as a method to model the discrepancy terms, with discrepancy with respect to time being considered an uncertain function,

$f_{dj}(t)$. We may wish to constrain the form of $f_{dj}(t)$, and in particular we may wish to ensure that $f_{dj}(t)$ is monotone with respect to t to reflect our belief that the probability of death increases with time. However, realisations of a Gaussian process tend to be 'wiggly' non-monotone functions, with the degree of 'wiggliness' controlled by the ω parameter. Increasing values of ω will result in an increasingly smooth functions, so by carefully choosing ω we can ensure that the realisations of the Gaussian process are constrained to be monotone to reflect our beliefs about the relationship between discrepancy and time.

Monotonicity with respect to t implies that, for a once differentiable function $f_{dj}(t)$ that $\partial f_{dj}(t)/\partial t > 0\ \forall t$, or $\partial f_{dj}(t)/\partial t < 0\ \forall t$. Informally then, we can ensure monotonicity by choosing hyperparameters for the mean and covariance functions such that this holds with some probability α.

It is a property of an n times differentiable Gaussian process $f(x) \sim GP\{m(x), c(x, x^*)\}$ with n times differentiable mean and covariance functions, that $\partial^n f(x)/\partial x^n$ is also a Gaussian process with mean function

$$E\left\{\frac{\partial^n}{\partial x^n} f(x)\right\} = \frac{\partial^n}{\partial x^n} m(x), \tag{6.10}$$

and covariance function

$$\operatorname{cov}\left\{\frac{\partial^n}{\partial x^n} f(x)\Big|_{x=x}, \frac{\partial^n}{\partial x^n} f(x)\Big|_{x=x^*}\right\} = \frac{\partial^{2n}}{\partial x^n \partial x^{*n}} c(x, x^*). \tag{6.11}$$

See O'Hagan (1992) for further details.

This implies that $\partial f_{dj}(t)/\partial t$ is the Gaussian process,

$$\frac{\partial}{\partial t} f_{dj}(t) \sim GP\left\{\frac{\partial}{\partial t} m(d, j, t), \frac{\partial^2}{\partial t \partial t^*} c(d, j, t, d, j, t^*)\right\}, \tag{6.12}$$

and we can ensure monotonicity of $f_{dj}(t)$ with some pre-specified probability α by choosing parameters of $m(\cdot)$ and $c(\cdot, \cdot)$ such that

$$\left|\frac{\partial}{\partial t} m(d, j, t)\right| - \Phi^{-1}(\alpha) \sqrt{\frac{\partial^2}{\partial t \partial t^*} c(d, j, t, d, j, t^*)} > 0, \tag{6.13}$$

where $\Phi^{-1}(\alpha)$ is the inverse normal cumulative distribution function (Montes Diez and Oakley, 2010).

Given a linear mean function, $m(d,j,t) = \beta_{0,dj} + \beta_{1,dj}t$, and a Gaussian form for the correlation function with respect to time, $c(d,j,t,d,j,t^*) = \sigma^2_{dj} \exp\left\{-\left(\frac{t-t^*}{\omega}\right)^2\right\}$, equation (6.13) becomes

$$\left|\beta_{1,dj}\right| - \Phi^{-1}(\alpha)\sqrt{\frac{2\sigma^2_{dj}}{\omega^2}} > 0, \tag{6.14}$$

which by solving for ω gives

$$\omega > \frac{\Phi^{-1}(\alpha)\sqrt{2}\sigma_{dj}}{\beta_{1,dj}}. \tag{6.15}$$

We can therefore, given $\beta_{1,dj}$ and σ^2_{dj}, ensure with some probability α that $f_{dj}(t)$ is monotone through a choice of correlation length parameter ω that obeys (6.15). If $\beta_{1,dj}$ is itself uncertain this approach is more difficult. In this case we may choose ω such that

$$\omega > \frac{\Phi^{-1}(\alpha)\sqrt{2}\sigma_{dj}}{\beta^{0.025}_{1,dj}}, \tag{6.16}$$

where $\beta^{0.025}_{1,dj}$ is the value of the 2.5th centile of the distribution of the random variable $\beta_{1,dj}$.

For scenario 1, $\beta_{1,dj}$ was considered known with certainty allowing us to use (6.15). We set $\alpha = 0.95$ and chose ω accordingly.

We must keep in mind that this approach relies on properties of the Gaussian process, which is only a model for our unknown function. We do not know for certain that the 'true' function that describes the relationship between discrepancy and time is smooth n times differentiable. Indeed, if we have reason to believe that it is not, then choosing a Gaussian process representation of the unknown function may well be inappropriate.

Ten samples from the Gaussian process for discrepancy term $\delta_{1,6,t}$ are shown in figure 6.3. Note the variation in functional form generated by the Gaussian process, reflecting our uncertainty about the relationship between probability of death and time, but with the constraint that the relationship between discrepancy and time should be monotone.

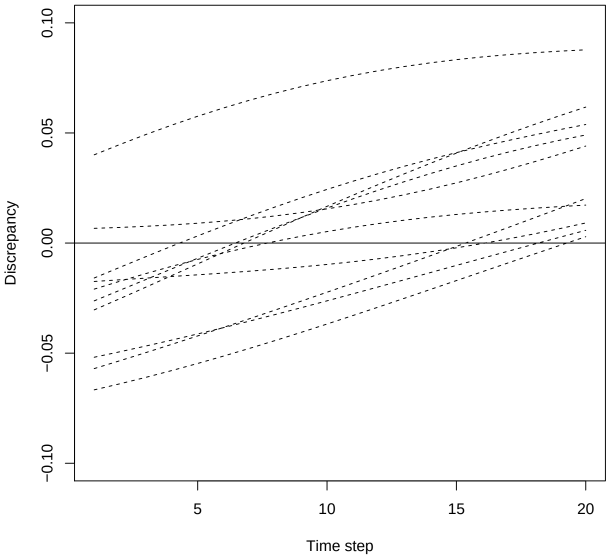


Figure 6.3: Ten samples from the distribution on discrepancy term $\delta_{1,6,t}$ in scenario 1

6.3.4 Sensitivity analysis to determine whether the discrepancies make any difference to the decision

Given our specification of discrepancy for our built model we can determine whether we should build a more complex model by examining the sensitivity of the decision to the discrepancy. We calculate, using standard Monte Carlo methods, the expected value of learning the true value of the discrepancy terms via the partial expected value of perfect information (EVPI),

$$\text{EVPI}(\boldsymbol{\delta}) = E_{\boldsymbol{\delta}}\{\max_d E_{\boldsymbol{X}|\boldsymbol{\delta}}(NB_d)\} - \max_d E(NB_d),$$

where $\boldsymbol{X}$ is the vector of model inputs, and $\boldsymbol{\delta}$ is the vector of discrepancy terms (see chapter 3 for discussion of EVPI). If EVPI($\boldsymbol{\delta}$) is large compared with the value of learning the inputs, EVPI($\boldsymbol{X}$), then we conclude that the potential structural error in adopting the simple Markov model is important.

The expected value of learning the discrepancy terms, EVPI($\boldsymbol{\delta}$), is the 'expected value of model improvement' (EVMI) under the assumption that any new input parameters that are introduced into the model during the structural improvement are known with certainty. It is likely however that model improvement will involve the addition of new *uncertain* input parameters. In this case the EVPI($\boldsymbol{\delta}$) provides an upper bound for the EVMI. If the EVPI($\boldsymbol{\delta}$) is small this offers us some reassurance that the model is good enough for the decision, whereas if it is large we know our uncertainty about the model structure is resulting in decision uncertainty. In the latter case improving the model may be worthwhile, but this will depend on the degree of decision uncertainty induced by any newly introduced uncertain inputs.

6.3.5 Scenario 2 - an uncertain relationship between efficacy and time since treatment commencement

The duration of effect of the combination therapy was a key uncertainty at the time of publication of Chancellor et al. (1997), and the authors presented results for three alternative scenarios: effectiveness lasting one year, two years and 20 years. We ask the following question: if our built model assumes that the combination therapy is effective over 20 years, but we are uncertain whether this is true, do we need to build a more complex model that incorporates an uncertain relationship between efficacy and time from commencement of treatment?

The treatment effect acts on six unconstrained terms in the transition matrix for the combination therapy (equation 6.2), but does not act on the transition matrix for the monotherapy, therefore resulting in six non-zero discrepancies per time step, $\delta_{2,1,t}, \ldots, \delta_{2,6,t}$. This specification of discrepancy is equivalent to incorporating a time varying treatment effect parameter (RR), but with the additional flexibility that allows the treatment effect to vary across the different transitions in the model (e.g. HIV to AIDS versus HIV to death).

We believe that efficacy falls over time, and therefore that the discrepancy between our built model and reality increases over time. We again chose a linear mean function $E(\delta_{2jt}) = \beta_{0,i} + \beta_{1,j}t$ with uncertain slope. The intercept parameter $\beta_{0,j}$ is zero in this case to reflect our judgement that during time step 1 the treatment effect parameter RR correctly determines the effectiveness of the combination therapy. We placed normal distributions on the six $\beta_{1,j}$ parameters with hyperparameters $\mu_{\beta_{1,j}}$ and $\sigma^2_{\beta_{1,j}}$ shown in table 6.4.

Next, we specify the covariance function. Our uncertainty about the six discrepancies $\delta_{2,1,t}, \ldots, \delta_{2,6,t}$ is controlled through variance terms $\sigma^2_{2,1}, \ldots, \sigma^2_{2,6}$, assumed to hold for all t. We specify our judgement about the dependency between the discrepancy terms for the six transitions through a single correlation parameter $\phi_{j,j^*} = \phi \; \forall j \neq j^*$ which we assume constant for all t. Since there is no discrepancy for the monotherapy option $d = 1$ in this scenario we do not need to specify between-decision correlations (i.e. there is no ψ_{d,d^*} correlation parameter). Finally we specify a correlation structure for the discrepancies as they evolve through time via a Gaussian form correlation function with parameter ω (equation 6.8), ensuring via equation (6.15) that discrepancy as a function of time is monotone with probability $\alpha = 0.95$. Values for all covariance function parameters are shown in table 6.5.

Ten samples from the Gaussian process for discrepancy term $\delta_{2,1,t}$ are shown in figure 6.4. Note the variation in functional form generated by the Gaussian process, reflecting our uncertainty about the relationship between efficacy and time.

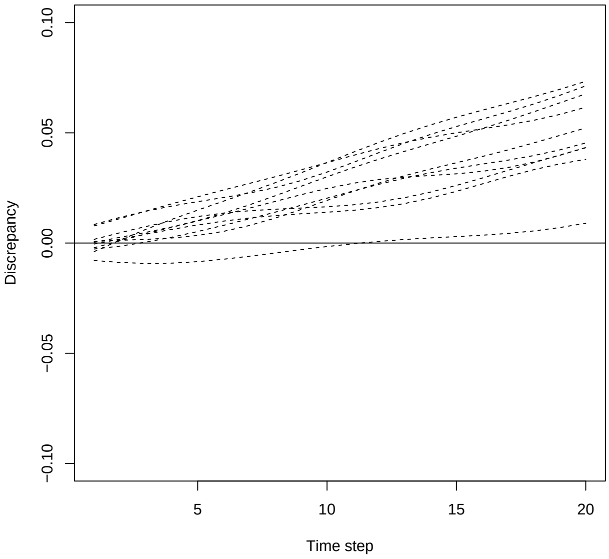


Figure 6.4: Ten samples from the distribution on discrepancy term $\delta_{2,1,t}$ in scenario 2

6.3.6 Scenario 3 - relaxation of the memoryless property

In scenario 3 we judge that the probability of transition to state y at time step $t+1$ is dependent not only on the state x at time t but on the states occupied at time steps $\leq t-1$. We therefore want to relax the Markov assumption and consider more complex time dependencies that would necessitate a more flexible modelling framework (for example using a discrete event or agent based approach). In order to judge whether this is necessary we add relatively unstructured discrepancy to allow for a wide range of possible deviations from the simple memoryless Markov process. Hyperparameters are shown in tables 6.4 and 6.5. Ten samples from the Gaussian process for discrepancy term $\delta_{2,1,t}$ are shown in figure 6.5.

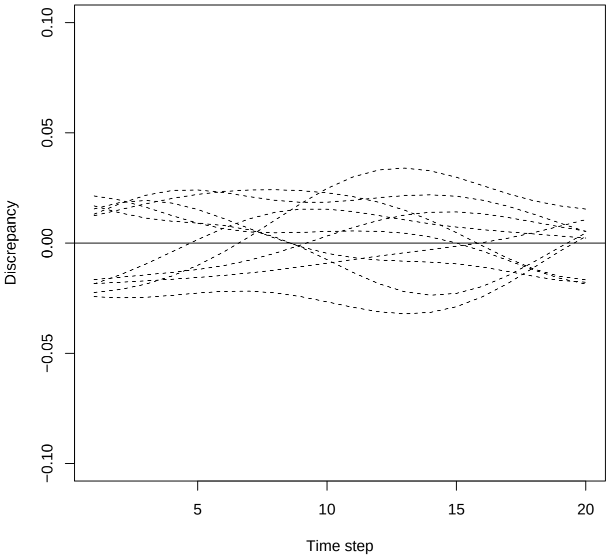


Figure 6.5: Ten samples from the distribution on discrepancy term $\delta_{2,1,t}$ in scenario 3

6.4 Results

6.4.1 Base case model

We implemented the model in R (R Development Core Team, 2011). We sampled from the base case model input parameters and ran the model 10,000 times. The mean incremental cost of combination therapy over monotherapy was £45,402 and the mean incremental benefit was 3.86 life years, representing a cost per life year gained of £11,749. Figure 6.6 shows the cost-effectiveness plane and cost-effectiveness acceptability curve (CEAC) for the base case, representing the uncertainty due to uncertainty in the model inputs.

Value of information analysis with $\lambda = £12,000$ per life year[1] suggests that decision uncertainty is being driven by uncertainty in the treatment effect parameter with EVPI(RR)=£169.91 (EVPI index, 46.5%), and uncertainty in the

[1] We assumed for the purposes of this case study that the value of one QALY is $\lambda = £12,000$ per life year to ensure that we were in the region of decision uncertainty. This is lower than the 'threshold' value that would be used for decisions in many Western health economies. At $\lambda = £30,000$ there is almost no decision uncertainty with EVPI negligible for all inputs and discrepancies.

cost parameters with EVPI(costs)=£194.41 (53.2%). See table 6.6. It is not the case that the partial EVPI values necessarily sum to the overall EVPI. The partial EVPI values are expressed as percentages of overall EVPI (the ‘EVPI index’) merely to aid comparison of their relative sizes.

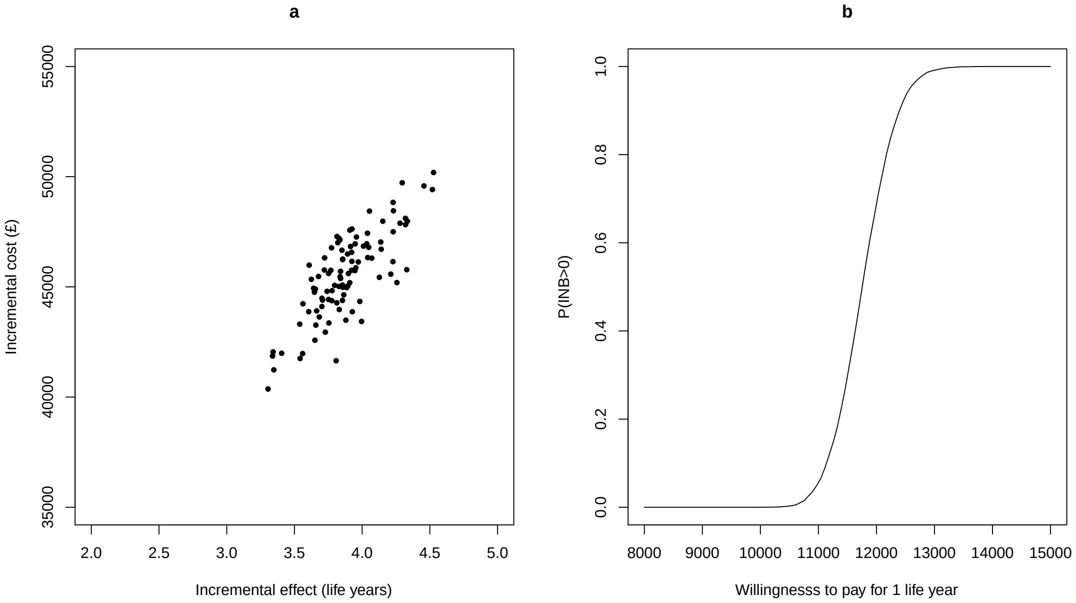


Figure 6.6: (a) cost-effectiveness plane (b) CEAC for base case model

Table 6.6: Partial EVPI results

Parameter	Partial EVPI (EVPI index†)			
	Base case	Scenario 1	Scenario 2	Scenario 3
Transition probabilities	£0 (0%)	£0 (0%)	£0 (0%)	£1.17 (0.1%)
Relative risk	£169.91 (46.5%)	£193.09 (48.1%)	£64.63 (19.4%)	£164.55 (17.2%)
Costs	£194.41 (53.2%)	£201.72 (50.2%)	£65.17 (19.55%)	£167.53 (17.5%)
Discrepancy terms	-	£7.86 (2.0%)	£110.21 (33.1%)	£699.06 (73.0%)
Overall EVPI	£365.42	£401.53	£333.43	£957.28

† The partial EVPI as a proportion of the overall EVPI

6.4.2 Scenario 1

After the addition of discrepancy to reflect the judgements about model error due to the time homogeneity assumption, the mean incremental cost of combination therapy over monotherapy was £44,697 and the mean incremental benefit was 3.80 life years, representing a cost per life year gained of £11,769. Figure 6.7 shows the cost-effectiveness plane and cost-effectiveness acceptability curve (CEAC) for scenario 1, overlaid on those for the base case. Note the similarity between scenario 1 and the base case suggesting that the discrepancy terms have not introduced significant new uncertainty.

Value of information analysis suggests that the decision uncertainty is still dominated by the uncertainty in the inputs with EVPI(RR)=£193.09 (48.1%) and EVPI(costs)=£201.72 (50.2%). There is little value in learning $\boldsymbol{\delta}$ with EVPI($\boldsymbol{\delta}$) =£7.86 (2.0%), indicating that building a more complex model is not advisable at a willingness to pay for one life year of $\lambda = \pounds 12,000$.

It appears that uncertainty regarding the model error that results from the time homogeneity assumption is not a significant driver of decision uncertainty. This would suggest that our simple built model is 'good enough' for the decision in this scenario.

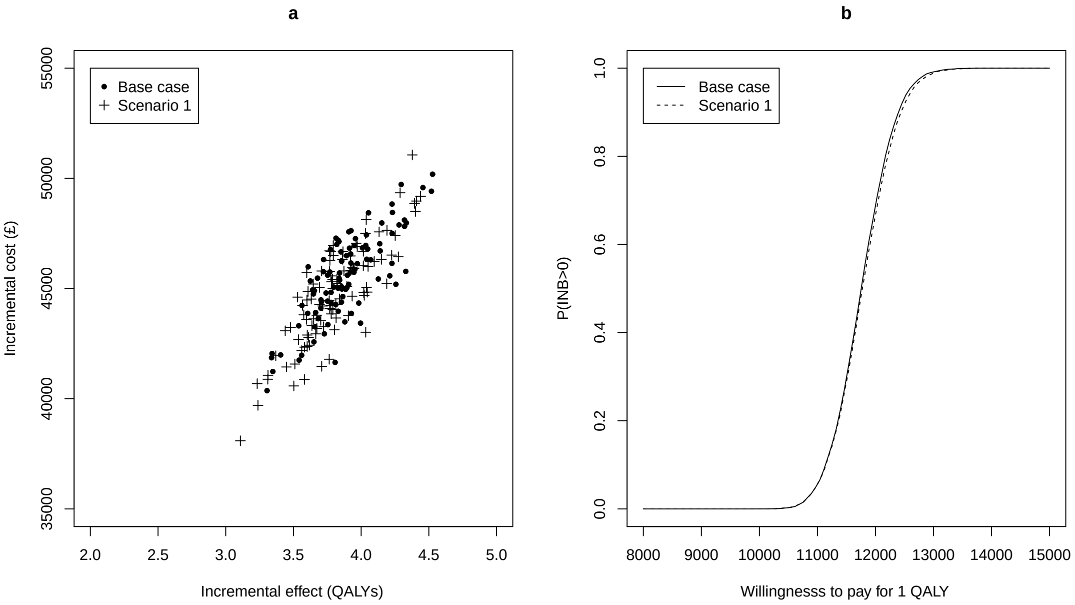


Figure 6.7: (a) cost-effectiveness plane (b) CEAC for scenario 1

6.4.3 Scenario 2

After the addition of discrepancy terms to reflect the judgements about model error due to the constant treatment efficacy assumption, the mean incremental cost of combination therapy over monotherapy was £39,741 and the mean incremental benefit was 3.20 life years, representing a cost per life year gained of £12,409.

Figure 6.8 shows the cost-effectiveness plane and cost-effectiveness acceptability curve (CEAC) for scenario 2, overlaid on those for the base case. Note the shift in the cloud of points on the CE plane towards the origin reflecting the reduced efficacy of the drug and consequent reduction in both life years gained, and care costs accrued. With smaller benefits *and* costs the combination therapy intervention will only now be cost effective at higher values of the willingness to pay, hence the shift of the CEAC curve to the right.

Value of information analysis (table 6.6) suggests that although there is still some value in learning the treatment effect and cost parameters, it is the discrepancy terms that are now most important in driving decision uncertainty at $\lambda = £12,000$. In this scenario there is value in improving the model such that it better reflects our judgements about the decision problem, as well as value in reducing parameter uncertainty.

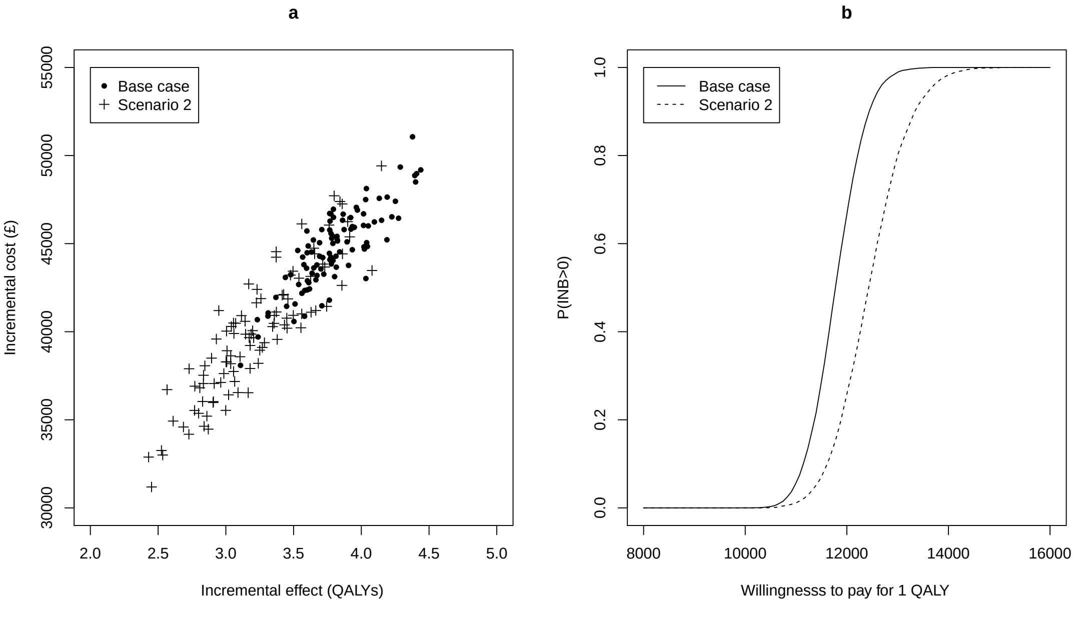


Figure 6.8: (a) cost-effectiveness plane (b) CEAC for scenario 2

6.4.4 Scenario 3

After the addition of discrepancy terms to reflect the judgements about model error due to the Markovian assumption of memorylessness, the mean incremental cost of combination therapy over monotherapy was £45,111 and the mean incremental benefit was 3.84 life years, representing a cost per life year gained of £11,744. Figure 6.9 shows the cost-effectiveness (CE) plane and cost-effectiveness acceptability curve (CEAC) for scenario 3, overlaid on those for the base case. Note the somewhat larger cloud of points on the CE plane and flatter CEAC reflecting the additional uncertainty.

Value of information analysis (table 6.6) suggests that the decision is again sensitive to the discrepancy terms, and that building a more complex model to better represent non-Markovian transitions between health states may be worthwhile.

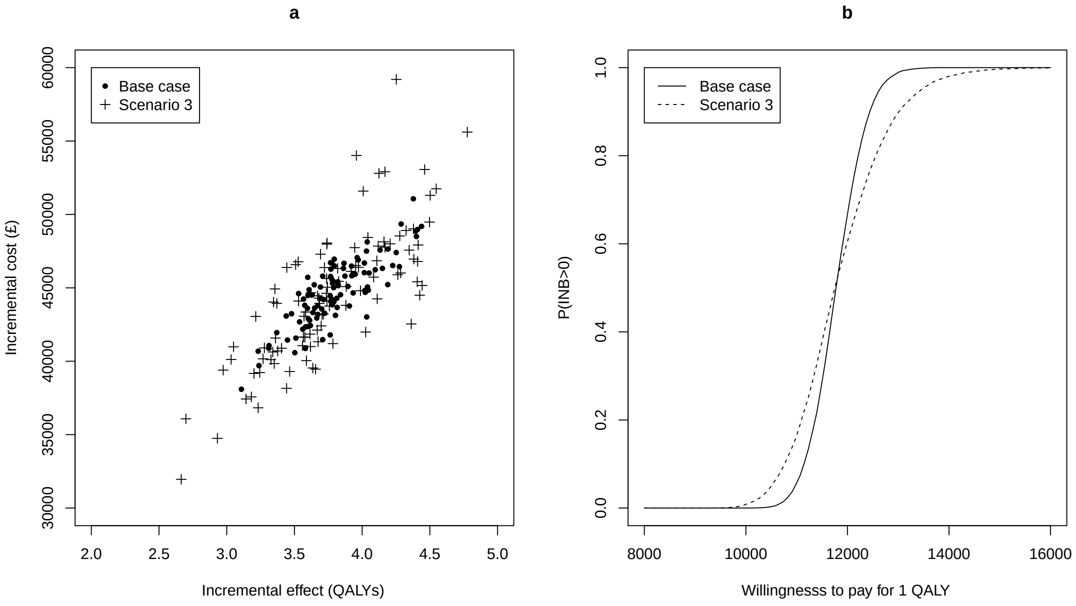


Figure 6.9: (a) cost-effectiveness plane (b) CEAC for scenario 3

6.5 Could we have added fewer discrepancy terms?

Adding discrepancies to the transition probabilities for each time step resulted in a large number of terms. In order to manage this large number of uncertain terms we specified a stochastic model for their joint distribution, the Gaussian process. An alternative would have been to consider adding the discrepancy further 'downstream' in the model. So, instead of adding discrepancies to the transition matrix, giving

$$NB_d = \lambda e_d^{tot} - c_d^{tot} = \lambda \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}(\boldsymbol{M}_d + \boldsymbol{\Delta}_{dt})^t \boldsymbol{e}_{dt} - \sum_{t=0}^{20} \boldsymbol{\pi}'_{d0}(\boldsymbol{M}_d + \boldsymbol{\Delta}_{dt})^t \boldsymbol{c}_{dt} \quad (6.17)$$

we might add discrepancies to total costs and effects at time t.

$$NB_d = \lambda e_d^{tot} - c_d^{tot} = \lambda \sum_{t=0}^{20} \left(\boldsymbol{\pi}'_{d0} \boldsymbol{M}_d^t \boldsymbol{e}_{dt} + \delta_t^e \right) - \sum_{t=0}^{20} \left(\boldsymbol{\pi}'_{d0} \boldsymbol{M}_d^t \boldsymbol{c}_{dt} + \delta_t^c \right). \quad (6.18)$$

There is a trade off here. There may be fewer discrepancies to specify downstream, but the discrepancies may be more difficult to make judgements about. Eventually we reach the model output, and here we are back in the position we discussed in section §3.4.3. At this level it might be very difficult to say anything particularly meaningful about the discrepancy. By adding discrepancy 'upstream' we can potentially utilise our detailed knowledge about how specific parts of the model might differ from our 'best conceptual' model of the problem.

Where to add discrepancies is a judgement in itself. We could start at the model output and consider whether we are able to make useful judgements at this point. If not, we would then work upstream towards the model inputs until we came to a level in the model at which we can make useful judgements about discrepancy. This is one possible approach. Finding an the best solution to this problem requires future research.

6.6 Conclusion

In this chapter we applied the discrepancy method to a Markov model. This necessitated making judgements about a large number of discrepancy terms. We imposed structure on the set of uncertain discrepancy terms by respecifying them as a Gaussian process. We determined the value of learning the uncertain discrepancy terms by calculating the partial expected value of perfect information. We interpreted this as an upper bound on the expected value of improving the model (EVMI).

As we noted at the end of chapter 5, specifying a meaningful distribution on the discrepancy terms is likely to be hard. We return to this point in chapter 8. In the next chapter we describe a novel method for calculating conditional expectations for the purposes of sensitivity analysis when inputs are correlated.

Chapter 7

Efficient Computation of the Main Effect Index and Partial Expected Value of Perfect Information when Inputs are Correlated

7.1 Introduction

In this chapter[1] we describe a novel method for efficiently computing the main effect index and the partial expected value of perfect information when inputs are correlated. In the context of the discrepancy analyses in chapters 5 and 6, each discrepancy term would be considered as just another 'input' to the model for the purposes of calculating these two sensitivity measures. In both our case studies discrepancy terms were correlated.

The standard two level Monte Carlo approach to calculating both the main effect index and the partial EVPI is to sample a value of the input parameter of interest in an outer loop, and then to sample values from the joint conditional

[1] A paper based on the content of this chapter was submitted to Medical Decision Making in October 2011.

distribution of the remaining parameters and run the model in an inner loop (Brennan et al., 2007; Koerkamp et al., 2006). Sufficient numbers of runs of both the outer and inner loops are required to insure that these quantities are estimated with sufficient precision. For the computation of partial EVPI there is the added complication that without sufficient numbers of inner loop samples the EVPI will be estimated with an unacceptable level of upward bias due to the maximisation step (Oakley et al., 2010).

We recognise two important practical limitations to the standard two level Monte Carlo approach to calculating the main effect index and partial EVPI. Firstly, the nested two level nature of the algorithm with a model run at each inner loop step can be highly computationally demanding for all but very small loop sizes if the model is expensive to run. Secondly, we require a method of sampling from the joint distribution of the inputs (excluding the parameter of interest) conditional on the input parameter of interest. If the input parameter of interest is independent of the remaining parameters then we can simply sample from the unconditional joint distribution of the remaining parameters. Indeed, Ades et al. (2004) show that in certain classes of model, most notably decision tree models with independent inputs, the Monte Carlo inner loop is unnecessary since the target inner expectation has a closed form solution. However, if inputs are not independent we may need to resort to Markov chain Monte Carlo (MCMC) methods if there is no closed form analytic solution to the joint conditional distribution. Including an MCMC step in the algorithm is likely to increase the computational burden considerably, as well as requiring additional programming.

In this chapter we present a simple one level 'ordered input' algorithm for calculating the main effect index and the partial EVPI that takes into account any dependency in the inputs. The method avoids the need to sample directly from the conditional distributions of the inputs, and instead requires only a single set of the sampled inputs and corresponding outputs in order to calculate the main effect index and partial EVPI values for all input parameters.

We introduce the method in section §7.2 and present a theoretical justification in section §7.4. We derive the sampling distribution of the estimator and discuss sample size choices in section §7.5, followed by a case study in section §7.6 where we use the method to calculate EVPI. In section §7.7 we discuss some strengths and limitations of the approach.

7.2 Method for partial EVPI

We assume we are faced with D decision options, indexed $d = 1, \ldots, D$, and have built a computer model $y_d = f(d, \boldsymbol{x})$ that aims to predict the net benefit of decision option d given a vector of input parameter values $\boldsymbol{x}$. We denote the true unknown values of the inputs $\boldsymbol{X} = \{X_1, \ldots, X_p\}$, and the uncertain net benefit under decision option d as Y_d. We denote the parameter for which we wish to calculate the partial EVPI as X_i and the remaining parameters as $\boldsymbol{X}_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p\}$. We denote the expectation over the full joint distribution of $\boldsymbol{X}$ as $E_{\boldsymbol{X}}$, over the marginal distribution of X_i as E_{X_i}, and over the conditional distribution of $\boldsymbol{X}_{-i}|X_i$ as $E_{\boldsymbol{X}_{-i}|X_i}$. The partial EVPI for input X_i is

$$EVPI(X_i) = E_{X_i}\left[\max_d E_{\boldsymbol{X}_{-i}|X_i}\left\{f(d, X_i, \boldsymbol{X}_{-i})\right\}\right] - \max_d E_{\boldsymbol{X}}\{f(d, \boldsymbol{X})\}. \tag{7.1}$$

We wish to evaluate the partial EVPI for each input X_i without sampling directly from the conditional distribution $\boldsymbol{X}_{-i}|X_i$, since this may require computationally intensive numerical methods if inputs are correlated.

Our method for avoiding this difficulty rests on recognising the following. Given a Monte Carlo sample of S input parameter vectors drawn from the joint distribution $p(\boldsymbol{X})$, we can order the set of sample vectors with respect to the parameter of interest X_i, i.e.,

$$\begin{pmatrix} x_1^{(1)} & \ldots & x_i^{(1)} & \ldots & x_p^{(1)} \\ x_1^{(2)} & \ldots & x_i^{(2)} & \ldots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{(S)} & \ldots & x_i^{(S)} & \ldots & x_p^{(S)} \end{pmatrix}, \tag{7.2}$$

where $x_i^{(1)} \leq x_i^{(2)} \leq \ldots \leq x_i^{(S)}$. Then, for some small integer δ and index k where $\delta < k \leq S - \delta$ the vectors $\boldsymbol{x}_{-i}^{(k-\delta)}, \ldots, \boldsymbol{x}_{-i}^{(k)}, \ldots, \boldsymbol{x}_{-i}^{(k+\delta)}$ are approximate samples from the conditional distribution $\boldsymbol{X}_{-i}|X_i = x_i^{(k)}$ if S is large compared to δ. We can approximate the problematic expectation in equation (7.1) by

$$E_{\boldsymbol{X}_{-i}|X_i}\left\{f(d, X_i, \boldsymbol{X}_{-i})\right\} \simeq \frac{1}{\delta+1}\sum_{j=k-\delta}^{k+\delta} f\left(d, \boldsymbol{x}^{(j)}\right). \tag{7.3}$$

The second term in the RHS of equation (7.1) can be estimated simply via Monte Carlo sampling, i.e.

$$\max_d E_{\boldsymbol{X}}\{f(d, \boldsymbol{X})\} \simeq \max_d \frac{1}{N}\sum_{n=1}^{N} f(d, \boldsymbol{X}). \tag{7.4}$$

7.2.1 Algorithm for calculating partial EVPI via the one stage ‘ordered input’ method

We propose the following algorithm for computing the first term in the RHS of equation (7.1). Code for implementing the algorithm in R (R Development Core Team, 2011) is shown in section §7.2.2.

Stage 1

We first obtain a single Monte Carlo sample $M = \{(\boldsymbol{x}^s, y_1^s, \ldots, y_D^s), s = 1, \ldots, S\}$ where $\boldsymbol{x}^s$ are drawn from the joint distribution of the inputs, $p(\boldsymbol{X})$, and $y_d^s = f(d, \boldsymbol{x}^s)$ is the evaluation of the model output at $\boldsymbol{x}^s$ for decision option $d = 1, \ldots, D$. Note the use of superscripts to index the randomly drawn sample sets. We let M be the matrix of inputs and corresponding outputs

$$M = \begin{pmatrix} x_1^1 & \ldots & x_p^1 & y_1^1 & \ldots & y_D^1 \\ x_1^2 & \ldots & x_p^2 & y_1^2 & \ldots & y_D^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^S & \ldots & x_p^S & y_1^S & \ldots & y_D^S \end{pmatrix}. \tag{7.5}$$

Stage 2

For parameter of interest i, we extract the x_i and $y_1, \ldots, y_D$ columns and reorder with respect to x_i, giving

$$M^* = \begin{pmatrix} x_i^{(1)} & y_1^{(1)} & \ldots & y_D^{(1)} \\ x_i^{(2)} & y_1^{(2)} & \ldots & y_D^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_i^{(S)} & y_1^{(S)} & \ldots & y_D^{(S)} \end{pmatrix}, \tag{7.6}$$

where $x_i^{(1)} \leq x_i^{(2)} \leq \ldots \leq x_i^{(S)}$.

Stage 3

We partition the resulting matrix into $k = 1, \ldots, K$ sub matrices $M^{*(k)}$ of J rows each,

$$M^{*(k)} = \begin{pmatrix} x_i^{(1,k)} & y_1^{(1,k)} & \ldots & y_D^{(1,k)} \\ x_i^{(2,k)} & y_1^{(2,k)} & \ldots & y_D^{(2,k)} \\ \vdots & \vdots & \vdots & \vdots \\ x_i^{(J,k)} & y_1^{(J,k)} & \ldots & y_D^{(J,k)} \end{pmatrix}, \tag{7.7}$$

retaining the ordering with respect to x_i, and where the row indexed (j, k) in equation (7.7) is the row indexed $(j + (k-1)J)$ in equation (7.6). Note that $J \times K$ must equal the total sample size S.

Stage 4

For each $M^{*(k)}$ we estimate the conditional expectation $\mu_d^{(k)} = E_{\boldsymbol{X}_{-i}|X_i = x_i^{*(k)}} \{f(d, X_i, \boldsymbol{X}_{-i})\}$ for each decision option by

$$\hat{\mu}_d^{(k)} = \frac{1}{J} \sum_{j=1}^{J} y_d^{(j,k)}, \tag{7.8}$$

where $x_i^{*(k)} = \sum_{j=1}^{J} x_i^{(j,k)}/J$. The maximum $m^{(k)} = \max_d E_{\boldsymbol{X}_{-i}|X_i=x_i^{*(k)}} \{f(d, X_i, \boldsymbol{X}_{-i})\}$ is estimated by

$$\hat{m}^{(k)} = \max_d \hat{\mu}_d^{(k)}. \tag{7.9}$$

Finally, we estimate the first term in the RHS of equation (7.1) by

$$\bar{\hat{m}} = \frac{1}{K}\sum_{k=1}^{K} \hat{m}^{(k)}. \tag{7.10}$$

Stages 2 to 4 are repeated for each parameter of interest. Note that only a single set of model runs (stage 1) is required.

7.2.2 R code for implementing the partial EVPI algorithm

The `partial.evpi.function` function as written below takes as inputs the costs and effects rather than the net benefits. This allows the partial EVPI to be calculated at any value of willingness to pay, λ.

```
partial.evpi.function<-function(inputs,input.of.interest,costs,effects,lambda,J,K)
{
  S <- nrow(inputs) # number of samples
  if(J*K!=S) stop("The number of samples does not equal J times K")
  D <- ncol(costs) # number of decision options

  nb <- lambda*effects-costs
  baseline <- max(colMeans(nb))
  perfect.info <- mean(apply(nb,1,max))
  evpi <- perfect.info-baseline

  sort.order <- order(inputs[,input.of.interest])
  sort.nb <- nb[sort.order,]

  nb.array <- array(sort.nb,dim=c(J,K,D))
  mean.k <- apply(nb.array,c(2,3),mean)
  partial.info <- mean(apply(mean.k,1,max))
  partial.evpi <- partial.info-baseline
  partial.evpi.index <- partial.evpi/evpi

  return(list(
    baseline = baseline,
    perfect.info = perfect.info,
    evpi = evpi,
    partial.info = partial.info,
    partial.evpi = partial.evpi,
    partial.evpi.index = partial.evpi.index
  ))
}
```

7.3 Method for main effect index

We assume we have built a computer model $y = f(\boldsymbol{x})$ with a scalar output, and a vector of input parameter values $\boldsymbol{x}$. We denote the true unknown values of the inputs $\boldsymbol{X} = \{X_1, \ldots, X_p\}$, and the uncertain output Y. We denote the parameter of interest as X_i and the remaining parameters as $\boldsymbol{X}_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p\}$. We denote the expectation over the full joint distribution of $\boldsymbol{X}$ as $E_{\boldsymbol{X}}$, over the marginal distribution of X_i as E_{X_i}, and over the conditional distribution of $\boldsymbol{X}_{-i}|X_i$ as $E_{\boldsymbol{X}_{-i}|X_i}$. The main effect index for X_i is

$$\frac{\text{var}_{X_i}\left[E_{\boldsymbol{X}_{-i}|X_i}\left\{f(X_i, \boldsymbol{X}_{-i})\right\}\right]}{\text{var}(Y)}. \tag{7.11}$$

7.3.1 Algorithm for calculating the main effect index via the one stage ‘ordered input’ method

We propose the following algorithm for computing the numerator of (7.11). Code for implementing the algorithm in R is shown in section §7.3.2.

Stage 1

We first obtain a single Monte Carlo sample $M = \{(\boldsymbol{x}^s, y^s), s = 1, \ldots, S\}$ where $\boldsymbol{x}^s$ are drawn from the joint distribution of the inputs, $p(\boldsymbol{X})$, and $y^s = f(\boldsymbol{x}^s)$ is the evaluation of the model output at $\boldsymbol{x}^s$. Note the use of superscripts to index the randomly drawn sample sets. We let M be the matrix of inputs and corresponding outputs

$$M = \begin{pmatrix} x_1^1 & \dots & x_p^1 & y^1 \\ x_1^2 & \dots & x_p^2 & y^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^S & \dots & x_p^S & y^S \end{pmatrix}. \tag{7.12}$$

Stage 2

For parameter of interest i, we extract the x_i and y columns and reorder with respect to x_i, giving

$$M^* = \begin{pmatrix} x_i^{(1)} & y^{(1)} \\ x_i^{(2)} & y^{(2)} \\ \vdots & \vdots \\ x_i^{(S)} & y^{(S)} \end{pmatrix}, \tag{7.13}$$

where $x_i^{(1)} \leq x_i^{(2)} \leq \ldots \leq x_i^{(S)}$.

Stage 3

We partition the resulting matrix into $k = 1, \ldots, K$ sub matrices $M^{*(k)}$ of J rows each,

$$M^{*(k)} = \begin{pmatrix} x_i^{(1,k)} & y^{(1,k)} \\ x_i^{(2,k)} & y^{(2,k)} \\ \vdots & \vdots \\ x_i^{(J,k)} & y^{(J,k)} \end{pmatrix}, \tag{7.14}$$

retaining the ordering with respect to x_i.

Stage 4

For each $M^{*(k)}$ we estimate the conditional expectation $\mu^{(k)} = E_{\boldsymbol{X}_{-i}|X_i=x_i^{*(k)}}\left\{f(X_i, \boldsymbol{X}_{-i})\right\}$ for each decision option by

$$\hat{\mu}^{(k)} = \frac{1}{J}\sum_{j=1}^{J} y^{(j,k)}, \tag{7.15}$$

where $x_i^{*(k)} = \sum_{j=1}^{J} x_i^{(j,k)}/J$.

Finally we estimate the variance of the conditional expectation by

$$\hat{\text{var}}(\hat{\mu}^{(k)}) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\mu}^{(k)} - \bar{\hat{\mu}}^{(k)}\right)^2, \tag{7.16}$$

Stages 2 to 4 are repeated for each parameter of interest. Note that only a single set of model runs (stage 1) is required.

7.3.2 R code for implementing the Main Effect Index algorithm

The `main.effect.index.function` function as written below takes as inputs the incremental costs and incremental effects rather than the incremental net benefit. This allows the main effect index to be calculated at any value of willingness to pay, λ.

```
main.effect.index.function<-function(inputs,input.of.interest,costs,effects,lambda,J,K)
{
  S <- nrow(inputs) # number of samples
  if(J*K!=S) stop("The number of samples does not equal J times K")

  inb <- lambda*effects-costs
  var.Y <- var(inb)

  sort.order <- order(inputs[,input.of.interest])
  sort.inb <- inb[sort.order]

  inb.array <- array(sort.inb,dim=c(J,K,1))
  mean.k <- apply(inb.array,c(2,3),mean)
  var.exp <- var(mean.k)

  return(list(
    var.Y = var.Y,
    var.exp = var.exp,
    main.effect.index = var.exp/var.Y
  ))
}
```

7.4 Theoretical justification

The ordered algorithm is a method for efficiently computing the inner expectation in the first term of the RHS in the EVPI equation (7.1), and the numerator in the main effect index expression (7.11). Dropping the decision option index d for clarity but without loss of generality, our target is $E_{\boldsymbol{X}_{-i}|X_i=x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\}$ where x_i^* is a realised value of the parameter of interest, and $\boldsymbol{X}_{-i}$ are the remaining (uncertain) parameters with joint conditional distribution $p(\boldsymbol{X}_{-i}|X_i = x_i^*)$.

Given a sample $\left\{\boldsymbol{x}_{-i}^{(1)}, \ldots, \boldsymbol{x}_{-i}^{(J)}\right\}$ from $p(\boldsymbol{X}_{-i}|X_i = x_i^*)$, the Monte Carlo estimator for $E_{\boldsymbol{X}_{-i}|X_i=x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\}$ is

$$\hat{E}_{\boldsymbol{X}_{-i}|X_i=x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\} = \frac{1}{J}\sum_{j=1}^{J} f\left(x_i^*, \boldsymbol{x}_{-i}^{(j)}\right). \tag{7.17}$$

In our ordered approximation method we replace (7.17) with

$$\hat{E}_{\boldsymbol{X}_{-i}|X_i=x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\} = \frac{1}{J}\sum_{j=1}^{J} f\left(x_i^* + \varepsilon_j, \tilde{\boldsymbol{x}}_{-i}^{(j)}\right), \tag{7.18}$$

where $\{x_i^* + \varepsilon_1, \ldots, x_i^* + \varepsilon_J\} = \{x_i^{(1)}, \ldots, x_i^{(J)}\}$ is an ordered sample from $p(X_i|X_i \in [x_i^* \pm \zeta])$ for some small ζ (and therefore $\bar{\varepsilon} \simeq 0$), and $\tilde{\boldsymbol{x}}_{-i}^{(j)}$ is a sample from $p(\boldsymbol{X}_{-i}|X_i = x_i^* + \varepsilon_j)$.

The expression (7.18) is an unbiased Monte Carlo estimator of

$$\begin{aligned}&E_{X_i\in[x_i^*\pm\zeta]}\left\{E_{\boldsymbol{X}_{-i}|X_i}f(X_i,\boldsymbol{X}_{-i})\right\}\\ &=\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}f(X_i,\boldsymbol{X}_{-i})p(\boldsymbol{X}_{-i}|X_i)p(X_i|X_i\in[x_i^*\pm\zeta])dX_i d\boldsymbol{X}_{-i},\end{aligned}\tag{7.19}$$

which we can rewrite by introducing an importance sampling ratio as

$$\begin{aligned}&\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}f(X_i,\boldsymbol{X}_{-i})p(\boldsymbol{X}_{-i}|X_i)p(X_i|X_i\in[x_i^*\pm\zeta])dX_i d\boldsymbol{X}_{-i}\\ &=\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)p(X_i|X_i\in[x_i^*\pm\zeta])}{p(\boldsymbol{X}_{-i}|X_i)p(X_i|X_i=x_i^*)}p(\boldsymbol{X}_{-i}|X_i)p(X_i|X_i=x_i^*)dX_i d\boldsymbol{X}_{-i}\\ &=\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i=x_i^*)}p(X_i|X_i\in[x_i^*\pm\zeta])dX_i\; p(\boldsymbol{X}_{-i}|X_i=x_i^*)\;d\boldsymbol{X}_{-i}.\end{aligned}\tag{7.20}$$

We write the terms $f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i=x_i^*)}$ within the inner integral as a function $g(\cdot)$, i.e.

$$f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i=x_i^*)}=g(X_i,x_i^*,\boldsymbol{X}_{-i}).$$

If $g(\cdot)$ is approximately linear in the small interval $X_i\in[x_i^*\pm\zeta]$ then we can express $g(X_i,x_i^*,\boldsymbol{X}_{-i})$ as a first order Taylor series expansion about $g(x_i^*,x_i^*,\boldsymbol{X}_{-i})$, giving

$$\begin{aligned}f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i=x_i^*)} &= g(X_i,x_i^*,\boldsymbol{X}_{-i}),\\ &\simeq g\left(x_i^*,x_i^*,\boldsymbol{X}_{-i}\right)+(X_i-x_i^*)\frac{\partial g\left(X_i,x_i^*,\boldsymbol{X}_{-i}\right)}{\partial X_i}\Big|_{X_i=x_i^*}\\ &= f(x_i^*,\boldsymbol{X}_{-i})+(X_i-x_i^*)\frac{\partial g\left(X_i,x_i^*,\boldsymbol{X}_{-i}\right)}{\partial X_i}\Big|_{X_i=x_i^*}.\end{aligned}$$

Substituting back into (7.20) with $c=\frac{\partial g\left(X_i,x_i^*,\boldsymbol{X}_{-i}\right)}{\partial X_i}\Big|_{X_i=x_i^*}$ gives

$$\begin{aligned}&\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}f(X_i,\boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i=x_i^*)}p(X_i|X_i\in[x_i^*\pm\zeta])dX_i\; p(\boldsymbol{X}_{-i}|X_i=x_i^*)\;d\boldsymbol{X}_{-i}\\ &\simeq\int_{\mathcal{X}_{-i}}\int_{\mathcal{X}_i}\left\{f(x_i^*,\boldsymbol{X}_{-i})+c(X_i-x_i^*)\right\}p(X_i|X_i\in[x_i^*\pm\zeta])\;dX_i\;p(\boldsymbol{X}_{-i}|X_i=x_i^*)\;d\boldsymbol{X}_{-i}.\end{aligned}$$

Since $\int_{\mathcal{X}_i} c(X_i - x_i^*)p(X_i|X_i \in [x_i^* \pm \zeta])dX_i = E_{X_i \in [x_i^* \pm \zeta]}\{c(X_i - x_i^*)\} \simeq 0$ and $\int_{\mathcal{X}_i} p(X_i|X_i \in [x_i^* \pm \zeta])\, dX_i = 1$, then

$$\begin{aligned} &\int_{\mathcal{X}_{-i}} \int_{\mathcal{X}_i} \{f(x_i^*, \boldsymbol{X}_{-i}) + c(X_i - x_i^*)\}\, p(X_i|X_i \in [x_i^* \pm \zeta])\, dX_i\, p(\boldsymbol{X}_{-i}|X_i = x_i^*)\, d\boldsymbol{X}_{-i}, \\ &\simeq \int_{\mathcal{X}_{-i}} f(x_i^*, \boldsymbol{X}_{-i}) p(\boldsymbol{X}_{-i}|X_i = x_i^*)\, d\boldsymbol{X}_{-i}, \\ &= E_{\boldsymbol{X}_{-i}|X_i = x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\}. \end{aligned}$$

Hence, we have shown that as long as $g(X_i, x_i^*, \boldsymbol{X}_{-i}) = f(X_i, \boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i = x_i^*)}$ is sufficiently smooth such that it is approximately linear in some small interval $X_i \in [x_i^* \pm \zeta]$, the ordered approximation method (7.18) will provide a good estimate of our target conditional expectation $E_{\boldsymbol{X}_{-i}|X_i = x_i^*}\{f(x_i^*, \boldsymbol{X}_{-i})\}$. For $f(X_i, \boldsymbol{X}_{-i})\frac{p(\boldsymbol{X}_{-i}|X_i)}{p(\boldsymbol{X}_{-i}|X_i = x_i^*)}$ to be smooth both the model function $f(X_i, \boldsymbol{X}_{-i})$ and the conditional probability density function $p(\boldsymbol{X}_{-i}|X_i)$ must be smooth with respect to X_i in the interval $[x_i^* \pm \zeta]$. Economic models tend to be smooth functions of their inputs, and it is also likely that in most health economic modelling scenarios that the conditional density $p(\boldsymbol{X}_{-i}|X_i)$ will be smooth with respect to X_i.

7.5 Sample size considerations for the estimation of partial EVPI

In this section we derive the sampling distribution of the estimator for the partial EVPI, and suggest a method for choosing optimum values of J and K.

7.5.1 Estimating the precision of the partial EVPI estimator

For the purposes of this section we assume that we can estimate the second term in the RHS of equation (7.1) with sufficient accuracy by choosing large N in equation (7.4), and therefore that this second term does not contribute significantly to the variance of the estimate of the partial EVPI.

If we denote $d_k^* = \arg\max_d \left(\hat{\mu}_d^{(k)}\right)$ we can rewrite equation (7.10) as

$$\begin{aligned}
\hat{E}_{X_i}(\hat{m}^{(k)}) = \bar{\hat{m}} &= \frac{1}{K}\sum_{k=1}^{K} \hat{m}^{(k)}, \\
&= \frac{1}{K}\sum_{k=1}^{K} \hat{\mu}_{d_k^*}^{(k)}, \\
&= \frac{1}{K}\sum_{k=1}^{K} \left(\frac{1}{J}\sum_{j=1}^{J} y_{d_k^*}^{(j,k)} \right), \\
&= \frac{1}{S}\sum_{k=1}^{K}\sum_{j=1}^{J} y_{d_k^*}^{(j,k)}. \qquad (7.21)
\end{aligned}$$

The variance of $\bar{\hat{m}}$ is

$$\begin{aligned}
\text{var}(\bar{\hat{m}}) &= \text{var}\left(\frac{1}{S}\sum_{k=1}^{K}\sum_{j=1}^{J} y_{d_k^*}^{(j,k)} \right), \\
&= \frac{1}{S^2}\sum_{k=1}^{K}\sum_{j=1}^{J} \text{var}\left(y_{d_k^*}^{(j,k)} \right), \qquad (7.22)
\end{aligned}$$

since the $y_{d_k^*}^{(j,k)}$ are independent. The estimator for $\text{var}(\bar{\hat{m}})$ is therefore simply

$$\widehat{\text{var}}(\bar{\hat{m}}) = \frac{1}{S(S-1)}\sum_{k=1}^{K}\sum_{j=1}^{J} \left(y_{d_k^*}^{(j,k)} - \bar{\hat{m}} \right)^2. \qquad (7.23)$$

We see therefore that the precision of the estimator does not depend on the individual choices of J and K, but only on $S = J \times K$.

7.5.2 Choosing values for J and K

We assume that we have a fixed number of model evaluations S and wish to choose values for J and K subject to the constraint $J \times K = S$.

Firstly we note that for small values of J the EVPI estimator is upwardly biased due to the maximisation in equation (7.9) (Oakley et al., 2010). Indeed for $J = 1$ (and $K = S$) our ordered input estimator for the first term in the RHS of equation (7.1) reduces to

$$\frac{1}{S}\sum_{s=1}^{S}\max_{d}(y_d^s), \tag{7.24}$$

which is the Monte Carlo estimator for the first term in the expression for the *overall* EVPI, $E_{\boldsymbol{X}}\left\{\max_d f(d,\boldsymbol{X})\right\} - \max_d E_{\boldsymbol{X}}\{f(d,\boldsymbol{X})\}$.

Secondly we note that for very large values of J, and hence small values of K, the EVPI estimator is downwardly biased, and converges to zero when $J = S$. In this case our ordered input estimator for the first term in the RHS of equation (7.1) reduces to

$$\max_{d}\frac{1}{S}\sum_{s=1}^{S}y_d^s, \tag{7.25}$$

which is the Monte Carlo estimator for the second term in the RHS of equation (7.1).

Given that the algorithm is computationally inexpensive we can find appropriate values for J and K empirically by running the algorithm at a range of values of J and K, subject to $J \times K = S$ (in practice we only need choose $J \times K \leq S$). Figure 7.1 shows values for the estimated partial EVPI against J (on the $\log_{10}$ scale) for input X_6 in scenario 1 of the case study that we introduce in section §7.6. The total number of model evaluations, S, is 1,000,000. Note the upward and downward biases at extreme values of J, but also the large region of stability between $J = 100$ ($K = 10,000$) and $J = 100,000$ ($K = 10$).

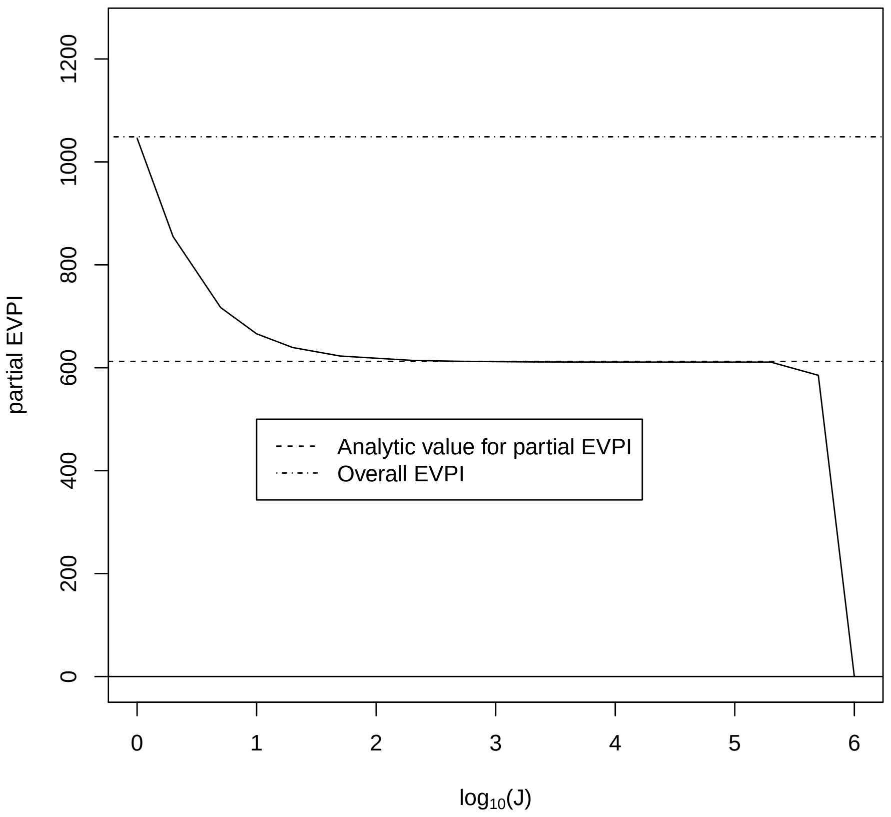


Figure 7.1: Partial EVPI at values of J ranging from 1 to 10^6 where the total number of model evaluations, S, is 10^6.

7.6 Case study

In this case study we compare the ordered input method with the standard two level Monte Carlo method for calculating partial EVPI. The model for the case study is a hypothetical decision tree model previously used for illustrative purposes in Brennan et al. (2007), Oakley et al. (2010) and Kharroubi et al. (2011). The model predicts monetary net benefit, Y_d, under two decision options ($d = 1, 2$) and can be written in sum product form as

$$\begin{aligned} Y_1 &= \lambda(X_5X_6X_7 + X_8X_9X_{10}) - (X_1 + X_2X_3X_4), &(7.26)\\ Y_2 &= \lambda(X_{14}X_{15}X_{16} + X_{17}X_{18}X_{19}) - (X_{11} + X_{12}X_{13}X_4), &(7.27)\end{aligned}$$

where $\boldsymbol{X} = \{X_1, \ldots, X_{19}\}$ are the 19 uncertain input parameters listed in table 7.1, and the willingness to pay for one unit of health output in QALYs is $\lambda = \pounds 10,000$/QALY. We implemented the model in R (R Development Core Team, 2011).

Parameter	Mean (sd)	
	$d = 1$	$d = 2$
Cost of Drug (X_1, X_{11})	£1000 (£1)	£1500 (£1)
% Admissions (X_2, X_{12})	10% (2%)	8% (2%)
Days in Hospital (X_3, X_{13})	5.20 (1.00)	6.10 (1.00)
Cost per day (X_4)	£400 (£200)	£400 (£200)
% Responding (X_5, X_{14})	70% (10%)	80% (10%)
Utility Change if respond (X_6, X_{15})	0.30 (0.10)	0.30 (0.05)
Duration of response (years) (X_7, X_{16})	3.0 (0.5)	3.0 (1.0)
% Side effects (X_8, X_{17})	25% (10%)	20 (5%)
Change in utility if side effect (X_9, X_{18})	-0.10 (0.02)	-0.10 (0.02)
Duration of side effect (years) (X_{10}, X_{19})	0.50 (0.20)	0.50 (0.20)

Table 7.1: Summary of input parameters

7.6.1 Scenario 1: correlated inputs with known conditional distributions

In scenario 1 we assume that a subset of the inputs are correlated, but with a joint distribution such that we can sample from the conditional distributions of the correlated inputs without the need for MCMC. We assume that the inputs are jointly normally distributed, with X_5, X_7, X_{14} and X_{16} all pairwise correlated with a correlation coefficient of 0.6, and with all other inputs independent. In a simple sum product form model the assumption of multivariate normality allows us to compute the inner conditional expectation analytically, as well as allowing us to sample directly from the conditional distribution $\boldsymbol{X}_{-i}|X_i$ in the standard nested two level method, but this will not necessarily be the case in models with additional non-linearity.

We calculated partial EVPI using three methods. Firstly, we calculated the partial EVPI for each parameter using a single loop Monte Carlo approximation for the outer expectation in the first term of the RHS of equation (7.1) with 10^6

samples from the distribution of the parameter of interest, and an analytic solution to the inner conditional expectation. Next, we calculated the partial EVPI values using the standard two level Monte Carlo approach with 1,000 inner loop samples and 1,000 outer loop samples (i.e 10^6 model evaluations in total). Finally, we computed the partial EVPI values using the ordered sample method with a single set of 10^6 samples and a value of $J = 1,000$.

Standard errors for the two level method estimates were obtained using the method presented in Oakley et al. (2010), and for the ordered input method estimates via equation (7.23). We measured the total computation time for obtaining EVPI values for all 19 parameters. We performed the computations on a single processor core on a 2.93GHz Intel Core i7 machine running 64 bit Linux.

Results for scenario 1

Calculating the expected net benefits for decision options 1 and 2 analytically results in values of £5057.00 and £5584.80 respectively, indicating that decision option 2 is optimal. Running the model with 10^6 Monte Carlo samples from the joint distribution of the input parameters results in option 2 having greater net benefit than option 1 in only 54% of samples, suggesting that the input uncertainty is resulting in considerable decision uncertainty. This is confirmed by the relatively large overall EVPI value of £1046.10.

The partial EVPI values for parameters X_1 to X_4, X_8 to X_{13} and X_{17} to X_{19} were all less than $\pounds 0.01$ and therefore considered unimportant in terms of driving the decision uncertainty. Results for the remaining parameters are shown in table 7.2. The standard errors of the EVPI values estimated via the ordered input method are considerably smaller than those estimated via the two level method, and computation time is reduced by a factor of five.

Parameter	Partial EVPI (SE), £		
	Analytic conditional expectation	Two level Monte Carlo	Ordered input method
X_5	22.50	9.52 (65.20)	25.29 (3.26)
X_6	612.38	614.76 (33.16)	612.63 (3.15)
X_7	11.56	77.65 (66.38)	14.86 (3.28)
X_{14}	230.94	312.39 (69.59)	233.63 (3.19)
X_{15}	271.52	315.02 (29.52)	273.00 (3.30)
X_{16}	458.97	502.91 (77.98)	462.42 (3.12)
Computation time†		57 seconds	12 seconds

† Computation time is for all 19 input parameters

Table 7.2: Partial EVPI values for scenario 1

7.6.2 Scenario 2: correlated inputs with conditional distribution sampling requiring MCMC

In scenario 2 we assume that a subset of the inputs are correlated, but with a joint distribution such that we can only sample from the conditional distributions of the correlated inputs using MCMC. We assume, as in scenario 1, that X_5, X_7, X_{14} and X_{16} are pairwise correlated, but with a more complicated dependency structure based on an unobserved bivariate normal latent variable $\boldsymbol{Z} = (Z_1, Z_2)$ that has expectation zero, variance 1 and correlation 0.6. Conditional on this latent variable, which represents some measure of effectiveness, the proportions of responders (X_5 and X_{14}) are assumed beta distributed, and the durations of response (X_7 and X_{16}) assumed gamma distributed. The hyperparamters of the beta and gamma distributions are defined in terms of $\boldsymbol{Z}$ such that X_5, X_7, X_{14} and X_{16} have the means and standard deviations in table 7.1.

We calculated partial EVPI for each parameter using a the standard two level Monte Carlo approach with 1,000 inner loop samples and 1,000 outer loop samples (i.e 10^6 model evaluations in total) using OpenBUGS (Lunn et al., 2009) to sample from the conditional distribution of $\boldsymbol{X}_{-i}|X_i$.

Parameter	Partial EVPI (SE), £	
	Two level Monte Carlo with MCMC inner loop	Ordered input method
X_5	102.55 (34.48)	34.65 (3.26)
X_6	610.82 (38.02)	618.80 (3.10)
X_7	132.16 (36.10)	56.25 (3.25)
X_{14}	334.13 (51.94)	368.87 (3.18)
X_{15}	223.09 (25.73)	275.78 (3.25)
X_{16}	554.20 (64.00)	663.25 (3.13)
Computation time†	2.7 hours	12 seconds

† Computation time is for all 19 input parameters

Table 7.3: Partial EVPI values for scenario 2

Results for scenario 2

Running the model with 10^6 samples from the joint distribution of the input parameters resulted in expected net benefits of £5043.12 and £5549.93 for decision options 1 and 2 respectively, indicating that decision option 2 is optimal, but again with considerable decision uncertainty. Based on this sample, the probability that decision 2 is best is 54% and the overall EVPI £1240.33.

Partial EVPI results are shown in table 7.3. Values for parameters X_1 to X_4, X_8 to X_{13} and X_{17} to X_{19} were again all less than $\pounds 0.01$ and are not shown. Standards errors for the partial EVPI values estimated via the order input method are again smaller than those estimated via the two level method. The total time required to compute partial EVPI for all 19 inputs was approximately 2.7 hours. In comparison, the ordered input method with a single set of 10^6 samples and a value of $J = 1,000$ took just 12 seconds, an approximately 1,000 fold reduction in computation time.

7.7 Conclusion

We have presented a method for calculating the main effect index and partial expected value of perfect information that is simple to implement, rapid to compute, and does not require an assumption of independence between inputs. In a case study we showed that the saving in computational time is particularly marked if the alternative is to use a nested two level approach in which the conditional expectations are estimated using MCMC. The method is straightforward to apply, even with little programming knowledge in a spreadsheet application.

Our approach requires only a single set of model evaluations in order to calculate the main effect index and/or partial EVPI for all inputs, allowing a complete separation of the sensitivity analysis from the model evaluation. This separation may be particularly useful when the model has been evaluated using specialist software (e.g. for discrete event or agent based simulation) that does not allow easy implementation of the sensitivity analysis, or where those who wish to perform the sensitivity analysis do not 'own' (and therefore cannot directly evaluate) the model.

As presented, the method calculates the main effect index and the partial EVPI for single inputs one at a time. We may however wish to calculate the value of learning groups of inputs simultaneously. There are good reasons for this. Firstly, for certain forms of model we may find that learning single inputs alone has little value, but learning a group of inputs has high value due to the interactions between those inputs within the model. It is important to note that interactions result from non-additive effects within the model, and can occur even if inputs are uncorrelated. Secondly, a certain subset of model inputs may be derived from a single study, and therefore learning one input in this set (by conducting the 'perfect' study) implies learning them all. If we are considering the value of a study in reducing uncertainty about inputs, we will consider the value of *all* the information that arises from the study, not just the information which informs a single input.

The value of our method may then be in 'drilling down' to specific inputs, or small groups of inputs within some larger group of inputs that is judged to be policy relevant. If inputs can be partitioned into broad 'policy relevant' groups (i.e. those which might be considered together when a decision is made to commission further research), and if these groups can be treated as uncorrelated, then calculating the EVPI for each group of inputs using two level Monte Carlo methods is straightforward. At this point, the ordered approximation method could be used to compute the value of single inputs (or small groups of inputs) if this was felt necessary.

Although it is possible to extend our approach to groups of inputs, we quickly come up against the 'curse of dimensionality'. This is because the method relies on partitioning the input space into a large number of 'small' sets such that in each set the parameter of interest lies close to some value. This works well where there is a single parameter of interest, but if we wish to calculate the sensitivity measures for a group of parameters, the samples quickly become much more sparsely located in higher dimensional space. Given a single parameter of interest imagine that we obtain adequate precision if we partition the input space into $K = 1,000$ sets of $J = 1,000$ samples each. With two parameters of interest, we would need to order and partition the space in two dimensions, meaning that to retain the same marginal probabilistic 'size' for each set we now require $K^2 = 1,000,000$ sets of $J = 1,000$ samples each.

Another problem arises when considering more than one input. How do we partition an n-dimensional space into sets of equal probabilistic size (i.e. containing equal numbers of samples) when the inputs are correlated? If inputs are uncorrelated this is straightforward. In two dimensions we can imagine placing a grid over the cloud of points in two dimensions, partitioning the space (figure 7.2). But the same method will not work when inputs are correlated, as in figure 7.3. In this case we must apply some kind of transformation to the inputs; in effect distorting the grid (figure 7.4). One approach is to transform the grid based on the principal components of the correlated variables of interest.

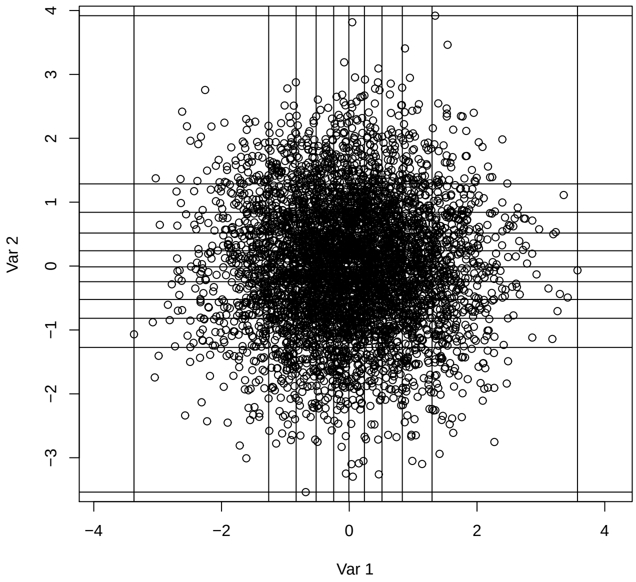


Figure 7.2: Uncorrelated variables with grid based on marginal empirical deciles. This partitions the space into 10 x 10 equal probability sub-spaces with equal numbers of samples in each sub-space.

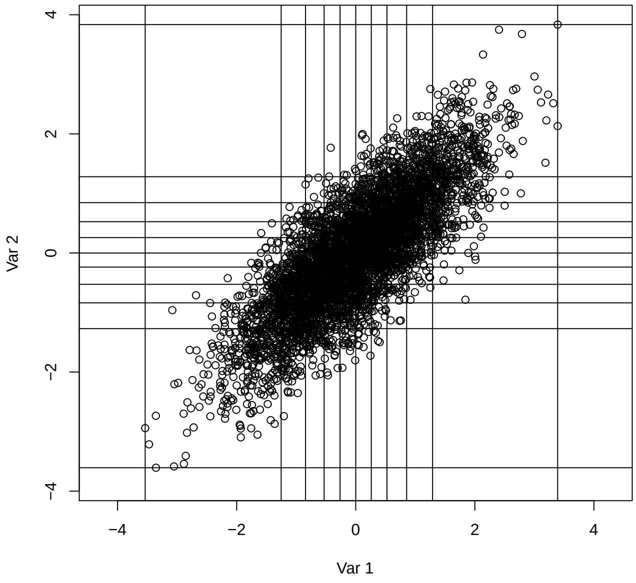


Figure 7.3: Correlated variables with grid based on empirical deciles that are correct for margins. This results in different numbers of samples in each sub-space.

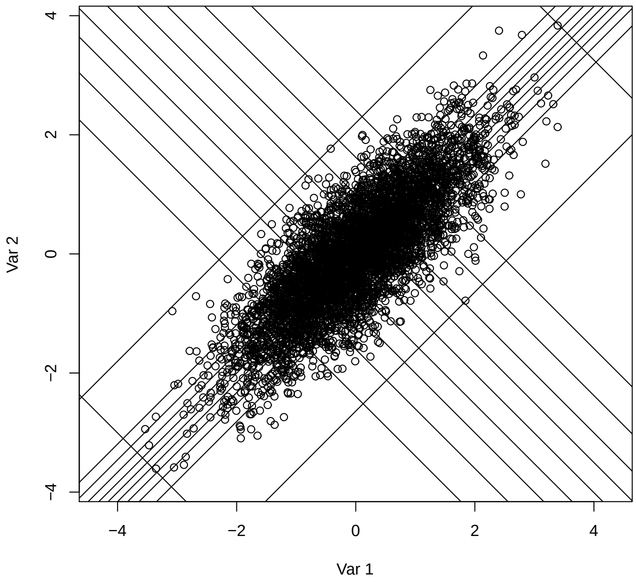


Figure 7.4: The same correlated variables as in figure 7.3, but now with a grid based on empirical deciles derived from the principal components of Var 1 and Var 2. There are equal numbers of samples in each sub-space.

For groups of more than two or three inputs, the standard two level approach is likely to be more efficient due to the curse of dimensionality that plagues our ordered input method. If this is impractical an alternative such as emulation may be necessary (Oakley and O'Hagan, 2004; Oakley, 2009).

Chapter 8

Conclusion

8.1 Summary

We have considered the problem of managing structural uncertainty in health economic decision models. We have seen that we will almost always be in a position of making judgements about the ‘correct’ model structure, or the size of the model error, in the absence of observations on the model output. This difficulty has motivated the development of the model discrepancy approach that we have presented in chapters 5 and 6 of the thesis. In our method we incorporated beliefs about structural error through the addition of discrepancy terms at the *sub-function* level in the model because this was easier than making similar judgements at the level of the model output. Adding discrepancy terms at the sub-function level allowed us to understand the relative effect of the different structural uncertainties on the model output and on the decision. This was helpful in guiding choices about model improvement.

In the first case study we used the discrepancy method to determine the sources of structural error that had an important impact on the output uncertainty and hence were able to make a rational choice about how best to improve the model. In a complex model it may not be at all obvious which are the most important sources of structural error, and so the method reveals features of the model that are otherwise hidden.

In the second case study we showed how it is possible using the sub-function discrepancy method to establish an upper bound on the expected value of model improvement (EVMI) using value of information methods. This approach will be most valuable in cases where the decision problem is complex, but due to difficulties in obtaining input parameter estimates or lack of time or resources we have built a simple model. We feel that this may be of particular relevance in the emerging field of economic evaluation of public health interventions where decision problems generally have many complex elements, but models are often relatively simple (for good examples see descriptions of the models that have been used by the National Institute for Health and Clinical Excellence to support public health intervention resources allocation decisions in England[1]).

We believe the approach offers some advantages over model averaging methods where, in the absence of data, elicitation of model weights is required. Making probability statements about models, which are by definition abstract non-observables is likely to be very difficult. The sub-function discrepancy terms identified in our method are, by contrast, defined such that they relate to observables, precisely so that judgements about them are easier to elicit.

A model's structure rests upon a series of assumptions regarding the relationships between the inputs, the intermediate parameters and the output. In any modelling process it is unavoidable that such assumptions are made, and in one sense model building is just a formal representation of a set of assumptions in mathematical functional form. Health economic modellers sometimes explore the sensitivity of the model prediction to underlying assumptions in a 'what if' scenario analysis in which sets of alternative assumptions are modelled. However, this process cannot in any formal sense quantify the sensitivity of the results to the assumptions, and nor can it quantify any resulting prediction uncertainty. Our method is an attempt to formally quantify the effect of all assumptions in the model about which we do not have complete certainty.

The method is most useful as a sensitivity analysis tool, highlighting areas of the model that may require further thought. However, if the modeller can satisfactorily specify a joint distribution for the inputs and the discrepancies,

[1] http://www.nice.org.uk/Guidance/PHG/Published

then the method results in a proper quantification of uncertainty about the ‘true’ incremental net benefit of one decision over an alternative, taking into account judgements about both parameters and structure.

We determined the sensitivity of the model to the discrepancies from two perspectives, variance based, and value of information. The variance based approach, which we applied in case study 1 in chapter 5, quantifies the expected reduction in variance on learning the value of an uncertain input or discrepancy term. This gives us a direct measure of the sensitivity of the output to variations in an input or discrepancy, taking into account any correlations. However, it does have limitations. In this formulation the main effect index can only be calculated for a scalar model output, and more importantly it does not tell us about *decision* uncertainty. If there is very little decision uncertainty in a particular decision problem, then even an input that has a very large main effect index will have a very small EVPI. This limitation does not matter if the primary purpose of the discrepancy analysis is to manage the uncertainty in the *prediction* that arises due to uncertainty about model structure, but for a decision model, calculating EVPI will usually be more appropriate.

In case study 2 in chapter 6 we adopted a decision theoretic perspective and quantified the sensitivity of the decision to the uncertain structure by calculating the partial EVPI for the discrepancy terms. We interpret this a being an upper bound on the expected value of model improvement (EVMI). Reviewing the structure of a model may introduce new uncertain inputs, which is why the EVPI for the discrepancies will not necessarily equal the EVMI. The *value* of modelling is rarely discussed in the literature, but implicit in all decisions to commission a model is the belief that it will be worth the resources that are committed for the purpose.

This is an interesting area for further exploration. Is it possible to meaningfully quantify the expected value of building a model in the first place? Given no model, my beliefs about some quantity under decision $d \in \mathcal{D}$ are $p(\boldsymbol{Z}_d)$. If I decide to commission a model $\boldsymbol{Y} = f(\boldsymbol{X})$ to tell me about $\boldsymbol{Z} = (\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_D)$, then my posterior beliefs will be $p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, f)$. The expected value of the model is therefore

$$E_{\boldsymbol{Y},\boldsymbol{X},f}[\max_{d} E_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{X},f}\{U(\boldsymbol{Z}_d)\}] - \max_{d} E\{U(\boldsymbol{Z}_d)\}. \tag{8.1}$$

Can I meaningfully determine $p(\boldsymbol{Y}, \boldsymbol{X}, f)$ and $p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, f)$ without building the model?

In both case studies we specified a joint distribution for the discrepancies in which terms were correlated. This motivated the development of the 'ordered input' method for quantifying the main effect index and the EVPI for correlated terms. In chapter 7 we demonstrated the method in a small case study along with presenting a theoretical justification. We showed that not only does the method allow for correlation, it also provides a gain in efficiency (both statistical and computational) over the standard two level Monte Carlo method.

8.2 Model complexity and parsimony

Current good practice guidance on modelling for health economic evaluation states that a model should only be as complex as necessary (Weinstein et al., 2003), but this well intentioned advice does not actually help us make judgements about how complex any particular model should be. Another guiding principle is the requirement for a model to be comprehensible to the non-modeller: a decision maker's trust in a model can easily be eroded if the model is so complicated that its features cannot be easily communicated (Taylor-Robinson et al., 2008).

Our view is that, in the health economic context, increasing the model complexity can have the effect of transferring uncertainty about structural error, which we express through the specification of model discrepancy terms, to uncertainty about model input parameters. Structural error often arises when a simple model is used to a model a complex real world process, thereby omitting aspects that could effect costs or consequences. If we make the model more complex by including such omitted features, typically we will then have more input parameters in the model.

Increasing the complexity of a model will therefore be desirable if the additional complexity relates to parts of the model in which discrepancy terms are influential, and if we have suitable data to tell us about any extra parameters that are required. This is because, to the decision-maker, data-driven probability distributions for model parameters will be preferable to distributions on discrepancy terms based solely on subjective judgements of the modeller.

Our framework can help guide the choice of model complexity by identifying which discrepancy terms are likely to be important. If we are satisfied that a structural error will have little effect on the model output, then increasing the complexity of the model to reduce such an error is likely to have little benefit.

8.3 How might this work in practice?

We envisage that the sub-function discrepancy approach has the greatest potential if used prospectively during model building. This will allow the modeller to incorporate judgements about structural error as they construct the model, encouraging an explicit recognition of the potential impact of the structural choices.

Model development is a sequential, hierarchical, iterative process of uncovering and evaluating options regarding structure, parameterisation and incorporation of evidence (Chilcott et al., 2010a). The process depends on the modeller developing an understanding of the decision problem, which is by its nature subjective. This understanding of the decision problem is the foundation upon which judgements are made in the model building process, and also provides the basis for making judgements about the likely discrepancy inherent in different model formulations. The essence of the discrepancy approach is that it allows a *formal quantification* of the impact of the choices made throughout the model building process.

Ultimately, the validity of the method relies on the ability to meaningfully specify the joint distribution of inputs and discrepancies, $p(\boldsymbol{X}, \boldsymbol{\delta})$. In both case studies we represented our beliefs about $p(\boldsymbol{X}, \boldsymbol{\delta})$ fairly crudely, making assumptions of independence between inputs and discrepancies and independence between groups of discrepancies that were not otherwise constrained. Whilst we felt that this was sufficient in the case studies for the purposes of identifying important model sub-functions we recognise that making defensible judgements about model discrepancies is in general likely to be difficult. If we wish to proceed to a full quantification of our uncertainty about the target quantity then a more sophisticated specification of $p(\boldsymbol{X}, \boldsymbol{\delta})$ will typically be required.

We could choose to make only a crude specification of uncertainty, as long as we are 'generous' with our uncertainty. The expected value of learning $\boldsymbol{\delta}$ will then provide an upper bound on the value of better modelling. If EVPI($\boldsymbol{\delta}$) is small compared with the value of learning the inputs, even with the generous estimate of uncertainty about the structural error, then we can be reassured that the current model as 'good enough'. In contrast, if EVPI($\boldsymbol{\delta}$) dominates EVPI($\boldsymbol{X}$) then we conclude that it is worthwhile either to think a little harder about the model discrepancy, or to rebuild the model so that it better reflects our beliefs about the relationships between the inputs and the target quantities we wish to predict. Developing practical methods for making helpful judgements about $p(\boldsymbol{X}, \boldsymbol{\delta})$ is an area for future research.

References

Ades, A. E., Lu, G. and Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling, *Medical Decision Making*, **24 (2)**: 207–227.

Bastos, L. S. and O'Hagan, A. (2009). Diagnostics for Gaussian process emulators, *Technometrics*, **51 (4)**: 425–438.

Bauch, C. T., Anonychuk, A. M., Effelterre, T. V., Pham, B. Z. and Merid, M. F. (2009). Incorporating herd immunity effects into cohort models of vaccine cost-effectiveness, *Medical Decision Making*, **29 (5)**: 557–569.

Bayarri, M. J., Berger, J. and Steinberg, D. M. (2009). Special issue on computer modeling, *Technometrics*, **51**: 353–353.

Beck, J. and Pauker, S. G. (1983). The Markov process in medical prognosis, *Medical Decision Making*, **3 (4)**: 419–458.

Bedford, T., Quigley, J. and Walls, L. (2006). Expert elicitation for reliable system design, *Statistical Science*, **21 (4)**: 428–450.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996). A new perspective on priors for generalized linear models, *Journal of the American Statistical Association*, **91 (436)**: 1450–1460.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Chichester: John Wiley.

Bojke, L., Claxton, K., Bravo-Vergel, Y., Sculpher, M., Palmer, S. and Abrams, K. (2010). Eliciting distributions to populate decision analytic models, *Value in Health*, **13 (5)**: 557–564.

Bojke, L., Claxton, K., Sculpher, M. and Palmer, S. (2009). Characterizing structural uncertainty in decision analytic models: A review and application of methods, *Value in Health*, **12 (5)**: 739–749.

Bowling, A. (1997). *Measuring Disease: A Review of Quality of Life Measurement Scales*, Buckingham: Open University Press, second edn.

Box, G. E. P. (1976). Science and Statistics, *Journal of the American Statistical Association*, **71 (356)**: 791–799.

Brazier, J. (1993). The SF-36 health survey questionnaire - a tool for economists, *Health Economics*, **2 (3)**: 213–215.

Brazier, J., Ratcliffe, J., Salomon, J. and Tsuchiya, A. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*, Oxford: Oxford University Press.

Brennan, A. and Akehurst, R. (2000). Modelling in health economic evaluation: What is its place? What is its value?, *PharmacoEconomics*, **17 (5)**: 445–459.

Brennan, A., Chick, S. E. and Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies, *Health Economics*, **15 (12)**: 1295–1310.

Brennan, A., Kharroubi, S., O'Hagan, A. and Chilcott, J. (2007). Calculating partial expected value of perfect information via Monte Carlo sampling algorithms, *Medical Decision Making*, **27 (4)**: 448–470.

Briggs, A. (2000). Handling uncertainty in cost-effectiveness models, *PharmacoEconomics*, **17 (5)**: 479–500.

Briggs, A., Sculpher, M. and Claxton, K. (2006). *Decision Modelling for Health Economic Evaluation*, Oxford: Oxford University Press.

Brooks, R. (1996). EuroQol: the current state of play, *Health Policy*, **37 (1)**: 53–72.

Bunge, M. (1967). *Foundations of Physics, Springer Tracts in Natural Philosophy*, Berlin: Springer-Verlag, second edn.

Buxton, M. J., Drummond, M. F., Van Hout, B. A., Prince, R. L., Sheldon, T. A., Szucs, T. and Vray, M. (1997). Modelling in economic evaluation: An unavoidable fact of life, *Health Economics*, **6 (3)**: 217–227.

Cagliano, A. C., Grimaldi, S. and Rafele, C. (2011). A systemic methodology for risk management in healthcare sector, *Safety Science*, **49 (5)**: 695–708.

Cameron, C. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.

Chancellor, J. V., Hill, A. M., Sabin, C. A., Simpson, K. N. and Youle, M. (1997). Modelling the cost effectiveness of lamivudine/zidovudine combination therapy in HIV infection, *PharmacoEconomics*, **12 (1)**: 54–66.

Chilcott, J., Brennan, A., Booth, A., Karnon, J. and Tappenden, P. (2003). The role of modelling in prioritising and planning clinical trials, *Health Technology Assessment*, **7 (23)**.

Chilcott, J., Tappenden, P., Paisley, S., Kaltenthaler, E. and Johnson, M. (2010a). Choice and judgement in developing models for health technology assessment; a qualitative study, *ScHARR Discussion Paper*, Available from `http://www.shef.ac.uk/scharr/sections/heds/dps-2010.html`.

Chilcott, J., Tappenden, P., Rawdin, A., Johnson, M., Kaltenthaler, E., Paisley, S., Papaioannou, D. and Shippam, A. (2010b). Avoiding and identifying errors in health technology assessment models, *Health Technology Assessment*, **14 (25)**.

Claxton, K. (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies, *Journal of Health Economics*, **18 (3)**: 341–364.

Claxton, K. (2008). Exploring uncertainty in cost-effectiveness analysis, *PharmacoEconomics*, **26 (9)**.

Claxton, K. and Posnett, J. (1996). An economic approach to clinical trial design and research priority-setting, *Health Economics*, **5 (6)**: 513–524.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., Brazier, J. and O'Hagan, T. (2005). Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra, *Health Economics*, **14 (4)**: 339–347.

Claxton, K., Walker, S., Palmer, S. and Sculpher, M. (2010). Appropriate perspectives for health care decisions, Tech. rep., Centre for Health Economics, University of York.

Cooke, R. and Goossens, L. (2000). Procedures guide for structural expert judgement in accident consequence modelling, *Radiation Protection Dosimetry*, **90 (3)**: 303–309.

Craig, P. S., Goldstein, M., Rougier, J. C. and Seheult, A. H. (2001). Bayesian forecasting for complex systems using computer simulators, *Journal of the American Statistical Association*, **96 (454)**: 717–729.

Culyer, A., McCabe, C., Briggs, A., Claxton, K., Buxton, M., Akehurst, R., Sculpher, M. and Brazier, J. (2007). Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence, *Journal of Health Services Research & Policy*, **12 (1)**: 56–58.

Da Veiga, S., Wahl, F. and Gamboa, F. (2009). Local polynomial estimation for sensitivity analysis on models with correlated inputs, *Technometrics*, **51 (4)**: 452–463.

Daneshkhah, A. and Oakley, J. (2010). Eliciting multivariate probability distributions, in *Rethinking Risk Measurement and Reporting: Volume I*, edited by Böcker, K., London: Risk Books.

DeGroot, M. (1970). *Optimal Statistical Decisions*, New York, NY: McGraw-Hill.

Devooght, J. (1998). Model uncertainty and model inaccuracy, *Reliability Engineering and System Safety*, **59 (2)**: 171–185.

Dijkstra, T. and Dixon, N. (2010). Climate change and slope stability in the UK: challenges and approaches, *Quarterly Journal of Engineering Geology and Hydrogeology*, **43 (4)**: 371–385.

Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P. and McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation, *Medical Decision Making*, **5 (2)**: 157–177.

Dowie, J. (2006). The Bayesian approach to decision-making, in *Public Health Evidence: Tackling Health Inequalities*, pp. 309–321, Oxford: Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57 (1)**: 45–97.

Drummond, M. F., O'Brien, B., Stoddart, G. and Torrance, G. (1997). *Methods for the Economic Evaluation of Health Care Programmes*, Oxford: Oxford University Press, second edn.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. and Stoddart, G. (2005). *Methods for the Economic Evaluation of Health Care Programmes*, Oxford: Oxford University Press, third edn.

Eckermann, S. and Willan, A. R. (2008). The option value of delay in health technology assessment, *Medical Decision Making*, **28 (3)**: 300–305.

Eddy, D. M., Hasselblad, V. and Shachter, R. (1990). An introduction to a Bayesian method for meta-analysis, *Medical Decision Making*, **10 (1)**: 15–23.

Eddy, D. M., Hasselblad, V. and Shachter, R. (1992). *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*, Boston: Academic Press.

Felli, J. C. and Hazen, G. B. (1998). Sensitivity analysis and the expected value of perfect information, *Medical Decision Making*, **18 (1)**: 95–109.

de Finetti, B. (1974). *Theory of Probability 1*, New York: Wiley.

Foster, I. (1995). *Designing and Building Parallel Programs*, Addison-Wesley.

Garthwaite, P. H. and Al-Awadhi, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **63 (1)**: 95–110.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, **100 (470)**: 680–700.

Garthwaite, P. H. and O'Hagan, A. (2000). Quantifying expert opinion in the UK water industry: An experimental study, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49 (4)**: 455–477.

Gelman, A. (2008). Objections to Bayesian statistics, *Bayesian Analysis*, **3 (3)**: 445–450.

Girling, A. J., Freeman, G., Gordon, J. P., Poole-Wilson, P., Scott, D. A. and Lilford, R. J. (2007). Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy, *International Journal of Technology Assessment in Health Care*, **23 (2)**: 269–277.

Gold, M. R., Siegel, J. E., Russell, L. B. and Weinstein, M. C. (1996). *Cost-Effectiveness in Health and Medicine*, Oxford: Oxford University Press.

Goldstein, M. (1992). Bayes linear methods 1 - adjusting beliefs: concepts and properties, Tech. rep.

Goldstein, M. (2011). External Bayesian analysis for computer simulators, in *Bayesian Statistics*, edited by Bernardo, J. M., Bayarri, M., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M. and West, M., vol. 9, Oxford: Oxford University Press.

Goldstein, M. and Rougier, J. (2004). Probabilistic formulations for transferring inferences from mathematical models to physical systems, *SIAM Journal on Scientific Computing*, **26 (2)**: 467–487.

Goldstein, M. and Rougier, J. (2006). Bayes linear calibrated prediction for complex systems, *Journal of the American Statistical Association*, **101 (475)**: 1132–1143.

Goldstein, M. and Rougier, J. (2009). Reified Bayesian modelling and inference for physical systems, *Journal of Statistical Planning and Inference*, **139 (3)**: 1221–1239.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data (with discussion), *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **168 (2)**: 267–306.

Griffin, S., Claxton, K., Hawkins, N. and Sculpher, M. (2006). Probabilistic analysis and computationally expensive models: Necessary and required?, *Value in Health*, **9 (4)**: 244–252.

Griffin, S. C., Claxton, K. P., Palmer, S. J. and Sculpher, M. J. (2011). Dangerous omissions: the consequences of ignoring decision uncertainty, *Health Economics*, **20 (2)**: 212–224.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*, Oxford: Oxford University Press, third edn.

Groot Koerkamp, B., Weinstein, M. C., Stijnen, T., Heijenbrok-Kal, M. H. and Hunink, M. G. M. (2010). Uncertainty and patient heterogeneity in medical decision models, *Medical Decision Making*, **30 (2)**: 194–205.

Hiance, A., Chevret, S. and Lévy, V. (2009). A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial, *Journal of Clinical Epidemiology*, **62 (4)**: 431–437.e2.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A. and Ryne, R. D. (2005). Combining field data and computer simulations for calibration and prediction, *SIAM Journal on Scientific Computing*, **26 (2)**: 448–466.

Howard, R. A. (1966). Information value theory, *IEEE Transactions on Systems Science and Cybernetics*, **2 (1)**: 22–26.

Jackson, C. H., Bojke, L., Thompson, S. G., Claxton, K. and Sharples, L. D. (2011). A framework for addressing structural uncertainty in decision models, *Medical Decision Making*, **31 (4)**: 662–674.

Jackson, C. H., Sharples, L. D. and Thompson, S. G. (2010). Structural and parameter uncertainty in Bayesian cost-effectiveness models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59 (2)**: 233–253.

Jackson, C. H., Thompson, S. G. and Sharples, L. D. (2009). Accounting for uncertainty in health economic decision models by using model averaging, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172 (2)**: 383–404.

Jacques, J., Lavergne, C. and Devictor, N. (2006). Sensitivity analysis in presence of model uncertainty and correlated inputs, *Reliability Engineering and System Safety*, **91 (10-11)**: 1126–1134.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T. and Feldman, B. M. (2010). Methods to elicit beliefs for Bayesian priors: a systematic review, *Journal of Clinical Epidemiology*, **63 (4)**: 355–369.

Kadane, J. (1994). An application of robust Bayesian analysis to a medical experiment, *Journal of Statistical Planning and Inference*, **40 (2-3)**: 221–232.

Kadane, J. and Wolfson, L. J. (1998). Experiences in elicitation, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47 (1)**: 3–19.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association*, **75 (372)**: 845–854.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection, *Journal of the American Statistical Association*, **99 (465)**: 279–290.

Kandlikar, M., Ramachandran, G., Maynard, A., Murdock, B. and Toscano, W. A. (2007). Health risk assessment for nanoparticles: A case for using expert judgment, in *Nanotechnology and Occupational Health*, edited by Maynard, A. D. and Pui, D. Y. H., pp. 137–156, Springer Netherlands.

Karnon, J. (2003). Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation, *Health Economics*, **12 (10)**: 837–848.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association*, **90 (430)**: 773–795.

Keeney, R. and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, New York: John Wiley.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63 (3)**: 425–464.

Kharroubi, S. A., Brennan, A. and Strong, M. (2011). Estimating expected value of sample information for incomplete data models using Bayesian approximation, *Medical Decision Making*, Epub ahead of print http://mdm.sagepub.com/content/early/2011/04/21/0272989X11399920.abstract.

Kim, S.-Y., Goldie, S. J. and Salomon, J. A. (2010). Exploring model uncertainty in economic evaluation of health interventions: The example of rotavirus vaccination in Vietnam, *Medical Decision Making*, **30 (5)**: E1–E28.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting, *Journal of the American Statistical Association*, **52 (278)**: 133–142.

Kinney, P., Roman, H., Walker, K., Richmond, H., Conner, L. and Hubbell, B. (2010). On the use of expert judgment to characterize uncertainties in the health benefits of regulatory controls of particulate matter, *Environmental Science and Policy*, **13 (5)**: 434–443.

Klügel, J.-U. (2008). Seismic hazard analysis — Quo vadis?, *Earth-Science Reviews*, **88 (1-2)**: 1–32.

Koerkamp, B. G., Myriam Hunink, M. G., Stijnen, T. and Weinstein, M. C. (2006). Identifying key parameters in cost-effectiveness analysis using value of information: a comparison of methods, *Health Economics*, **15 (4)**: 383–392.

Kotiadis, K. and Robinson, S. (2008). Conceptual modelling: Knowledge acquisition and model abstraction, in *Winter Simulation Conference, 2008. WSC 2008*, pp. 951 –958.

Krahn, M. and Gafni, A. (1993). Discounting in the economic evaluation of health care interventions, *Medical Care*, **31 (5)**: 403–418.

Kuhnert, P. M., Martin, T. G. and Griffiths, S. P. (2010). A guide to eliciting and using expert knowledge in Bayesian ecological models, *Ecology Letters*, **13 (7)**: 900–914.

Lamb, S. E., Bartlett, H. P., Ashley, A. and Bird, W. (2002). Can lay-led walking programmes increase physical activity in middle aged adults? a randomised controlled trial, *Journal of Epidemiology and Community Health*, **56 (4)**: 246–252.

Leal, J., Wordsworth, S., Legood, R. and Blair, E. (2007). Eliciting expert opinion for economic models: An applied example, *Value in Health*, **10 (3)**: 195–203.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: Evolution, critique, and future directions, *Statistics in Medicine*, **28 (25)**: 3049–3067.

MacDonald, J. A., Small, M. J. and Morgan, M. G. (2008). Explosion probability of unexploded ordnance: Expert beliefs, *Risk Analysis*, **28 (4)**: 825–841.

McCabe, C., Claxton, K. and Culyer, A. J. (2008). The NICE cost-effectiveness threshold: what it is and what that means, *PharmacoEconomics*, **26 (9)**: 733–44.

McDowell, I. (2006). *Measuring Health: A Guide to Rating Scales and Questionnaires*, New York: Oxford University Press, third edn.

McKay, M. and Morrison, J. (1997). Structural model uncertainty in stochastic simulation, in *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics*.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21 (2)**: 239–245.

McKenna, C. and Claxton, K. (2011). Addressing adoption and research design decisions simultaneously, *Medical Decision Making*, **31 (6)**: 853–865.

Miners, A., Sabin, C., Trueman, P., Youle, M., Mocroft, A., Johnson, M. and Beck, E. (2001). Assessing the cost-effectiveness of HAART for adults with HIV in England, *HIV Medicine*, **2 (1)**: 52–58.

Moala, F. A. and O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach, *Journal of Statistical Planning and Inference*, **140 (7)**: 1635–1655.

Montes Diez, R. and Oakley, J. E. (2010). Gaussian processes priors for monotone functions, Tech. rep., Rey Juan Carlos University / University of Sheffield, Unpublished.

Morita, S. (2011). Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West, *Statistics in Medicine*, **30 (17)**: 2090–2097.

Morris, P. A. (1974). Decision analysis expert use, *Management Science*, **20 (9)**: 1233–1241.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion), *Bayesian Statistics*, **6**: 475–501.

Negrín, M. A. and Vázquez-Polo, F.-J. (2008). Incorporating model uncertainty in cost-effectiveness analysis: a Bayesian model averaging approach, *Journal of health economics*, **27 (5)**: 1250–9.

Nestorov, I., Rowland, M., Hadjitodorov, S. and Petrov, I. (1999). Empirical versus mechanistic modelling: Comparison of an artificial neural network to a mechanistically based model for quantitative structure pharmacokinetic relationships of a homologous series of barbiturates, *The AAPS Journal*, **1 (4)**: 5–13.

Neumann, P. J. (2005). *Using Cost-Effectiveness Analysis in Health Care*, Oxford: Oxford University Press.

NICE (2006). Four commonly used methods to increase physical activity: PH2, Tech. rep., NICE.

NICE (2008). Guide to the methods of technology appraisal, Tech. rep., NICE.

NICE (2009). The guidelines manual 2009, Tech. rep., NICE.

Nilsen, T. and Aven, T. (2003). Models and model uncertainty in the context of risk analysis, *Reliability Engineering and System Safety*, **79 (3)**: 309–317.

Oakley, J. (2010). Eliciting univariate probability distributions, in *Rethinking Risk Measurement and Reporting: Volume I*, edited by Böcker, K., London: Risk Books.

Oakley, J. and O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs, *Biometrika*, **89 (4)**: 769–784.

Oakley, J. E. (2009). Decision-theoretic sensitivity analysis for complex computer models, *Technometrics*, **51 (2)**: 121–129.

Oakley, J. E. (2011). Modelling with deterministic computer models, in *Simplicity, Complexity and Modelling*, edited by Christie, M., Cliffe, A., Dawid, P. and Senn, S. S., John Wiley and Sons.

Oakley, J. E., Brennan, A., Tappenden, P. and Chilcott, J. (2010). Simulation sample sizes for Monte Carlo partial EVPI calculations, *Journal of Health Economics*, **29 (3)**: 468–477.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity of complex models: a Bayesian approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**: 751–769.

Oakley, J. E. and O'Hagan, A. (2007). Uncertainty in prior elicitations: a nonparametric approach, *Biometrika*, **94**: 427–441.

O'Connor, G. T. and Sox, H. C. (1991). Bayesian reasoning in medicine, *Medical Decision Making*, **11 (2)**: 107–111.

O'Hagan, A. (1992). Some Bayesian numerical analysis, in *Bayesian Statistics*, edited by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., vol. 4, pp. 345–363, Oxford University Press.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications, *Journal of the Royal Statistical Society. Series D (The Statistician)*, **47 (1)**: 21–35.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*, Chichester: John Wiley and Sons.

O'Hagan, A. and Oakley, J. E. (2004). Probability is perfect, but we can't elicit it perfectly, *Reliability Engineering and System Safety*, **85 (1-3)**: 239–248.

O'Neill, S. J., Osborn, T. J., Hulme, M., Lorenzoni, I. and Watkinson, A. R. (2008). Using expert knowledge to assess uncertainties in future polar bear populations under climate change, *Journal of Applied Ecology*, **45 (6)**: 1649–1659.

Oremus, M., Collet, J.-P., Corcos, J. and Shapiro, S. H. (2002). A survey of physician efficacy requirements to plan clinical trials, *Pharmacoepidemiology and Drug Safety*, **11 (8)**: 677–685.

Orton, T. G., Goulding, K. W. T. and Lark, R. M. (2011). Geostatistical prediction of nitrous oxide emissions from soil using data, process models and expert opinion, *European Journal of Soil Science*, **62 (3)**: 359–370.

Paddock, S. M. and Ebener, P. (2009). Subjective prior distributions for modeling longitudinal continuous outcomes with non-ignorable dropout, *Statistics in Medicine*, **28 (4)**: 659–678.

Possolo, A. (2010). Copulas for uncertainty analysis, *Metrologia*, **47 (3)**: 262–271.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raiffa, H. (1968). *Decision Analysis. Introductory Lectures on Choices Under Uncertainty*, Reading, Massachusetts: Addison-Wesley.

Ramachandran, G., Banajee, S. and Vincent, J. (2003). Expert judgement and occupational hygiene: Application to aerosol speciation in the nickel primary production industry, *Annals of Occupational Hygiene*, **47 (6)**: 461–475.

Robinson, S. (2008). Conceptual modelling for simulation Part I: definition and requirements, *Journal of the Operational Research Society*, **59 (3)**: 278–290.

Rojnik, K. and Naversnik, K. (2008). Gaussian process metamodeling in Bayesian value of information analysis: A case of the complex health economic model for breast cancer screening, *Value in Health*, **11 (2)**: 240–250.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, **81 (3-4)**: 247–264.

Ryan, M., Scott, D. A., Reeves, C., Bate, A., van Teijlingen, E. R., Russell, E. M., Napper, M. and Robb, C. M. (2001). Eliciting public preferences for healthcare: a systematic review of techniques., *Health Technology Assessment*, **5 (5)**: 1–186.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*, Chichester: John Wiley and Sons Ltd.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*, New York: Springer.

Savage, L. (1954). *The Foundations of Statistics*, New York: Wiley.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations, *Journal of the American Statistical Association*, **66 (336)**: 783–801.

Senn, S. (2008). Comment on article by Gelman, *Bayesian Analysis*, **3**: 459–462.

Shardell, M., Scharfstein, D. O., Vlahov, D. and Galai, N. (2008). Sensitivity analysis using elicited expert information for inference with coarsened data: Illustration of censored discrete event times in the AIDS link to intravenous experience (ALIVE) study, *American Journal of Epidemiology*, **168 (12)**: 1460–1469.

Simpson, K. N., Luo, M. P., Chumney, E., Sun, E., Brun, M. and Ashraf, T. (2004). Cost-effectiveness of Lopinavir/Ritonavir versus Nelfinavir as the first-line highly active antiretroviral therapy regimen for HIV infection, *HIV Clinical Trials*, **5 (5)**: 294–304.

Smith, J. Q. (2010). *Bayesian Decision Analysis. Principles and Practice*, Cambridge: Cambridge University Press.

Soares, M. O., Bojke, L., Dumville, J., Iglesias, C., Cullum, N. and Claxton, K. (2011). Methods to elicit experts' beliefs over uncertain quantities: application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration, *Statistics in Medicine*, e-pub ahead of print.

Sonnenberg, F. A. and Beck, J. R. (1993). Markov models in medical decision making, *Medical Decision Making*, **13 (4)**: 322–338.

Spiegelhalter, D. J. and Best, N. G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling., *Statistics in Medicine*, **22 (23)**: 3687–709.

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling, *Technometrics*, **29 (2)**: 143–151.

Stevens, J. W. 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 (4)**: 782–790.

Stevenson, M. D., Oakley, J. and Chilcott, J. B. (2004). Gaussian process modeling in conjunction with individual patient simulation modeling: A case study describing the calculation of cost-effectiveness ratios for the treatment of established osteoporosis, *Medical Decision Making*, **24 (1)**: 89–100.

Stevenson, M. D., Oakley, J. E., Chick, S. E. and Chalkidou, K. (2008). The cost-effectiveness of surgical instrument management policies to reduce the risk of vCJD transmission to humans, *Journal of the Operational Research Society*, **60 (4)**: 506–518.

Stevenson, M. D., Oakley, J. E., Lloyd Jones, M., Brennan, A., Compston, J. E., McCloskey, E. V. and Selby, P. L. (2009). The cost-effectiveness of an RCT to establish whether 5 or 10 years of bisphosphonate treatment is the better duration for women with a prior fracture, *Medical Decision Making*, **29 (6)**: 678–689.

Stinnett, A. A. and Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis, *Medical Decision Making*, **18 (2)**: S68–80.

Streiner, D. L. and Norman, G. R. (1995). *Health Measurement Scales: A Practical Guide to Their Development and Use*, Oxford: Oxford University Press, second edn.

Tappenden, P., Chilcott, J. B., Eggington, S., Oakley, J. and McCabe, C. (2004). Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-γ and glatiramer acetate for multiple sclerosis, *Health Technology Assessment*, **8 (27)**.

Taylor-Robinson, D., Milton, B., Lloyd-Williams, F., O'Flaherty, M. and Capewell, S. (2008). Policy-makers' attitudes to decision support models for coronary heart disease: a qualitative study, *Journal of Health Services Research Policy*, **13 (4)**: 209–214.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2009). Bias modelling in evidence synthesis, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172 (1)**: 21–47.

van der Voet, H., van der Heijden, G. W., Bos, P. M., Bosgra, S., Boon, P. E., Muri, S. D. and Bruschweiler, B. J. (2009). A model for probabilistic health impact assessment of exposure to food chemicals, *Food and Chemical Technology*, **47 (12)**: 2926–2940.

Weinstein, M. C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C. and Luce, B. R. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices-modeling studies, *Value in Health*, **6 (1)**: 9–17.

Weinstein, M. C. and Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices, *New England Journal of Medicine*, **296 (13)**: 716–721.

White, I. R., Higgins, J. P. T. and Wood, A. M. (2008). Allowing for uncertainty due to missing data in meta-analysis Part 1: Two-stage methods, *Statistics in Medicine*, **27 (5)**: 711–727.

Wilson, A. G., Anderson-Cook, C. M. and Huzurbazar, A. V. (2011). A case study for quantifying system reliability and uncertainty, *Reliability Engineering and System Safety*, **96 (9)**: 1076–1084.

Zaric, G. S. (2003). The impact of ignoring population heterogeneity when Markov models are used in cost-effectiveness analysis, *Medical Decision Making*, **23 (5)**: 379–386.

Zio, E. and Apostolakis, G. E. (1996). Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories, *Reliability Engineering and System Safety*, **54 (2-3)**: 225–241.

Zohar, S., Baldi, I., Forni, G., Merletti, F., Masucci, G. and Gregori, D. (2011). Planning a Bayesian early-phase phase I/II study for human vaccines in HER2 carcinomas, *Pharmaceutical Statistics*, **10 (3)**: 218–226.