USING EXPERTS' BELIEFS TO INFORM PUBLIC POLICIES IN HEALTH: CAPTURING AND USING THE VIEWS OF MANY

Dina Jankovic

PhD

Health Sciences

University of York

October 2018

ABSTRACT

Cost-effectiveness decision modelling (CEDM) is widely used to inform healthcare resource allocation, however there is often a paucity of data to quantify the level of uncertainty around model parameters. Expert elicitation has been proposed as a method for quantifying uncertainty when other sources of evidence are not available.

Elicitation refers to formal processes for quantifying experts' beliefs about uncertain quantities, typically as probability distributions. It is generally conducted with multiple experts to minimise bias and ensure representation of experts with different perspectives.

In CEDM, priors are most commonly elicited from individual experts then pooled mathematically into an aggregate prior that is subsequently used in the model. When pooling priors mathematically, the investigator must decide whether to weight all experts equally or assume that some experts in the sample should be given 'more say'. The choice of method for deriving weights for experts' priors can affect the resulting estimates of uncertainty, yet it is not clear which method is optimal.

This thesis develops an understanding of the methods for deriving weights in opinion pooling.

A literature review is first conducted to identify the existing methods for deriving weights. Differences between the identified methods are then analysed and discussed in terms of how they affect the role of each method in elicitation.

The developed principles are then applied in a case study, where experts' priors on the effectiveness of a health intervention are elicited, and used to characterise parametric uncertainty in a CEDM. The findings are used to analyse and compare different methods for weighting priors, and to observe the consequences of using different methods in the decision model.

The findings improve the understanding of how different weighting methods capture experts' 'contributions' while the choice of methods for deriving weights is found to influence the decision generated by the model.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The completion of this thesis was made possible by a studentship awarded by the Centre for Health Economics, University of York, and I would like to thank them sincerely for this opportunity. Taking on this challenge was an easy decision to make given CHE's fantastic reputation, and indeed, everything I had heard about the department turned out to be true - working here has been inspiring, educational and in no small part fun!

I am particularly grateful to my supervisors, Dr Laura Bojke and Dr Mona Kanaan. While their technical knowledge was invaluable, it is their approach to work – pragmatism, ambition and uncompromising focus on the 'bigger picture' – that has been truly inspiring, and is the most valuable lesson I take from this experience. I will strive to apply the same approach and values throughout my career, and who knows, maybe even my time management will live up to the standard one day!

I would also like to thank my Thesis Advisory Panel – Prof. Katherine Payne, Prof. Gerry Richardson and Dr Marta Soares – for their patience, feedback and thought provoking discussions. I am particularly grateful to Katherine for pushing me to apply scientific rigour in every aspects of my work and for emphasising the importance of presenting work clearly. I thank Gerry (the good cop) for making me perfect my thesis 'pub pitch' and Marta (the other good cop) for her perfectionism, unprecedented attention to detail, and always finding the time to read my drafts.

I take this opportunity to acknowledge other departments and individuals who have made the completion of this thesis possible. I thank the York Trials Unit, particularly Prof. David Torgerson and Belen Corbacho Martin for providing a case study for my thesis – the REFORM trial – and for helping me organise the pilot elicitation exercise. I would like to thank all experts who took time out of their busy days to take part in my study, and those who encouraged them to do so. I am particularly grateful to Katie Robinson at AGILE, and Geraint Collingridge, Dr James Reid, Dr Huma Naqvi and Dr David Broughton at the British Geriatric Society.

Finally, I would like to thank my friends and family for continued support and encouragement over the years, and most of all, my loving, encouraging, supportive, and patient husband-to-be Gareth, who was there unreservedly through my ups and downs, and kept the house in one piece during my write up. Thank you.

SEMINAR PRESENTATIONS

‘Expert elicitation for informing uncertainty in cost-effectiveness decision modelling: an application to evaluate the cost-effectiveness of a podiatry intervention for reducing the rate of falls’ Centre for Research on Health and Social Care Management Seminar, Bocconi University, Milan (Forthcoming, June 2017).

‘Choosing experts for expert elicitation: an application to evaluate the cost-effectiveness of a podiatry intervention for reducing the rate of falls’ Economic Evaluation Seminar, University of York (April 2017).

‘Expert elicitation for informing uncertainty in cost-effectiveness decision modelling: an application to evaluate the cost-effectiveness of a podiatry intervention for reducing the rate of falls’ Academic Unit of Health Economics Seminar, University of Leeds (February 2017)

‘Expert elicitation for informing temporal uncertainty: effect of expert characteristics on perception of uncertainty’ Evidence Synthesis & Modelling for Health Improvement Seminar, University of Exeter (March 2016)

‘A policy model of alcohol-related harms for predicting life years and quality-adjusted life years’ Discussant at the Health Economics Study Group Conference, University of Lancaster (June 2015).

DECLARATION

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1. Introduction

1.1. Introduction

The aim of this thesis is to improve the methodology for expert elicitation when used to characterise uncertainty in cost-effectiveness decision modelling in health.

This chapter sets the scene for the thesis. First, section 1.2 gives context by introducing the role of economic evaluation in medical decision making and discusses the use of cost-effectiveness decision modelling. The section concludes by highlighting the need to represent uncertainty in cost-effectiveness analysis. Section 1.3 introduces expert elicitation as a tool for informing uncertainty when data is sparse and highlights methodological challenges that constrain its implementation. Section 1.4 then sets the thesis objectives and outlines the structure.

1.2. Context: decision making for resource allocation in healthcare

1.2.1. Rationale for efficient resource use in healthcare

Spending on healthcare is rising. According to the World Health Organisation (WHO) Health Expenditure Database, 30 out of 35 countries in the Organisation for Economic Co-operation and Development (OECD) increased spending on healthcare as percentage of their national Gross Domestic Product (GDP) between 2000 and 2015 (World Health Organization, 2017). In the UK, healthcare spending increased from 6.9% to 9.1% of GDP between 2000 and 2015.

The rise in spending has largely been attributed to rising demand (Sorenson, Drummond and Bhuiyan Khan, 2013). An ageing population, the availability of new technologies and rising public expectations have all contributed to an increase in demand for healthcare.

Given the increase in spending, there is continued pressure on efficient resource allocation. A report by Monitor, the regulator for health services in England, has estimated that failing to increase annual efficiencies and real term funding would produce a £30 billion gap in funding by 2020/21 given the growing demand (NHS England, 2014).

Economic evaluation is one of the most widely used methods for resource allocation in healthcare, where there are constrained budgets. The next section (1.2.2) defines economic evaluation and describes its role.

1.2.2. The role of economic evaluation in resource allocation in healthcare

Economic evaluation is defined as the comparison of the costs and consequences of alternative competing options. The principle of economic evaluation is that, for an intervention to be funded, its benefits should be greater than its opportunity cost. (Drummond *et al.*, 2015)

Economic evaluation was first conceptualised in 1976 as part of health technology assessment (HTA). HTA refers to a multidisciplinary approach to assessing health technologies to take into account the medical, social, ethical and economic implications of developing and implementing health technologies (Drummond *et al.*, 2015; WHO, 2015). Economic evaluation was first formally introduced in medical decision-making in Australia and Ontario, Canada (Commonwealth Department of Health, 1992; Ministry of Health, 1994). In 1999, the UK established the National Institute for Health and Care Excellence (NICE), an independent body that requires pharmaceutical companies to demonstrate value for money for newly developed medicines in order to be funded by the National Health Service (NHS). Since 1999, its role has expanded from decisions about whether to approve new medication to resource allocation for public health interventions, diagnostics and medical devices. In this thesis, the term 'technologies' collectively refers to medications, interventions, services, devices and diagnostics.

1.2.3. Analytical frameworks for economic evaluation

There are three broad approaches to economic evaluation described in the literature: cost-benefit analysis, cost-effectiveness analysis and cost-utility analysis (Drummond *et al.*, 2015).

Cost-benefit analysis takes a wide approach of including all societal costs and benefits pertaining to an intervention, including effects on health, time out of paid and unpaid employment, cost to the health system and cost to patients (Drummond *et al.*, 2015). The benefits are valued in monetary units, or willingness to pay (Johanesson, 1995; Pauly, 1995), and an intervention is considered to be good value for money if the willingness to pay for it is greater than the cost of the intervention. There are several practical challenges in CBA (Drummond *et al.*, 2015) and these have been discussed extensively. Its use in HTA has been

limited, likely because of the values that underpin it- publicly funded health systems such as the NHS in the UK are tasked with maximising the population health with a fixed healthcare budget, and so do not consider resource use outside the health budget (e.g. cost to employers) and non-health related benefits (e.g. effect on productivity) in their decision making.

Cost-effectiveness analysis (CEA) compares the cost per unit of effect between competing alternatives (Drummond *et al.*, 2015). Effectiveness is measured in terms of clinical outcomes common to all comparators, such as life years saved or infection cases averted, thus comparison across diseases/populations is challenging. A technology is considered to be cost-effective if its marginal product (health gained) is greater than its opportunity cost (health that would be gained if the resources required to fund the technology were spent elsewhere).

Cost-utility analysis is a form of cost-effectiveness analysis where effects are measured in terms of a broad measure of health gain, comparable across therapeutic areas (Drummond *et al.*, 2015). For example, in the UK health system the effect is measured in terms of Quality Adjusted Life Years (QALYs), a measure that considers both the length and the Health Related Quality of Life (HRQoL) on a unified scale, allowing comparison of interventions from all disease areas (Klarman, Francis and Rosenthal, 1968).

The next section describes how cost-effectiveness (including cost-utility) analysis is used to make decisions on resource allocation in healthcare.

1.2.4. Framework for using cost-effectiveness analysis to make resource allocation decisions

One way of expressing the cost-effectiveness of a health technology is to present the Incremental Cost-Effectiveness Ratio (ICER) compared to the marginal opportunity cost to the health care system, k. The ICER is the additional cost per unit of health (e.g. QALY) produced by the technology, calculated relative to the next best alternative, as shown in Equation 1.1. The marginal opportunity cost k is the inverse of the marginal productivity of the health system – the amount of health gained/lost with an increase/decrease in expenditure at the margin (e.g. QALYs gained /lost per £1 increase/decrease in expenditure). This has previously been referred to as the threshold. In the UK, the cost-effectiveness threshold has been assumed to be £20,000-£30,000; however, recent empirical studies have found that the average marginal opportunity cost of health technologies in England is lower (below £14,000) (Claxton *et al.*, 2015).

An intervention is considered to be cost-effective if the ICER<k.

$$ICER = (C_1 - C_2)/(H_1 - H_2)$$ **Equation 1.1**

Where C_1 and H_1 are the cost and effect of the new technology, and C_2 and H_2 are the cost and effect of the next best alternative.

Cost-effectiveness can also be expressed in terms of net benefit, either monetary or health. Net health benefit represents the difference between the health gained and offset through disinvestment elsewhere. It is calculated using Equation 1.2.

$$(H_1 - H_2) > (C_1 - C_2)/k$$ **Equation 1.2**

Net monetary benefit represents the monetary value of health gained, minus the opportunity cost of health lost. It is calculated using Equation 1.3.

$$(C_1 - C_2) < (H_1 - H_2)\ x\ k$$ **Equation 1.3**

These measures of cost effectiveness can be generated using within trial (study) methods or using decision analytic modelling. The remainder of section 1.2 describes the use of decision analytic modelling methods to conduct cost-effectiveness analysis.

1.2.5. Decision modelling as a framework for CEA

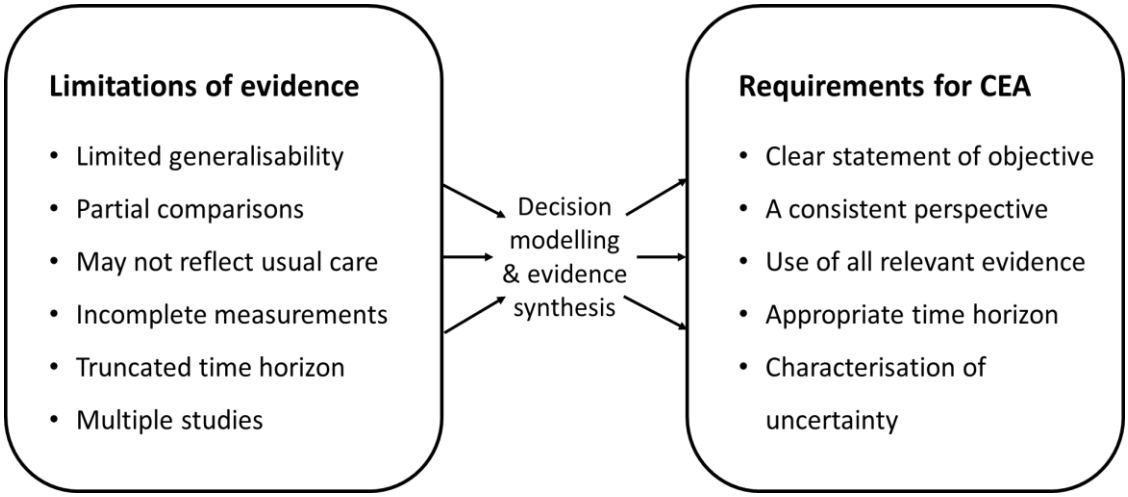
CEA aims to capture the expected costs to the healthcare system and health effects resulting from the technology of interest, and its competing alternatives. (Drummond *et al.*, 2015) While Randomised Controlled Trials (RCTs) can be designed to incorporate information on the HRQoL and resource use, they are rarely sufficient to capture all costs and effects for the target population. Potential limitations include a failure to include all relevant comparators, poor generalisability, or a truncated time horizon failing to capture all costs and effects (Sculpher *et al.*, 2006). Additional analysis is thus almost always required.

Decision modelling and evidence synthesis are the most commonly used frameworks for bridging the gap between the limitations of RCTs and the requirements for CEA. Decision models are a mathematical representation of events that combine multiple information sources to define the possible consequences of the alternative strategies being evaluated. Their role is illustrated in Figure 1.1.

Decision models are informed by a range of sources including RCTs, observational studies, and registries and synthesised data. Ultimately, the quality of evidence used to populate decision

models affects the accuracy of decisions generated by the model (Box, 1979). All evidence has some level of uncertainty and so the resulting decisions are inevitably uncertain.

Figure 1.1. The role of decision modelling and evidence synthesis in bridging the gap between limitations of evidence and requirements for CEA. Adapted from Mahon (2014).



Characterisation of uncertainty in decision models is considered an integral part of decision modelling (Griffin *et al.*, 2011). The next section discusses why representing uncertainty around decisions generated in cost effectiveness decision modelling (CEDM) is important, and section 1.2.7 describes the methods for characterising uncertainty.

1.2.6. The role of uncertainty in cost-effectiveness decision models

Uncertainty in CEDM results means that there is a certain probability that the decision to adopt a new technology will not be the correct one. In some circumstances, the implementation of new technologies can carry irreversible costs such as staff training and equipment. If the decision to implement a new technology is later reversed, these 'sunk' costs will represent a loss in health benefit (for example, QALY loss due to administering the ineffective treatment) (Claxton, Sculpher and Drummond, 2002). Furthermore, approving a new technology can dis-incentivise research, meaning that the cost-ineffective technology could remain in use. Therefore, if the net benefit of an intervention is uncertain, it may be better to delay the approval decision than to risk approving an ineffective technology. (Claxton, Sculpher and Drummond, 2002) Representing uncertainty in CEA can thus inform whether a technology should be implemented and whether further research should be conducted (Claxton, Sculpher and Drummond, 2002). The methods for characterising uncertainty in CEDM are described in the next section.

1.2.7. Methods for modelling uncertainty in CEDM

Uncertainty in cost-effectiveness decision modelling is defined as uncertainty due to limited data or knowledge (e.g. uncertainty around the average population level treatment effect of an intervention). It is important to distinguish uncertainty from variability, which refers to unexplained differences between patients in a population. Unlike variability, uncertainty can be reduced by conducting further research.

There are two approaches to informing uncertainty in decision models: probabilistic and deterministic sensitivity analysis.

Probabilistic sensitivity analysis (PSA) simultaneously combines any uncertainty around each model parameter to quantify the overall uncertainty in outputs, namely on estimates of cost-effectiveness (Claxton *et al.*, 2005). It is conducted by assigning probability distributions to each model parameter, then using Monte Carlo simulations to sample possible parameter values. Each sample thus yields a possible combination of parameter values. Their output (the resulting net benefit) is then recorded and used to calculate the expected net benefit across all random samples. The proportion of times that the intervention yields the highest net benefit (NB) represents the probability that the intervention will be cost effective.

Deterministic sensitivity analysis can be univariate or multivariate and involves changing the value of one or more parameters and observing the model results.

Claxton (2008) proposed three reasons why PSA is the preferred method for characterising uncertainty in decision models. Firstly, it is required to derive the expected net benefit, as CEDM often employs non-linear models and so the mean values of inputs do not necessarily lead to the mean net benefit. Secondly, while deterministic sensitivity analysis informs what the cost-effectiveness of an intervention is in a particular scenario, PSA informs uncertainty around the decision generated by the model (i.e. the probability that it will be cost-effective). Thirdly, PSA can be used to determine the value of further research given the uncertainty in the model. The methods for using PSA to determine the value of further research are described in the next section.

1.2.8. Value of Information analysis

Information on parameter uncertainty can be used to conduct Value of Information (VOI) analysis. The Estimated Value of Perfect Information (EVPI) (the outcomes of VOI analysis) represents the difference between the NB under the current state of knowledge, and the NB if

the optimum treatment was always chosen, i.e. if all uncertainty was resolved. Research required to resolve uncertainty is considered to be worthwhile only if its cost is lower than the value of resolving the uncertainty.

The EVPI is calculated using Equation 1.4. Since the net benefit of choosing the optimum treatment is not known with certainty, it is derived by calculating the net benefit for 'possible outcomes'. The methods are illustrated in more detail in Table 1.1, where the expected net benefit of two competing technologies, A and B, is used to derive the value of further research. Each row in the table represents a possible outcome, derived from a combination of possible parameter values (by sampling from their probability distributions). The expected net benefit of treatment B - the best treatment - is 13 but the probability of error is 0.4, as B leads to lower net benefit for 2 out of 5 samples. Choosing the best treatment every time (i.e. $E_{\Theta} max_j NB(j, \Theta)$ in Equation 1.4) results in NB of 13.6 QALYs. Knowing which treatment is optimal thus leads to a potential gain of 0.6 (13.6-13) QALYs.

$$EVPI = E_{\Theta} max_j NB(j, \Theta) - max_j E_{\Theta} NB(j, \Theta) \qquad \textbf{Equation 1.4}$$

Table 1.1. Expected net benefit of two competing technologies, A and B, used to derive the value of further research.

A sample of parameter values	Net benefit			Health benefit of the best choice
	Treatment A	Treatment B	Best choice	
1	9	12	B	12
2	12	10	A	12
3	14	17	B	17
4	11	10	A	11
5	14	16	B	16
Average	**12**	**13**	**B**	**13.6**

1.2.9. Methodological challenges in cost-effectiveness analysis

The benefits of PSA as a tool for exploring the impact of uncertainty in decision models is widely accepted; however, there is often a paucity of data to quantify the level of uncertainty around the missing parameters, particularly in economic evaluation of health interventions, diagnostics and medical devices where evidence of effectiveness is not required for market

approval and so tends to be sparse (Bojke *et al.*, 2017). While methods for incorporating uncertainty into decision models have received substantial attention over the past two decades, quantifying what this uncertainty looks like is more problematic.

Expert elicitation has been proposed as a method for quantifying uncertainty in decision models when other sources of evidence are not available (Bojke *et al.*, 2010). The next section discusses the role of expert elicitation in CEDM.

1.3. Expert elicitation as a tool for informing uncertainty in cost-effectiveness decision models

Expert elicitation refers to a formal, structured process to capture the beliefs of individuals considered to be experts in the relevant topic (O'Hagan *et al.*, 2006). Beliefs can be captured as probability distributions (experts' priors), as point estimate probabilities (e.g. experts' beliefs about the likelihood of distinct model scenarios) or by ranking in order to prioritise research questions or generate an appropriate set of comparators (O'Hagan *et al.*, 2006). In this thesis, elicitation is considered in the context of quantifying uncertainty in CEDM, in particular parameter second order uncertainty.

This section summarises the role of elicitation in CEDM. First, section 1.3.1 summarises its application to date, and section 1.3.2 discusses potential barriers to wider use. The section concludes that uncertainty around the accuracy of elicited priors could be a barrier to its implementation, and so the remainder of the section discusses the role of bias in elicitation: section 1.3.3 summarises causes of bias in decision and risk analysis, while section 1.3.4 describes the elicitation process, and discusses how each step in the process can be used to minimise bias in elicited priors. Section 1.4 then summarises the outstanding methodological challenges in expert elicitation and describes the aims and objectives of this thesis.

1.3.1. Role of elicitation in health care decision making

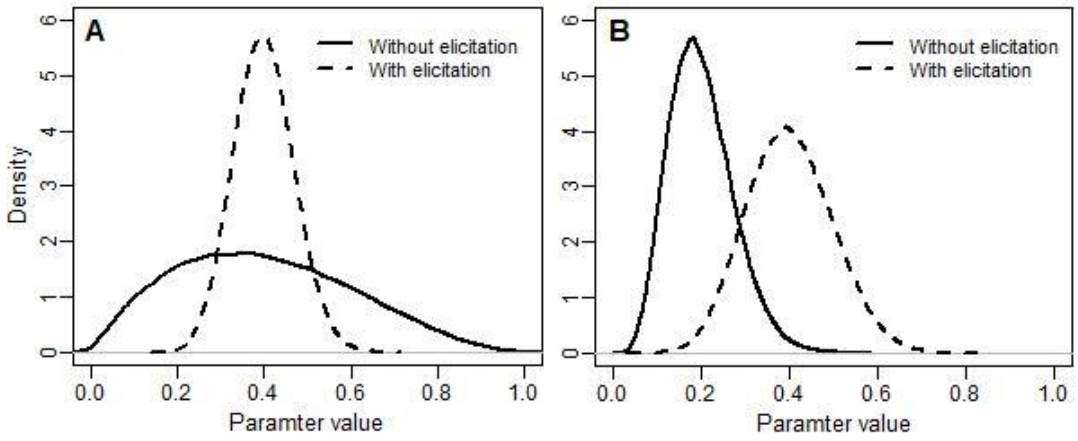
Elicitation has been used in a range of fields including weather forecasting, marine biology, environmental science and nuclear science. It has been used to capture initial estimates of model parameters when other data is not available, and is not attainable within the set time and resource constraints (Cooke, 1991; O'Hagan *et al.*, 2006).

Similarly, elicitation has been used in CEDM to inform uncertainty around various model parameters, including probabilities (or frequencies), time to event, dependency, relative

effectiveness, diagnostic accuracy, minimum important clinical difference, and diffusion (rate of technology adoption) when other data was not available. The role of elicitation in characterising uncertainty when there is no information available is illustrated in Figure 1.2.A, where the uncertainty around the value of a hypothetical parameter is shown with and without elicitation. If parameter uncertainty is high due to no or limited data, and the expert can resolve some uncertainty by eliminating some improbable values from the range or by indicating which values in the range are more likely, then elicitation may be useful.

Other proposed uses for elicitation include improving generalisability and validating or calibrating model estimates (Bojke *et al.*, 2017). Figure 1.2.B shows the probability distribution of a hypothetical parameter, illustrating how elicitation can be used to improve the generalizability of data. Uncertainty around the parameter appears to be lower than for the parameter in Figure 1.2.A, but the study that was used to inform the parameter was carried out on a sample unrepresentative of the target population for the analysis, and so there is uncertainty around its generalisability. Experts' beliefs can thus inform uncertainty around the parameter in the target population, given what has been observed in an alternative population.

Figure 1.2. Uncertainty around hypothetical parameters before and after elicitation. In scenario A elicitation is used to resolve uncertainty. In scenario B elicitation is used to improve the generalisability of a parameter.



The use of elicitation in CEDM has been limited to date. A recent systematic review identified 21 applied elicitation exercises (Soares *et al.*, 2018). However, interest is growing, likely to be driven by the global increased in the use of CEA in medical decision making (as discussed in section 1.2.2). In 2015 the Medical Research Council (MRC) published a call for research on methods for expert elicitation in HTA, and NICE has formally included the use of elicited priors

as a viable method for quantifying uncertainty in decision models in their Diagnostics Assessment Programme manual (National Institute for Health and Care Excellence (NICE), 2011).

The next section discusses the barriers to using expert elicitation in CEDM.

1.3.2. Barriers to using elicitation in CEDM

One of the key barriers to the use of elicitation in CEDM has been the reluctance to base policy on clinical opinion and scepticism around its accuracy. Indeed, elicitation is unlikely to replace clinical trials for the purpose of approving new technologies, but it has been proposed to be useful for quantifying the current state of knowledge in order to understand the key sources of model uncertainty and inform further research (Bojke *et al.*, 2017).

Nevertheless, the risk of inaccurate priors is non-negligible. Historically, there have been numerous occasions where clinicians revealed erroneous beliefs. A commonly cited example in medicine are nurses sabotaging the first RCT assessing the effect of oxygen concentration on retrolental fibroplasia in neonates because they erroneously believed that high oxygen concentrations were beneficial, consequently increasing the number of babies blinded by the condition (Silverman, 1980). While elicitation can help characterise uncertainty, the results of the CEDM are less likely to be useful for decision making if experts are biased.

Accuracy of judgement has been researched in a range of fields and the findings often suggest that the accuracy of predictions tends to be low. Daniel Kahneman, a psychologist specialising in human behaviour and decision making, famously concluded in a study about reliability of professional forecasts and predictions that experts 'produce poorer predictions than dart-throwing monkeys who would have distributed their choices evenly over the options' (Kahneman, Slovic and Tversky, 1982). In health research, Hoffmann and Del Mar (2017) conducted a systematic review of studies comparing clinicians' expectations of the effects of any treatment or test to the effects observed in studies, and found that clinicians' expectations are often inconsistent with observed outcomes: most participants provided a correct estimation of benefits in only 11% of outcomes and correct estimations of harm in only 13% of outcomes. The methods for comparing experts' expectations and observed outcomes will be discussed later in the thesis; however, the study highlights that expert opinion isn't accurate by default.

There is a large body of research on how to minimise bias in elicitation – it relies on understanding the sources of bias, and using formal processes to elicit experts' beliefs in a

manner that minimises bias (O'Hagan *et al.*, 2006). The following section (1.3.3) discusses the sources of bias in elicitation, while section 1.3.4 describes how elicitation methods can be used to minimise bias.

1.3.3. Sources of bias in expert elicitation

In statistics, bias refers to a systematic difference between an estimator of a parameter (in this case the subjective prior) from the true value of that parameter. However, the elicitation literature acknowledges that different experts have different knowledge, levels of experience and opinions, and thus their priors will inevitably differ, and some will be 'closer' to the truth than others. O'Hagan (O'Hagan *et al.*, 2006) proposes that subjectivity in elicited priors is not necessarily 'bad' (O'Hagan *et al.*, 2006). If experts' priors vary due to opinion (for example if two experts disagree about the duration of effect of a new medicine) recruiting multiple experts and combining their priors ensures that all possible views are taken into account and the resulting prior will be impartial. If, however, beliefs are based on inaccurate information (like the nurses in the RCT described in the previous section), then the aggregated priors will be biased, as will the resulting estimate of uncertainty in the model.

Thus in this thesis, the difference between a prior and the true value of a parameter are referred to as accuracy of experts' beliefs[1], while bias is defined as inaccuracy in beliefs caused by prejudice, superstition or irrational beliefs.

There is a vast body of research on the causes of bias. Traditionally, the elicitation literature cites two types of bias, based on their origin: cognitive and motivational (Montibeller and von Winterfeldt, 2015). Cognitive bias refers to illogical inferences arising from irrational processing of information, while motivational bias refers to erroneous, usually conscious, beliefs motivated by one's personal situation. The distinction between the two is not always clear; experts may base their beliefs on an emotional predisposition for or against a particular outcome even if they have no personal stake in it.

Finally, it is important to note that priors may be biased because they represent experts' beliefs inaccurately, rather than because experts' beliefs themselves are inaccurate (O'Hagan *et al.*, 2006). An expert may believe that a new treatment will be effective but may not be able to express their uncertainty accurately in the required format – for example using risk ratios.

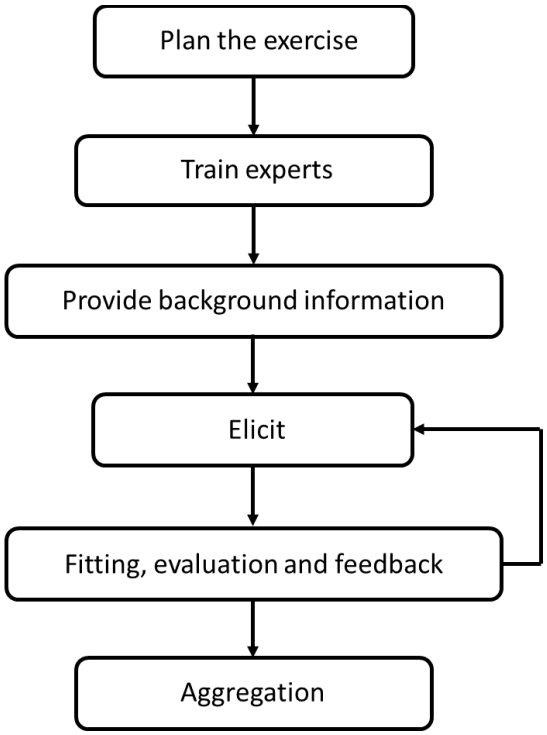
[1] Accuracy of beliefs and its measures are defined later in the thesis, in section 2.4.3.2 in Chapter 2.

The elicitation process aims to minimise the risk of bias in elicited priors by minimising the risk of irrational beliefs and by allowing experts to express their beliefs accurately (O'Hagan *et al.*, 2006). The process involves careful planning and conduct, and transparent reporting. The details of the elicitation process are discussed in section 1.3.4.

1.3.4. The elicitation process

O'Hagan argued that the aim of an elicitation exercise should be to 'ensure that the expert view the problem from as complete a perspective as possible, utilising all relevant information in an unbiased way' (O'Hagan *et al.*, 2006). In order to achieve this, the elicitation process typically consists of six steps as shown in Figure 1.3. The remainder of this section describes each step, in turn.

Figure 1.3. The elicitation process, consisting of six steps: 1) planning, 2) training, 3) assessment of background information, 4) elicitation, 5) fitting, evaluation and feedback, and 6) aggregation.



1.3.4.1. Planning

Elicitation requires careful planning to ensure that the priors represent experts' uncertainty around the parameters of interest, and that any bias is minimised. During the planning stage, a range of decisions have to be made, including who is an expert in the field, which parameters

to elicit, how to elicit them, which elicitation technique to use, how to format the questions and how to deliver the exercise. Each factor is discussed here in further detail.

Who is an expert?

Knowledge in the field for which elicitation is conducted is referred to as **substantive expertise**, and is proposed to minimise the risk of cognitive bias (Tetlock, Gardner and Richards, 2016). For example, clinical knowledge and experience can ensure that experts understand factors that can influence the value of a parameter and give informed, plausible estimates of it. In contrast, experts who lack such knowledge can be overconfident about 'wrong' values for a parameter (assuming the parameter can be measured) by missing important factors that can influence it. Furthermore, substantive knowledge is of value even if all experts are perfectly unbiased. If the aim of an elicitation exercise is to capture the current state of knowledge in the face of limited evidence, it could be argued that asking an uninformed or an inexperienced expert where a more experienced one exists will overestimate uncertainty in the model and the value of further research.

Definitions of substantive experts in the elicitation literature vary. For example, Jenkinson (2005) proposes that elicitation participants should be 'substantive experts in the particular area', Leal et al. (2007) define experts as individuals who have 'specialist knowledge' in the field and Garthwaite et al. (2005) define them as 'persons to whom society and/or his or her peers attribute special knowledge about the matters being elicited'. The exact skills and experience that experts should demonstrate are likely to vary between professions, and depend on experts' availability and willingness to participate.

How many experts?

As discussed in section 1.3.3, elicitation exercises tend to include multiple experts, to capture a range of views (Winkler and Poses, 1993; Clemen and Winkler, 1999). The optimal number of experts is not clear. Knol et al. (2010) argued that between six and twelve experts should be included, while Kattan et al. (2016) found that each additional expert (up to 24 experts in their sample) improved the accuracy of aggregate predictions, although marginal returns were diminishing. Kadane (1986) proposed that experts' perspective should be taken into account, as well as their number. The author argued that the sample of experts should represent the expert community. It is not clear how this can be achieved and demonstrated.

What to elicit?

While the aim of elicitation is to inform model parameters, it has been proposed that experts' ability to give accurate assessments also depends on the parameter itself (Morgan, 2014), and so consideration should be given to whether experts can reasonably assess uncertainty around the required parameter. Probabilities of rare events, parameters influenced by multiple factors, distribution moments (other than the mean) and unobservable parameters have been suggested to be difficult to estimate accurately, and so they tend to be elicited indirectly (Kadane and Wolfson, 1998). For example, a risk ratio cannot be observed directly, so it tends to be derived from elicited priors on the probability of an outcome in those who receive the intervention and those who do not (Bojke *et al.*, 2010; Soares *et al.*, 2011). Kleinmuntz (1996) found that the accuracy of predictions improved when the probabilities of rare events were elicited conditional on preceding events instead of joint.

Furthermore, there may be more than one way to elicit the same parameter: for example the probability that a patient will experience a particular symptom can be elicited as the probability itself, or the time required before a certain proportion of patients experience the symptoms (Bojke *et al.*, 2017). When there is more than one way to inform uncertainty around a parameter, it is not clear which format gives the most accurate estimates.

Elicitation technique

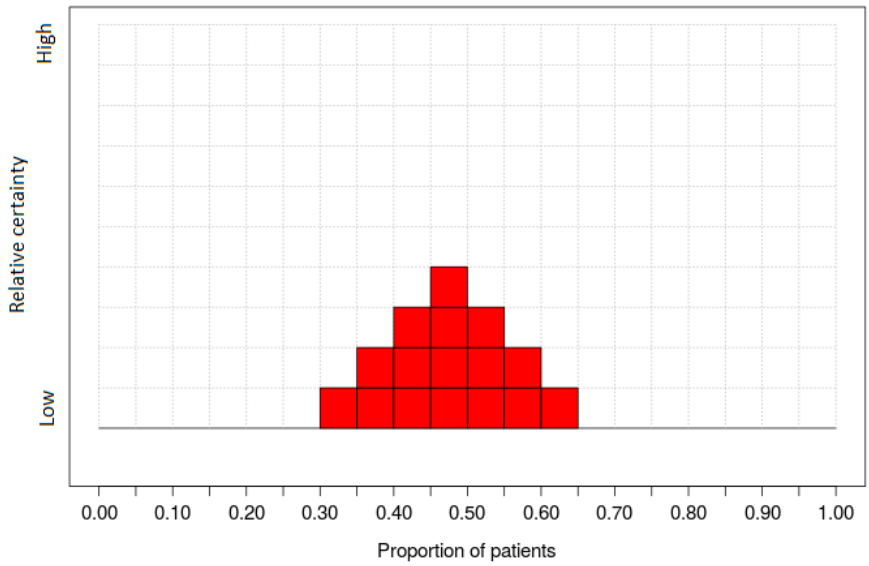
It is generally accepted that experts' beliefs should not be expressed verbally using terms such as 'likely', 'probable' or 'certain' as individuals attach different meaning to these. Thus techniques are employed to elicit their probability distributions numerically. Experts can be asked to specify their probability p that a parameter will take a particular value (or range of values) x, or they can be asked to specify a parameter value x given a probability p (O'Hagan *et al.*, 2006). These summaries can then be used to describe a distribution that represents the expert's beliefs.

For each approach, the investigator can chose from a number of techniques to elicit these values. The most commonly used techniques in HTA are the histogram (also referred to as the 'roulette' or 'chips and bins') method, fixed interval method, and the bisection method. (Soares *et al.*, 2018) The histogram technique involves providing experts with a range of possible values on a grid, and asking them to distribute 'chips' across the range to indicate their uncertainty, as illustrated in Figure 1.4-A. The fixed interval method involves eliciting the range of possible values, splitting it into intervals (for example, four equal intervals) and asking experts to express their probability that the value of the parameter will fall within each

interval. The method is illustrated in Figure 1.4-B. Alternatively, the bisection method asks experts for a set of parameter values that splits the distribution into specific quantiles. For example tertiles can be elicited by asking for two values x_1 and x_2 that split the distribution into three equally probable intervals, so that $P(X < x_1) = P(x_1 < X < x_2) = P(X > x_2) = 0.33$, where X is the expected value of the parameter.

Figure 1.4. Elicitation methods: a) the histogram (chips and bins) method; and b) the fixed interval method.

A Please consider the effectiveness of treatment X. What proportion of patients do you believe will be symptom-free after receiving treatment X? Please use the grid below to indicate your answer.



B Please consider the effectiveness of treatment X. What proportion of patients do you believe will be symptom-free after receiving treatment X?

What is the lowest likely value?	0.3
What is the highest likely value?	0.7
What is the most likely value?	0.5

What is the probability of your estimated value lying in the following intervals?

1) Between 0.3 and 0.4	
2) Between 0.4 and 0.5	
3) Between 0.5 and 0.6	
4) Between 0.6 and 0.7	

Note the values entered in A and B are hypothetical and do not represent the same distribution.

The chosen elicitation technique can affect the results, although it is not clear which method is 'the best'. The histogram technique is generally reported to be easier to use though ease of use can lead experts to put less thought into their responses (Grigore, Peters and Hyde, 2016).

How each technique is implemented can also affect the results (O'Hagan *et al.*, 2009). In the fixed interval and bisection methods, asking experts for their mode or median first has been suggested to lead to overconfidence due to anchoring, and so the range of plausible values tends to be elicited first (Kadane and Wolfson, 1998). The number, the range and the width of bins can affect priors elicited using the histogram technique (O'Hagan *et al.*, 2009), as can the number of quantiles elicited using the bisection method. Some researchers prefer to elicit tertiles (Garthwaite and O'Hagan, 2000) as the accuracy of assessments has been suggested to diminish when more extreme quantiles are elicited (Lichtenstein, Fischhoff and Phillips, 1982), although quartiles are considered to be more intuitive and are used more commonly in CEDM (Soares *et al.*, 2018).

Which quantity to elicit

Probability distributions can be elicited as different quantities, for example:

- **Probability**, e.g. the probability of having a fall in one year is 0.3 in elderly people.
- **Proportions**, e.g. the proportion of elderly people who have a fall every year is 0.3.
- **Percentage**, e.g. 30% of elderly people that have a fall every year.
- **Relative frequency**, e.g. 30 out of every 100 elderly people have a fall every year.
- **Odds**, e.g. the ratio of people who have a fall compared to those who do not is 3:7.
- **Natural frequency**, e.g. of 11.4 million people in the UK aged 60 or older, 3.42 million will have a fall this year.

The above examples are mathematically equivalent, however the elicitation literature suggests that using different quantities for representing uncertainty results in different probability distributions (Koehler, 2001b, 2001a; O'Hagan *et al.*, 2006). Gigerenzer (1996) found that priors elicited as frequencies were less likely to result in error due to miscalculation (e.g. by mistakenly interpreting the probability of 0.3 and 3%). However, Slovic et al. (2000) argued that the frequency effect can be eliminated if the problem is presented clearly, for example using a Venn diagram that clarifies the relations of the problem.

Clear phrasing of questions is important in questionnaire and survey design (Meyer and Booker, 1991). Providing multiple phrasings of the same question has been proposed to help with cognitive biases; examples include asking for the probability of a binary outcome and the counterfactual and highlighting the need for the probabilities to add up to one (Montibeller and von Winterfeldt, 2015).

Mode of delivery

Elicitation can be undertaken on a face-to-face basis with one or more experts and a facilitator, or at distance, for example by email or via the internet (Bojke *et al.*, 2017). Another possibility is to use video conferencing facilities rather than simple surveys to allow some level of interaction. Face to face interviews allow experts to ask for clarification along the way, and they are provided with immediate feedback and interpretation of their responses. Experts may feel more motivated to participate and provide more thoughtful answers than in a remote survey (Bowling, 2005). However, surveys ensure that all experts receive the same information and in the same way, while any interactive features have to be carefully built in a priori. If distance elicitation is to be undertaken, resource needs to be invested in a tool which experts can use themselves with minimal guidance (Grigore *et al.*, 2013; Bojke *et al.*, 2017). Remote elicitation (where the facilitator is not present) has been found to lead to more certain priors; it is not clear whether this is desirable (Nemet, Anadon and Verdolini, 2017).

1.3.4.2. Training

Elicitation requires experts to evaluate their knowledge and experience and then formulate this into beliefs relating to unknown or unobserved parameters. In doing so they must express their uncertainty regarding these parameters. The ability to express one's own beliefs quantitatively is referred to as normative expertise in elicitation (Ferrell, 1994; Stern and Fineberg, 1996; O'Hagan *et al.*, 2006), as it requires a different skillset to that required for substantive expertise in clinical work (such as knowledge of literature and care pathways, or communication skills). The training stage of elicitation should be used to ensure that experts have sufficient normative expertise to assess their beliefs free from bias and express them in the required format.

Training often includes teaching the concept of uncertainty (to ensure the experts are not expressing variability), teaching experts how to use the relevant elicitation technique and debiasing (P Garthwaite, J Kadane and O'Hagan, 2005).

The effectiveness of debiasing techniques varies. Montibeller and von Winterfeldt (2015) proposed that strategy-based (SB) biases, which occur when decision makers use a suboptimal cognitive strategy (such as assigning a higher probability to a conjunction of two events than to each of those events separately), are the easiest to eliminate, for example by teaching experts the meaning of conditional and joint probabilities. Biases that arise due to automatic mental associations are more difficult to eliminate (Montibeller and von Winterfeldt, 2015). Asking practice questions and highlighting bias in beliefs has been shown to be more effective than simply 'warning' experts against it (Fischhoff, Slovic and Lichtenstein, 1977; Siegel-Jacobs and Yates, 1996; Hammersley, Kadous and Magro, 1997), although there is no consensus on what training and practice questions should contain.

Furthermore, it is not clear whether experts require a certain level of normative skills priors to the exercise or whether they can be effectively trained under time constraints of an elicitation exercise (typically half a day). Soares et al. (2011) caution that experts in HTA are likely to be health care professionals who in general have more limited quantitative training (Sabin, 2001). There is a danger that, while they may be taught the constraints of the target parameter (e.g. 0-1 for probabilities) and what probability distributions tend to look like (for example, to avoid U-shaped and bimodal priors), it is not clear whether they can learn to meaningfully assess these.

1.3.4.3. Background information

Providing background information helps experts view the problem from as complete a perspective as possible, utilising all relevant information to minimise the risk of basing their priors on false or incomplete information. Background information can be provided in advance of or during the exercise. The former can allow time for experts to think and consult other references they feel are relevant.

Furthermore, it is not clear whether experts have the analytical skills required to draw correct inferences from evidence.

1.3.4.4. Conducting an elicitation exercise

Section 1.3.4.1 discussed that there are different modes of delivering an elicitation exercise: face-to-face or remote.

The interaction with experts can also affect the results, although the optimal level of interaction is not clear. Investigators' feedback and interpretation can minimise bias by

querying and clarifying apparently irrational beliefs. However, investigators can also lead experts, resulting in priors that do not necessarily represent their beliefs. Similarly, discussion between experts can encourage sharing of information, but some bias can also be induced. For example, dominance of strong-minded or strident individuals, (Knol *et al.*, 2010) common background of group members and socially reinforced irrelevance (i.e. 'taboos') (Ayyub, 2001) can all reinforce biases and lead to overconfidence. (Sniezek, 1992)

1.3.4.5. Fitting, evaluation and feedback

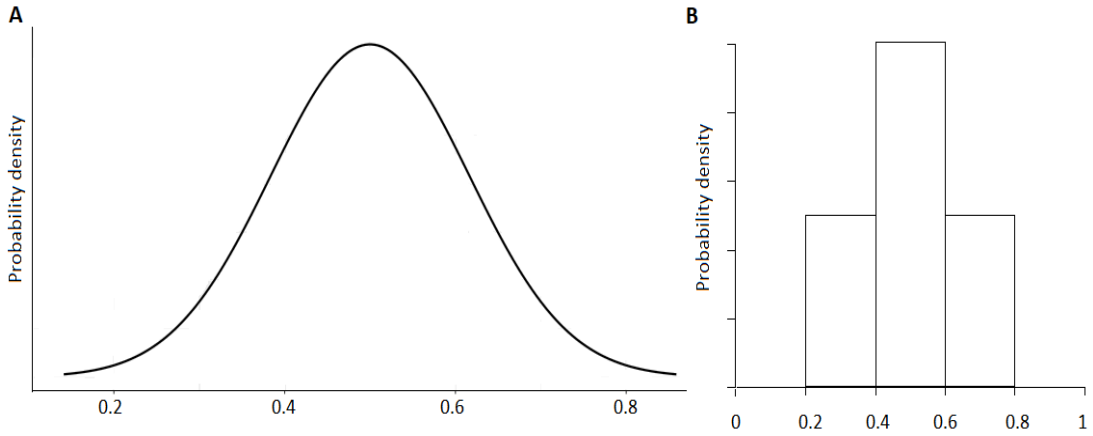
Fitting, evaluation and feedback consists of three steps that can vary in order. Here, each is described in turn.

Fitting

When fixed interval or bisection methods are used, probabilities are elicited on a limited number of intervals for possible parameter values (P Garthwaite, J Kadane and O'Hagan, 2005). The investigator can then either assume that that probability density is uniform across each interval or fit a (smooth) parametric probability distribution, as shown in Figure 1.5.

Fitting can lead to a discrepancy between elicited quantities and those derived from the fitted distribution. This is particularly likely with the histogram method due to the relatively large number of quantities that are elicited (i.e. the probability of each interval). If fitting is carried out during the elicitation exercise, the fitted probability distribution can be fed back to the expert for confirmation. If the distribution is fitted after the elicitation exercise however, the investigator must assume that the fitted probability distribution represents experts' beliefs.

Figure 1.5. Experts' priors when a) Beta distribution is fitted to elicited tertiles; and b) distribution is not fitted.



Evaluation

Evaluation aims to determine 'how well' the elicitation has been done. Evaluation can be based on internal consistency and seed-calibration. (Wallsten and Budescu, 1983)

Internal consistency is the extent to which experts' priors conform to laws of probability; for example, if an expert believes that the probability that a patient is cured is 0.7, then their probability of not being cured should be 0.3. If an expert is aware that the probabilities of all mutually exclusive outcomes of an event should add up to one but their priors suggest otherwise, then their priors can be said not to represent their beliefs.

Seed-calibration involves eliciting experts' beliefs about one or more parameters that are unknown to the expert but known to the investigator ('seed' parameters) (O'Hagan *et al.*, 2006). The closeness of experts' priors and the observed value of seeds is referred to as calibration, and can be used as an indicator of experts' accuracy in elicitation. Poor calibration can be due to inaccurate beliefs, or an inaccurate representation of experts' beliefs (O'Hagan *et al.*, 2006).

There are multiple methods for assigning numerical values, or 'scores' to experts' calibration performance, although it is not clear how they compare and how they should be selected – this topic will be revisited in Chapter 2.

Feedback

Feedback refers to the process of showing experts the fitted distribution or interpreting their probabilities to ensure they represent their beliefs (P Garthwaite, J Kadane and O'Hagan, 2005). The investigator can use this opportunity to identify any irrational beliefs and verify with experts what they thought, or discuss any discrepancies in priors from different experts. Feedback is generally only delivered when elicitation is conducted in person.

1.3.4.6. Aggregation

Aggregation involves combining priors elicited from multiple experts to obtain an 'overall' representation of uncertainty (Clemen and Winkler, 2007). There are two approaches to aggregation: behavioural and mathematical.

Behavioural aggregation focuses on eliciting a single probability distribution from a group of experts. In practice there are several ways to achieve this (Clemen and Winkler, 2007). O'Hagan recommends eliciting experts' individual distributions then using them as a basis for a group discussion (Rohrbaugh, 1981) – differences in priors can be highlighted and used to

identify bias or knowledge asymmetry. For example, if some experts are identified to be basing their priors on false information and other experts are aware of this, they can be prompted in the discussion to update their priors in light of new information. Similarly, if one of the experts possesses more normative skills they can help other experts represent their beliefs. Eventually, experts are asked to agree on a 'rational impartial' prior that represents their combined views.

As discussed in section 1.3.4.4, discussion between experts can lead to bias, and it is not clear whether facilitators can effectively prevent this by managing the discussion. Furthermore, it can be practically challenging to conduct a group discussion, not least because of the practical difficultly in convening all experts on one occasion. This is a particularly common challenge when eliciting the beliefs of clinicians, where experts tend to have irregular and unpredictable working patterns. Grigore et al. (2016) reported that organising one-on-one elicitations took approximately three months with each expert – requiring all experts to be in the same room is likely to take longer, if at all possible.

The DELPHI method was introduced to overcome some of the challenges associated with O'Hagan's method whilst encouraging knowledge sharing between experts, in a more controlled way (Mullen, 2003). The method involves eliciting priors of individual experts then providing controlled, usually written feedback, followed by opportunity to adjust personal priors. Several rounds of feedback and adjustment are repeated until experts reach consensus. DELPHI can be delivered remotely if organising a face-to-face discussion is not possible, and if experts struggle to reach consensus their priors can be aggregated mathematically.

Mathematical aggregation refers to eliciting probability distributions from each expert individually, then aggregating their priors mathematically. (Clemen and Winkler, 2007) It can be applied to priors elicited from each expert independently, or as part of DELPHI if consensus is not reached after a particular number of rounds. There are two methods for combining experts' priors mathematically: Bayesian approaches and opinion pooling.

The Bayesian approach involves using experts' probability assessments to update the decision makers' own prior beliefs about an uncertain parameter. (Moatti *et al.*, 2013) These methods have not yet been applied in HTA and the need for the decision makers input is difficult to implement in practice.

Opinion pooling assumes that the aggregate prior is a function of individual priors. The relationship can be linear (shown in Equation 1.5) or logarithmic (shown in Equation 1.6) (Stone, 1961; Genest and Zidek, 1986).

$$p(\Theta) = \sum_{i=1}^{E} w_i \, p_i(\Theta)$$ **Equation 1.5**

Where i represents one of E experts;

w_i represents the weight assigned to expert i;

$p_i(\Theta)$ is expert i 's prior on parameter Θ.

$$p(\Theta) = k \prod_{i=1}^{E} p_i(\Theta)^{w_i}$$ **Equation 1.6**

Where k is a normalising constant to ensure that the distribution integrates to 1.

Mathematical aggregation requires the investigator to determine the contribution of each expert. They can either ensure all experts contribute equally so that w_i=1/E for all i. Alternatively, weights can be based on some explicit measure of experts' contribution. There are multiple methods for deriving weights, including scoring experts' calibration (discussed in the 'Fitting, evaluation and feedback' section) and clinical experience. It is not clear what the optimal method for deriving weights is – this will be discussed in Chapter 2.

1.3.5. Methodological uncertainties in elicitation

Section 1.3 introduced expert elicitation and discussed its role in characterising uncertainty in CEDMs. The aim of elicitation is to capture the current state of knowledge around uncertain quantities, and the structured elicitation processes described are used to achieve this. In particular, the elicitation process can ensure that experts base their priors on all available information to avoid bias due to inaccurate information and to minimise uncertainty, and to help experts assess uncertainty in their beliefs and express them in the required format, free from bias. The former is achieved by recruiting substantive and impartial experts, providing background information, recruiting multiple experts with different perspectives, and encouraging information sharing. The latter is achieved through careful planning and delivery of the exercise to minimise bias, training, debiasing, delivery, as well as evaluation and feedback.

However, section 1.3.4 highlighted many methodological uncertainties that make it difficult to decide on which methods to use to achieve the stated objectives. For example, it is not clear how to identify substantive experts, how to train experts to ensure they have the normative expertise required to complete the exercise, how to debias effectively, which technique elicits

experts' beliefs most accurately, the optimal mode of delivery, and how experts' priors should be aggregated.

This thesis explores a particular aspect of the elicitation process: the methods for opinion pooling when experts' priors are aggregated mathematically. Section 1.4 provides impetus for studying pooling methods, and outlines the thesis objectives.

1.4. Thesis aims and objectives

This thesis aims to improve elicitation methodology, in particular the methods for opinion pooling in mathematical aggregation.

Mathematical aggregation is appealing in CEDM because it may be more practically feasible as well as methodologically superior in this context. As highlighted in section 1.3.4, summoning multiple clinicians for a group elicitation may not always be possible and mathematical aggregation may be the only viable option. While one-to-one elicitation and behavioural aggregation can be achieved using DELPHI, mathematical aggregation is arguably more transparent than behavioural aggregation, as experts' contribution to the aggregate probability distribution is not based on personality or peer-assessed expertise, both of which have been suggested to be poor indicators of expertise (Bolger, 2017). Indeed, opinion pooling is the most commonly used method for aggregating priors in CEDM (Soares *et al.*, 2018).

Despite the relatively high use of opinion pooling, there is no consensus or guidance on how pooling should be performed. There are choices to be made, in particular where the investigator must decide whether to weight experts' priors (P Garthwaite, J Kadane and O'Hagan, 2005), giving some experts 'more say' than others. In theory, differential weighting can adjust for shortcomings in the elicitation process. For example, if some experts are believed to be more experienced or their experience is more relevant to the topic for which elicitation is being conducted they can contribute to the final (aggregate) probability distribution more. The choice of method for deriving weights for experts' priors can affect the resulting estimates of uncertainty (Cooke, ElSaadany and Huang, 2008), yet it is not clear which method is optimal.

1.4.1. Thesis objectives

This thesis develops methods for opinion pooling to ensure that the aggregate priors are an unbiased representation of the current state of knowledge. Specifically, the thesis addresses the following three objectives:

1. To develop a set of guiding principles for deriving weights in opinion pooling.

This is explored in Chapter 2. First a literature review is conducted to identify existing weighting methods. Then, principles and assumptions that underpin each method are analysed and their role in capturing experts' contribution is discussed. A set of guiding principles is then developed for using each method.

2. To apply the principles developed in Chapter 2 to a case study.

To achieve this, Chapter 3 identifies an appropriate case study (involving a CEDM) and designs an elicitation exercise to inform uncertainty around the model parameters. Chapter 4 provides an overview of the results of the elicitation exercise, while Chapter 5 applies and analyses different weighting methods specific to that example.

3. To observe the consequences of using different methods for opinion pooling.

Chapter 6 applies the different weighting methods analysed in Chapter 5 to a CEDM and observes their effect on the cost-effectiveness decision generated by the model in the case study, and the resulting value of further research.

Chapter 7 then discusses the findings of the thesis, draws conclusions and makes recommendations for further research.

Chapter 2. Methods for opinion pooling in expert elicitation in CEDM

2.1. Introduction

Chapter 1 introduced the role of expert elicitation as a tool for characterising uncertainty in cost-effectiveness decision models (CEDM) and proposed that the aim of an elicitation exercise is to capture the current state of knowledge around uncertain quantities (such as model parameters). To do this, elicitation is conducted using formal processes that encourage experts to use all available information and express their priors in an unbiased way. However, in navigating the choices available in designing and conducting an elicitation exercise, there are many methodological uncertainties.

This chapter explores a particular aspect of the elicitation process: the methods for deriving weights in opinion pooling.

As discussed in Chapter 1, mathematical aggregation is the process of combining priors elicited from multiple experts individually into a single probability distribution that captures uncertainty in the parameter of interest. Opinion pooling, specifically, assumes that the resulting aggregate probability distribution is a function of individual priors (Stone, 1961; Genest and Zidek, 1986).

In opinion pooling, the investigator must decide whether to weight experts' priors (P Garthwaite, J Kadane and O'Hagan, 2005). Differential weighting assumes that some experts should be 'given more say' than others and there are multiple methods for deriving weights for experts. The choice of method for deriving weights for experts' priors can affect the resulting estimates of uncertainty (Cooke, ElSaadany and Huang, 2008), yet it is not clear which method is optimal.

The aim of this chapter is to develop a set of guiding principles for deriving weights. In order to achieve this, the following objectives were set:

1) **To identify existing methods for deriving weights**
2) **To discuss the role of weighting in elicitation**
3) **To evaluate and compare methods for deriving weights**

Section 2.2 describes a literature review conducted to identify existing methods for deriving weights (objective 1). Given the lack of guidance on how to use different methods for deriving weights, section 2.3 discusses factors that provide a basis for differential weighting in opinion pooling (objective 2). To do so, the section revisits the aims of elicitation proposed in Chapter 1, and then discusses factors that could affect experts' contribution towards achieving those aims. Section 2.4 discusses the assumptions, advantages and limitations of existing methods and how these affect their role in elicitation (objective 3).

Section 2.5 then discusses the findings and outlines the structure of the remaining six chapters in the thesis.

2.2. Identification of existing methods for deriving weights

A literature review was conducted to identify existing methods for deriving weights. Section 2.2.1 describes the search strategy and reports the number of studies identified in the review, section 2.2.2 reports the findings, and section 2.2.3 then summarises the findings and outlines the structure of the discussion in the remaining three sections.

2.2.1. Literature review

A literature review was conducted to identify existing methods for deriving weights in opinion pooling. The chosen search strategy in the review was bidirectional citation searching to completion (BCSC).

BCSC is a 'pearl growing' method, where starting citations, or 'initial pearls', are identified through systematic literature searches, informal literature searches or on advice from experts in the field (Hinde and Spackman, 2015). Further literature is then identified by tracing references (backward searching) and citations (forward searching) from the initial pearls. The process is then repeated for the newly identified citations until saturation, when no new relevant citations are identified. BCSC has been proposed to be a useful, even superior alternative to traditional Boolean database searches when conducting explorative literature reviews (Hinde and Spackman, 2015) because Boolean searches rely on correctly identifying all relevant search terms and data bases where relevant literature can be found (Garfield, 2006).

Elicitation is applied in a wide range of fields; the diversity of potentially relevant literature meant that identifying relevant papers would have required searches in numerous databases,

with the use of terminology varying between them, and so BCSC was judged to be a more appropriate search strategy.

Sections 2.2.1.1 - 2.2.1.3 provide details of each step in the BCSC, while section 2.2.1.4 reports the number of studies identified in the review.

2.2.1.1. Identifying initial pearls

Initial pearls were identified on advice from experienced researchers in the field. A pragmatic approach was taken to identify health economists with interest in expert elicitation. Identified researchers were employed by, or had collaborated with, researchers at the University of York.

The initial pearls aimed to include the following types of publication:

1) Literature on the methods for deriving weights;
2) Applied elicitation exercises in HTA to understand how methods described in elicitation literature are applied in practice, and identify practical challenges and limitations associated with existing methods; and
3) Studies comparing and evaluating existing methods for deriving weights.

2.2.1.2. Backwards searching

The backward search was performed by extracting references from relevant areas of the text only. For example, Bojke et al. (2017) reviewed existing literature on all aspects of elicitation in the context of cost-effectiveness modelling, but only references in sections 4.1 (Synthesising Multiple Elicited Beliefs), 4.3 (Combining Probability Distributions) and 5 (Assessing the Elicitation Process) were considered for inclusion. Similarly, the systematic review of elicitation methods in CEDM by Soares et al. (2018) was used to identify applied elicitation exercises in the field, and only the studies that reported their pooling methods were shortlisted. The extracted references were then reviewed for relevance by reading the full text before continuing with the literature search.

2.2.1.3. Forward searching

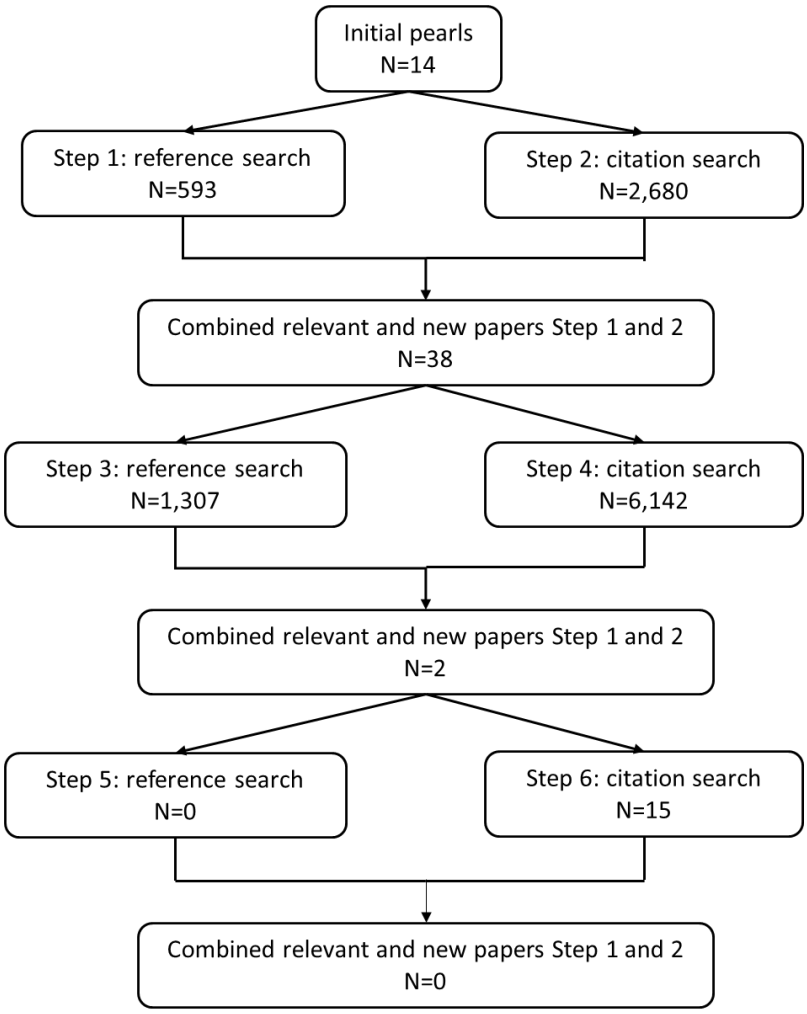
The forward search was performed by searching Google Scholar for citations that cited the initial pearls. Google Scholar was chosen based on the results of an informal search performed on three different search databases recommended by the University of York: Google Scholar, Web of Knowledge, and Scopus, where the former led to the most relevant citations and the least repetition. The identified citations were scanned for relevance in three stages: 1) by title

where only those that explicitly reported opinion pooling methods and applied examples were shortlisted; 2) by abstract; and 3) by full text.

2.2.1.4. Results

Fourteen initial pearls were used in the review, and three rounds of bidirectional searches were performed before saturation was reached. In total 55 citations were identified; the number identified in each round is shown in Figure 2.1.

Figure 2.1. Stepwise results of the BCSC.



The identified citations were classified into four categories:

1) Methods for deriving weights;
2) Applied exercises demonstrating the use of methods for deriving weights;
3) Papers reviewing the methods for expert elicitation that commented on the use of methods for deriving weights; and
4) Papers that evaluated, compared or critiqued the existing methods.

Two references spanned over two categories – they were applied exercises in HTA that also proposed a new weighting method (Bojke *et al.*, 2010; Shabaruddin *et al.*, 2010). Table 2.1 shows the number of references in each of the stated categories.

Table 2.1. References identified in the BCSC.

	N*	References
Methods for deriving weights	12	(Cooke, 1991) (Bojke *et al.*, 2010) (Shabaruddin *et al.*, 2010) (Brier, 1950) (Hallenbeck, 1920) (Murphy and Murphy, 1973) (Yates and F., 1994) (Epstein, 1969) (Murphy, 1970) (Murphy, 1971) (Brockhoff, 1975) (Degroot, 1974)
Applied elicitation exercises	14	(Soares, Dumville and Ashby, 2013) (Bojke *et al.*, 2010) (Speight *et al.*, 2006) (Haakma *et al.*, 2014) (Leal *et al.*, 2007) (D Sperber *et al.*, 2013) (Mckenna *et al.*, 2009) (Stevenson *et al.*, 2009) (Grigore, Peters and Hyde, 2016) (Fischer, Lewandowski and Janssen, 2013) (Shabaruddin *et al.*, 2010) (Chaloner *et al.*, 1993) (Hallenbeck, 1920) (Rakow *et al.*, 2005)
Reviews of existing methods	13	(Bojke *et al.*, 2017) (Grigore *et al.*, 2013) (Soares *et al.*, 2018) (O'Hagan *et al.*, 2006) (O'Hagan *et al.*, 2006) (Cooke, 2017) (Cooke, 2017) (Hartley and French, 2017) (Gosling, 2014) (Knol *et al.*, 2010) (P Garthwaite, J Kadane and O'Hagan, 2005) (Cooke and Goossens, 1999) (EFSA, 2014)
Evaluation, comparison and critique of existing methods	17	(Colson and Cooke, 2018) (Cooke, ElSaadany and Huang, 2008) (Bolger and Rowe, 2015) (Hammitt and Zhang, 2013) (Genest and McConway, 1990) (Cooke and Goossens, 2000) (Ferrell, 1985) (Clemen, 2008) (Shi-Woei Lin and Chih-Hsing Cheng, 2008) (Shi-Woei Lin and Cheng, 2009a) (Cooke, 2008) (Flandoli *et al.*, 2011) (Eggstaff, Mazzuchi and Sarkani, 2014) (Burgman *et al.*, 2011) (Brown and Aspinall, 2004) (Cooke and Goossens, 2006) (Aspinall and Cooke, 2013)

* The total number of references is higher than that reported in Figure 2.1 because two references were applied exercise in HTA that also proposed a new weighting method (Bojke *et al.*, 2010; Shabaruddin *et al.*, 2010).

The rest of section 2.2 reports the findings of the literature review. The methods for deriving weights (references in row 1 in Table 2.1) are described in section 2.2.2. References presented in row 2 and 3 in Table 2.1 are used along the way to demonstrate the application of each method. Then, section 2.2.3 summarises the findings and outlines the structure of the discussion. Literature evaluating, comparing and critiquing the existing methods (shown in the row 4 in Table 2.1) is discussed later in the chapter (in section 2.4), where methods for deriving weights are evaluated and compared.

2.2.2. Methods for deriving weights in opinion pooling

The literature review identified two general approaches for deriving weights: 1) based on experts' observed characteristics, and 2) based on experts' elicitation performance. Sections 2.2.2.1 and 2.2.2.2 describe each approach in turn.

2.2.2.1. Deriving weights from observed characteristics

Weights can be derived from some measure of professional status, seniority, education level or historical track record. For example, Haakma et al. (2014) asked experienced radiologists to rank tumour characteristics on their importance in detecting malignancies. The weights for radiologists were based on their experience:

- **45%** of the score was determined by their length of experience in the field (score 1 if <3 years or 2 if three years or more);
- **45%** of the score was determined by the average number of MRI images they see per week (score 1 if <5 MRIs per week, score 2 if 5-10 MRIs per week and score 3 if more than 10 MRIs per week); and
- **10%** of the score was determined by their experience in using MRI scans in other areas (score 1 if no, or 2 if yes).

Similarly, Shabaruddin et al. (2010) scored experts according to the number of patients they prescribe the treatment under consideration.

In both studies, the cut-off for each category was determined by the investigator and it is not clear whether the relationship between characteristics and weights is based on evidence or chosen arbitrarily.

Examples of characteristics used to derive weights in fields other than HTA include weights derived from experts' self-rated expertise on a scale of 1-7 (Brockhoff, 1975), and citation counts (Cooke, ElSaadany and Huang, 2008).

2.2.2.2. Deriving weights from measured performance

Weights can be derived by eliciting experts' beliefs about the value of one or more parameters unknown to them but known to the investigator, and comparing their priors to the observed value of that parameter once it becomes available (Cooke, 1991). The parameters used to measure experts' performances are referred to as 'seeds'. The consistency between experts' priors and observed values of the seeds can then be assigned a numerical value, or a 'score', which is then used to weight them.

Weighting experts by their performance in elicitation requires the investigator to decide on the seed parameter, the method used to score the discrepancy between experts' priors and the observed value of the seed, and the methods used to derive weights from these scores. Here, each is described in detail.

Choosing the seed parameters

Cooke (1991) argues that numerical scores should capture those skills that a good expert is expected to possess; however, he does not detail what skills this should entail.

Seed questions can be either predictions or retrodictions, and either domain or adjacent (Cooke, 2017). Predictions are questions about future quantities not known at the time of the elicitation that are observed within the timeframe of the study, while retrodictions refer to seeds based on previously collected data. Domain seeds refer to those within the same field of expertise as the target parameter, while adjacent seeds are related but not identical to the target. The author argues that domain predictions are the preferred type of seed, followed by domain retrodictions and adjacent predictions, and that adjacent retrodictions are the least desirable, although it is emphasised that the recommendation is based on practical experience rather than empirical evidence. Published estimates are thought to be inappropriate because the seed does not aim to capture their recall or subject specific knowledge, but their ability to assess uncertainty (Cooke, 2017).

According to Cooke (2017), the types of parameters that have been used as seeds in the past include the following:

- results of measurements performed within the study's timeframe (e.g. trial results);
- existing but unpublished measurement results;
- unfamiliar features of standard datasets; and
- combining and comparing values from disparate data sets.

The effectiveness of different types of seeds in capturing experts' ability to judge uncertainty is not clear.

In order to gain insight into the types of seeds used in HTA, applied exercises were extracted from the systematic review of reported practice of expert elicitation in HTA conducted by Grigore et al. (2013) and later updated by Soares et al. (2018). The review by Soares et al. (2018) identified five studies that explored the use of measured performance to weight experts; these are shown in Table 2.2. Out of the five studies one did not report the number of seeds, and the remaining four used between one and eight seeds (mean=3.75). The authors did not report how the seeds were selected, although all seeds were related to the target parameter for which elicitation was conducted.

Table 2.2. Seed and target parameters elicited in HTA context. NR=not reported.

Study	Number of seeds	Seed parameters	Number of target parameters	Target parameters
Bojke et al. (2010)	2	Expected response to three months of treatment for two treatments – infliximab and etanercept	4	Rate of progression while responding to treatment and after treatment failure – for both treatments (infliximab and etanercept)
Soares et al. (2011)	4	NR	10	Ten parameters relating to the effectiveness of different treatments for severe pressure ulceration
Fischer et al. (2013)	8	Data published in the literature (specific questions not reported)	15	'Uncertain parameters needed for [the] model for optimal treatment of haemophilia patients'
Sperber et al. (2013)	NR	NR	1	The percentage of patients who develop obstructive sleep apnoea within the first year after the onset of tetraplegia
Grigore et al. (2016)	1	The proportion of patients undergoing relevant treatment who experience clinically significant complications as a result of testosterone flare	1	The proportion of patients who experience spinal cord compression as result of testosterone flare, also experience paraplegia

Scoring methods

Three methods for deriving performance-based weights were identified from the available literature: Cooke's Classical Model (Cooke, 1991), Bojke's Best Estimate Fractions (Bojke *et al.*, 2010), and Budescu and Chen's (2015) Contribution-Weighted Model. All three methods differ in how they score experts' priors: the Classical Model uses Kullbeck-Liebler divergence and the Shannon relative information, the Contribution-Weighted Model uses Brier's probability scores, and the Best Estimate Fractions are based on absolute difference. This section describes each scoring method in turn. Furthermore, three additional scoring methods were identified that have been used to study the accuracy of probabilistic judgments in elicitation. The methods include Decomposition of Brier's probability score (Yates and F., 1994), Ranked Probability Score (Epstein, 1969), and Confidence Interval Probability Coverage (Murphy and Winkler, 1977). While these methods have not been explicitly used to weight experts, their role is to measure the closeness between experts' priors and observed values of the seed, and so they are also described in this section, and their role as potential methods for deriving weights is discussed later in this chapter, in section 2.4.3.

The Classical Model: Kullbeck Liebler divergence and Shannon's relative information

The Classical Model, introduced by Cooke (1991), was the first available method for deriving weights based on experts' elicitation performance. The weights are determined by two aspects of experts' performance: calibration (accuracy) and information (uncertainty).

The calibration score represents the closeness between experts' priors and observed parameter values (Cooke, 1991). An expert is thought to be well calibrated if the discrepancy between their probability distribution P and the observed probability distribution S is no greater than the discrepancy between two independent variables with distribution P.

The discrepancy between P and S is measured using Kullback–Leibler divergence, as shown in Equation 2.1 (Cooke, 1991).

$$I(S,P) = \sum_{i=1}^{M} S(i) ln \frac{S(i)}{P(i)} \qquad \textbf{Equation 2.1}$$

Where i is one of M outcomes;

$S(i)$ is the observed probability of i; and

$P(i)$ is the expert's probability of i.

In practice, the Classical Model involves eliciting five quantiles (5th, 25th, 50th, 75th and 95th percentiles) for a range of seed parameters, then observing the frequency with which the observed parameter values fall into each quantile (Quigley *et al.*, 2017). For example, if an expert is perfectly calibrated, then 5% of seeds would have a value lower than experts' stated 5th percentile, 20% of the seeds would have a value between 5th and 25th percentile, etc. An example is shown in Table 2.3. In the table, each row represents expert's assessment of one seed. Five different quantities are elicited for each seed – the values of the 5th, 25th, 50th, 75th and 95th percentiles – splitting the distribution into six intervals (0-5%, 5-25%, 25-50%, 50-75%, 75-95%, 95-100%). Each interval represents an outcome i, and so M=6. An expert's probability P that the seed value will be in each of the six intervals is (0.05, 0.2, 0.25, 0.25, 0.2, 0.05). Column 7 shows the observed value of that seed, while column 8 shows the interval in experts' priors where the observed value of the seed lies. The proportion of observed seed values S that fall within each interval is (0.2, 0.3, 0.2, 0.1, 0.2, 0). The experts' discrepancy score is 0.262[2].

The higher the $I(S,P)$ the greater the discrepancy between S and P (Cooke, 1991).

Cooke's calibration score is then the probability that the discrepancy between observed and subjective probability distributions $I(S,P)$ is due to sampling variation (see Equation 2.2) (Cooke, 1991).

$$Cooke's\ calibration\ score\ =\ Prob\{I(S',P) \geq\ I(S,P)|Cal(P),n \qquad \textbf{Equation 2.2}$$

Where $Cal(P)$ is the hypothesis that the elicited priors and observed probability distributions are independent and identically distributed with distribution P; and

n = number of observations, or seeds.

As the number of observations gets large, 2nI(S,P) follows a Chi-squared distribution with M-1 degrees of freedom (Cooke, 1991).

[2] $I(S,P)$ = 0.2*ln(0.2/0.05) + 0.3*ln(0.3/0.2) + 0.2*ln(0.2/0.25) +0.1*ln(0.1/0.25) + 0.2*ln(0.2/0.2) + 0*ln(0 /0.05)

Table 2.3. Hypothetical example of an expert's priors of ten seeds and observed values of those seeds used to derive Cooke's calibration score.

Seed	Elicited percentiles					Observed value	Experts' percentile interval in which the observed value of the seed lies
	5th	25th	50th	75th	95th		
1	1	3	5	6	8	4	25th-50th
2	7	12	14	16	20	8	5th-25th
3	1	4	6	8	10	7	50th-75th
4	5	6	7	9	12	5	5th-25th
5	3	7	9	10	15	8	25th-50th
6	3	6	7	8	9	5	5th-25th
7	4	9	11	12	15	3	5th-25th
8	1	2	3	4	6	5	75th-95th
9	6	8	10	12	17	7	5th-25th
10	1	3	4	5	7	6	75th-95th

Cooke (1991) argues that, given two identical calibration scores, the expert who expresses greater certainty should be assigned a higher (better) score, and proposes information, or entropy, as a measure of the spread of probability distributions. Cooke's information scores are derived relative to some other distribution referred to as 'the background'. The background is usually a uniform or log-uniform distribution and its range is the smallest possible range that contains all values believed to be plausible by the experts. For example, if priors are elicited from two experts – one who believes that the plausible range is 2-4 and the other believes that the range is 3-6 – then the background distribution would be a uniform distribution with a range of 2-6.

Cooke's information score is derived using Equation 2.3, based on Shannon's relative information (Cooke, 1991).

$$Cooke's\ information\ score = \frac{1}{N}\sum_{i=1}^{N} p_i \ln\left(\frac{p_i}{s_i}\right) \qquad \textbf{Equation 2.3}$$

N is the number of i outcomes;

p_i is experts' probability of outcome i; and

s_i is the background probability of outcome i.

The lower the score, the more uncertain the expert is. The minimum score is 0, when p_i = 1/n for all i. The maximum score depends on the range of the background distribution.

Information does not take into account the value the experts' distributions take, nor the observed value of the seed. This means that an expert who is completely certain about a wrong value would score a high (good) information score.

Cooke combines experts' calibration and information to weight experts – the methods for deriving weights are described later in this chapter (section 'Methods for converting scores into weights').

Best Estimate Fractions derived from absolute difference

Bojke et al. (2010) used the absolute difference to derive Best Estimate Fractions. The authors used continuous variables as seeds, and sampled from experts' priors and the observed the probability distribution of the seed (taking into account its parametric uncertainty) and derived the absolute difference between them. For each iteration the authors assigned one point to one expert whose value was closest to the observed value (i.e. the expert who minimised the absolute difference), while the remaining experts were assigned zero points. Overall scores were then derived by adding up the total score for each expert and dividing it by the total number of iterations. An example is presented in Table 2.4.

Table 2.4. Hypothetical example of Best Estimate Fraction scores for two experts, derived from four random samples drawn from their priors.

Iteration	Sample from parameter distribution	Expert 1		Expert 2	
		Sample from experts' prior	Score	Sample from experts' prior	Score
1	3	2	1	6	0
2	5	1	0	7	1
3	5	3	0	6	1
4	4	2	0	5	1
Overall score		0.25		0.75	

Brier's probability score

Probability scores were introduced by Brier (1950) to measure accuracy of meteorology forecasts. Brier's Probability Score (PS) is given in Equation 2.4.

$$\mathrm{PS} = \sum_{j=1}^{n} (p_j - d_j)^2 \qquad \textbf{Equation 2.4}$$

Where j is one of n possible outcomes;
p_j= expert's prior that outcome j will occur, such that $\sum p_j = 1$; and
$d_j$is a parameter that takes value of 1 if outcome j does occur and 0 if it does not, such that $\sum d_j = 1$, and d_j=1 for exactly one j.

For example, if the seed question is '*What number of days it will rain next week?*', there are eight possible outcomes (0-7 rainy days) and so n=8 and $j \in \{0, 1, 2, 3, 4, 5, 6, 7\}$. If rain is observed on four days, then d_j=0 for all $j \in \{0, 1, 2, 3, 5, 6, 7\}$, and d_4 = 1. If an expert believes that the probability of rain on exactly (0, 1, 2, 3, 4, 5, 6, 7) days is (0.05, 0.1, 0.2, 0.3, 0.2, 0.15, 0, 0), then their score is 0.64[3].

Deriving experts' scores on the basis of their performance against a single seed question can be unreliable. If an event results in a rare outcome (d_j=1 but p_j is small), then the expert who assigns a low probability to outcome j will have a bad (high) score, despite potentially being right (that the probability of the outcome was low). Thus when Brier's method is used to score experts, multiple seed questions are asked and experts are scored using mean PS ($\overline{PS}$) (see Equation 2.5).

$$\overline{PS} = \frac{1}{n} \sum PS \qquad \textbf{Equation 2.5}$$

Where n is the number of seed questions.

Brier's PS does not take into account the value of the parameter (O'Hagan *et al.*, 2006). If two experts are asked how many days in a week they believe will rain, one expert may believe that the probability of rain on (0, 1, 2, 3, 4, 5, 6, 7) days is (0.05, 0.1, 0.2, 0.3, 0.2, 0.15, 0, 0), while another expert believes that it is (0.05, 0.3, 0.2, 0.2, 0.15, 0.1, 0, 0). If rain is then observed on exactly one day both experts would have the same score, although the second expert placed a higher probability on the values closer to the observed number of

[3] $(0.05-0)^2 + (0.1-0)^2 + (0.2-0)^2 + (0.3-0)^2 + (0.2-1)^2 + (0.15-0)^2 + (0-0)^2 + (0-0)^2 = 0.64$

rainy days, and thus may be intuitively regarded as the more accurate of the two. Brier's PS thus may fail to take into account how wrong experts are when elicited parameters are continuous, ordinal or discrete interval variables where magnitude of 'error' is relevant.

Brier's scores have been used to weight experts in the Contribution-Weighted Model – the model is described in detail later in this chapter (in the section 'Methods for converting scores into weights').

Decomposition of Brier's score

Several researchers have decomposed Briers' PS into multiple components to understand what drives the score. For example, Murphy (1973) decomposed the PS into calibration, variance and resolution, where each component is assigned an individual score. Calibration captures the difference between experts' priors on the probability of outcomes and their observed relative frequencies. Variance indicates how the relative frequency of outcomes influences the PS, considering that outcomes with relative frequency of 0.5 result in lower (better) scores than those that occur with probabilities closer to 0 or 1. Resolution shows how well experts distinguish between outcomes with a high probability and outcomes with a low probability, so that when the calibration score is low (good), the resolution is also good, whereas good resolution alone could indicate that the expert places a high probability on outcomes with a low relative frequency and vice versa. Yates (1994) decomposed the PS into bias, slope and scatter, where bias measures experts' tendency to be consistently higher or lower than the observed relative frequency, slope measures a combination of calibration and resolution, and scatter measures the random error in experts' subjective probabilities.

The final scores derived using these methods are mathematically equivalent to Brier's PS (PS = calibration + variance – resolution = bias + slope + scatter), and so, when scores are used to weight experts, they lead to identical weights.

Ranked probability score

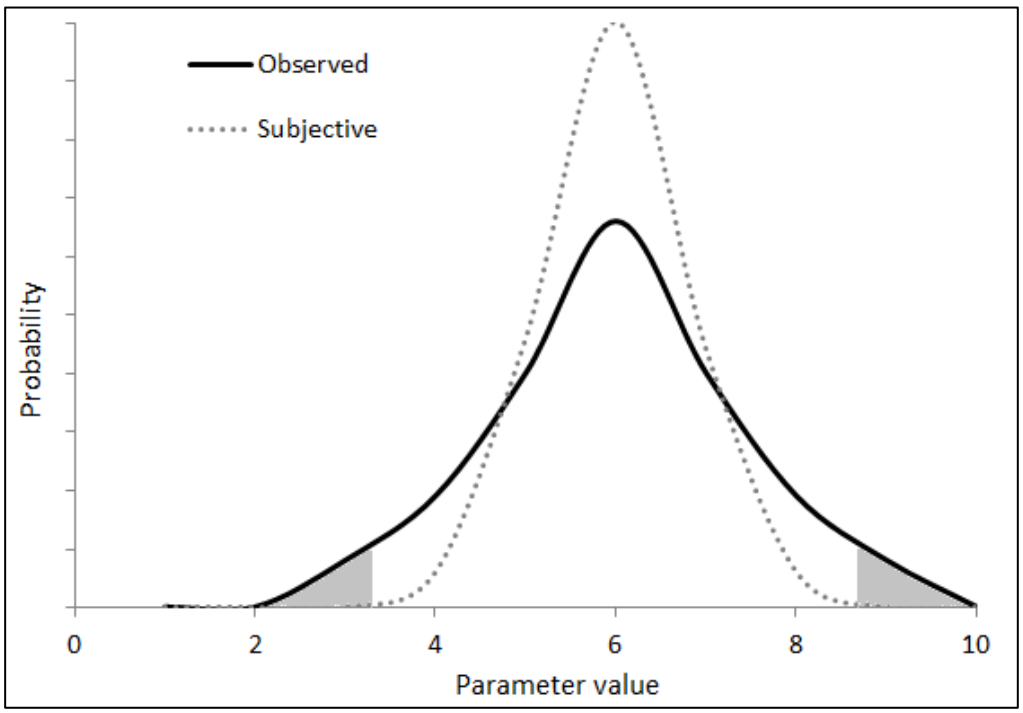
Ranked probability scores have been developed by Epstein (1969) to take into account the distance from true value when scoring experts on ordinal variables. The method was developed to score meteorological forecasts and assumes that some outcomes are better than others. For example, if there are four outcomes i (i=1,2,3,4), where outcome 1 has the least costly consequences and outcome 4 has the most costly consequences, predicting outcome 1 when 4 occurs is worse than predicting outcome 4 when outcome 1 occurs. The

equation for deriving ranked probability scores is specific to Epstein's (1969) weather example; applying the scoring method in HTA would require deriving a new equation specific to that example, taking into account the number of possible outcomes and the consequences of each of those outcomes.

Confidence interval coverage probability

The confidence interval coverage probability (CICP) measures the proportion of the time that the interval contains the 'true value of interest' (Murphy and Winkler, 1977). Murphy and Winkler (1977) have used it to score experts' predictions of temperature forecasts by eliciting priors on temperature predictions and calculating the proportion of observed temperatures which were within the elicited distributions. Uncertainty around the true value of the parameter can be taken into account by sampling from the probability distribution and deriving the proportion of random samples which fall within an experts' prior (see the shaded area in Figure 2.2). Experts are penalised for placing zero probability on observed values (i.e. for being overconfident).

Figure 2.2. Probability distributions derived from an observed sample and elicited from an expert. The highlighted area represents the observed values not included in the prior. The CICP is 1-the highlighted proportion of the observed probability distribution.



This literature review did not identify any examples where CICP scores were used to weight experts; their only application has been to compare the accuracy of priors elicited from different experts.

Methods for converting scores into weights

In the last section it was highlighted that three methods for deriving performance-based weights were identified from the available literature: Cooke's Classical Model (Cooke, 1991), Bojke's Best Estimate Fractions (Bojke *et al.*, 2010), and the Contribution-Weighted Model (Budescu and Chen, 2015), and the methods used to score experts' priors were described. This section describes how each method uses scores to derive weights.

The Classical Model

The previous section shoes how the Classical model derives calibration and information scores from experts' priors. The model then combines the two scores in order to derive weights, using Equation 2.6 (Cooke, 1991).

$$w_\alpha = Ind_\alpha \text{ (calibration score) x calibration score x information score} \qquad \textbf{Equation 2.6}$$

Where Ind_α is an indicator function such that $Ind_\alpha(x) = 0$ if $x<\alpha$ and $Ind_\alpha(x) = 1$ otherwise; and α is a threshold value an expert must score above in calibration, in order to have a non-zero score overall.

Ind_α was introduced to ensure that very high information cannot compensate for poor accuracy (calibration). It is not clear what the value of α should be; the author often uses α that maximises the weighted score of all experts combined (Colson and Cooke, 2018). Cooke argues that experts who have a score of 0 are not considered to be bad, or irrelevant, but their marginal contribution to the aggregate probability distribution when priors from all experts are combined is zero (Goossens, 2008a).

Contribution-Weighted Model

Budescu and Chen (2015) developed the Contribution-Weighted Model, which aims to take into account experts' performance relative to the crowd by capturing the effect of inclusion (or exclusion) of each expert in the sample. The method was designed for categorical variables with R outcomes, and used Brier's Probability Score (Brier, 1950) (see Equation 2.4) to value their accuracy.

To derive experts' weights, first the aggregate score S for event i is calculated using Equation 2.7 (Budescu and Chen, 2015).

$$Si = a + b\sum_{r=1}^{R_i} PS_{ir} \qquad \textbf{Equation 2.7}$$

Where PS is Brier's probability score;

i is one of N events;

r is one of R outcomes for event i (for a binary event R=2);

o_{ir} is the binary indicator of outcome r for event i (1=occur, 0=not occur);

m_{ir} is experts' average probability assigned to outcome r, of event I; and

a and b are constants introduced by Budescu and Chen (2015) used to scale experts' scores. The authors set a = 100 and b=-50 to yield scores ranging from 0 to 100, where a score of 100 indicates that all experts assigned probability 1 on every outcome r that occurred (m_{ir}=1 on all outcomes r for which o_{ir} =1) and zero to all other outcomes (m_{ir}=0 for all outcomes r for which o_{ir} =0), while a score of 0 indicates the opposite.

Budescu and Chen (2015) derived weights from experts' scores by measuring their contribution. The contribution C of each judge *j* is derived using Equation 2.8.

$$C_j = \sum_{i=1}^{N_j}(S_i - S_i^{-j})/N_j \qquad \textbf{Equation 2.8}$$

where *j* is one of N_j experts;

S_i^{-j} is the score of the unweighted aggregate prior for event i calculated when expert j's prior is removed from the sample; and

N_j is the number of seeds for which expert *j*'s beliefs were elicited. The term was introduced to allow for the possibility that, for some seeds, not all experts' beliefs are elicited.

The contribution, C_j, can take any value between -100 and 100, where positive values indicate that the experts' prior on average improved the crowd's S, while negative values suggest that the experts' prior reduces the score S of the crowd.

Best Estimate Fraction

As described in the section on scoring methods, Best Estimate Fractions are derived by calculating the proportion of random samples from each expert's prior that minimised the distance from the observed value of the seed in comparison to other experts (Bojke *et al.*, 2010). The scores from different experts summed to 1, and were then used directly to derive weights.

2.2.3. Summary of findings

Overall, the literature review described in this chapter has identified two approaches to deriving weights (based on experts' observed characteristics and their measured performance in elicitation) and multiple methods within each approach.

The aim of this chapter is to derive guiding principles for choosing between the various options for deriving weights. To achieve this, factors that could affect experts' contribution in elicitation are first identified. Section 2.3 revisits the aims of elicitation proposed in Chapter 1, then proposes and discusses factors that could affect experts' contribution towards achieving those aims. Section 2.4 discusses the assumptions, advantages and limitations of different methods for deriving weights in relation to the discussion provided in section 2.3. Section 2.5 summarises the findings from the chapter and discusses the role of each method in opinion pooling. The section concludes with highlighting research gaps and outlining the structure of the rest of the thesis.

2.3. Defining the role of weighting in elicitation

Chapter 1 proposed that elicitation can be used to characterise uncertainty in CEDM when other sources of information are unavailable, unattainable within the existing resource constraints, or are of uncertain generalisability. The aim of an elicitation exercise is to capture the current state of knowledge around uncertain quantities. Structured elicitation processes are used to ensure that experts base their priors on all available information to avoid bias due to inaccurate or incomplete information and to minimise uncertainty, and to help experts assess uncertainty in their beliefs and express it probabilistically, free from bias.

The different steps in the elicitation process used to elicit informed and unbiased priors were discussed in detail in Chapter 1, and a summary of this is shown in Table 2.5.

Table 2.5. Elicitation process steps taken to ensure experts use all available information, and assess and express their uncertainty free from bias.

Ensure complete information	Help experts assess the information
• Recruit substantive experts • Recruit multiple experts • Provide background information • Encourage information sharing	• Recruit impartial experts • Plan the elicitation process (choice of parameter, quantity, elicitation technique, delivery method) • Training and debiasing • Evaluation and feedback

However, methodological uncertainties and logistical challenges mean that it is not always clear how to achieve these. Table 2.6 provides examples of such challenges.

Differential weighting can potentially compensate for methodological challenges by giving 'more say' to experts who are believed to be less affected. For example, if information sharing is not possible, experts whose experience is more closely related to the target parameter can be assigned a greater weight.

In this chapter, four factors were identified to give a potential basis for differential weighting:

- substantive expertise;
- perspective;
- normative expertise; and
- the ability to make accurate probabilistic assessments independent of knowledge, perspective and normative expertise.

Table 2.6. Examples of methodological and logistical challenges in the elicitation process.

Step in elicitation process	Potential challenges in delivering the step
Recruit substantive experts	• It is unclear who is a substantive expert
Recruit impartial experts	• Impartiality is difficult to ascertain
Recruit multiple experts to represent the expert community	• Composition of the expert community unclear • Unable to recruit representative sample
Provide background information	• Background information not read by experts • Experts have difficulty interpreting the background information
Encourage information sharing	• Information sharing not possible because it is not possible to gather all experts simultaneously
Planning	• Methodological uncertainties around what to elicit, how to elicit and how to deliver the exercise to minimise bias
Training and debiasing	• Optimal method for training and debiasing is unclear • Experienced trainers may not be available • Unable to demonstrate effectiveness
Evaluation and feedback	• Can induce bias • Unclear if it leads to 'improvement'

Sections 2.3.1-2.3.4 define and provide justification for each of the proposed factors in turn. Section 2.3.5 then discusses what determines the importance of each of the four factors.

2.3.1. Substantive expertise

Chapter 1 established that experts' **substantive expertise** – expertise in the field for which elicitation is being conducted – can affect the outcomes of an elicitation exercise by ensuring that experts are not basing their priors on inaccurate information, and that uncertainty is not overestimated. A perfectly unbiased expert with no field specific experience could give unbiased quantities by expressing complete uncertainty, i.e. assigning equal probability to all possible outcomes. However, asking an expert with clinical (substantive) experience within the field may resolve some of that uncertainty by narrowing down the range of plausible values of the elicited parameter. If the aim of the

elicitation exercise is to capture the current state of knowledge based on limited evidence, it could be argued that asking an uninformed (inexperienced) expert, where a more experienced one exists, overestimates uncertainty in the model and the value of further research.[4]

If in a sample of experts some are more substantive than others, differential weighting can potentially ensure that experts who are less likely to base their priors on inaccurate beliefs are given more say.

2.3.2. Perspective

Perspective refers to an experts' point of view in relation to the target parameter. Perspective can vary between experts with different professions (for example, an epidemiologist may have a different perspective to a clinician) or between experts with the same profession but different settings (such as primary and secondary care nurses) (Soares *et al.*, 2011).

Perspective is of particular importance if it creates the possibility of motivational bias. Garthwaite et al. (2005) highlighted how perspective can induce partiality (even without an obvious incentive such as financial gain from a particular outcome) using an example of a radiation expert whose beliefs are elicited about the health effects of the radiation release at Chernobyl. The relevant individual is likely to have spent years becoming an expert and their pay off (in terms of social attention and grant funding) is likely to depend on the societal perception of the importance of radiation. As a result, the expert may be incentivised to accentuate the dangers of radiation.

Weights can be based on experts' perspective to ensure that the expert sample is representative of the expert community, as recommended in the elicitation literature (Kadane, 1986). For example, a general practitioner (GP) and a geriatrician can both be experienced in treating patients who have suffered a fall, but their experience could differ in the types of patients they treat. The GP is likely to be more familiar with patients who have suffered mild falls, whereas the geriatrician is likely to be more familiar with falls that

[4] It is important to note that the above example only suggests that the choice of experts can affect results of elicitation. It does not imply that experts who express more confidence (narrow confidence intervals) always lead to the best representation of uncertainty, as experts can be certain due to bias; for example, they may be overconfident due to irrational beliefs.

result in fractures and hospitalisation. If both perspectives are judged to be equally relevant in an elicitation exercise but the sample of experts contains more geriatricians than GPs, then assigning greater weights to the GPs could be used to ensure that experts with each of the two perspectives are given equal say.

It is important to note the similarity between perspective and substantive expertise – if a clinician is only experienced in treating one subgroup of patients they can be argued to be less substantive than a clinician with a more general patient population. In this chapter they are proposed to be separate concepts where substantive expertise is hierarchal, whereas perspective describes differences in experience that can affect experts' priors but where it is not clear which is better. For example, if two clinicians are compared who both have equal patient experience, but one is also an experienced researcher, than the researcher can be argued to be more substantive. If, however, we compare two experts where one has more patient contact but the other has more research experience, they are said to have different perspectives as it is not clear which type of experience is 'better'.

2.3.3. Normative expertise

Normative expertise refers to the ability of experts to express their beliefs in the required format, usually quantitatively (Ferrell, 1994; Stern and Fineberg, 1996). As discussed in Chapter 1, statistically incoherent or inconsistent priors are used as indicators of a lack of normative expertise. For example, if an expert knows that probabilities of mutually exclusive outcomes must add up to one, but their priors suggest otherwise, then it is likely that their priors do not represent their beliefs accurately.

Incoherent and inaccurate priors can be identified and corrected during the evaluation and feedback steps in the elicitation process; however, when elicitation is delivered remotely, as is the case in the majority of exercises in HTA (Soares *et al.*, 2018), evaluation tends to be carried out after the elicitation exercise and the findings are not fed back to the expert for clarification.

If subsequent priors elicited from some experts are believed not to represent their beliefs accurately, it may be desirable to assign those experts lower weights.

2.3.4. Ability to make accurate probabilistic assessments

In elicitation, experts are required to process available information to assess the plausible values of a parameter and uncertainty around it, free from bias. For example, assessing

what proportion of patients who take a particular medicine will experience side effects requires consideration of how many patients they observe in clinical practice, the proportion of their patients that have the condition, the proportion of patients who receive the relevant treatment, the proportion that report side effects and the proportion who may be experiencing side effects but do not report these to their clinician. Without careful consideration of each of these factors, a clinician could be susceptible to cognitive bias by basing their response on whether they can easily remember a patient who experienced side effects with the medicine in question.

Assessment of the plausible values of a parameter are very different tasks to those generally involved in clinical work, such as keeping up to date with relevant literature, knowledge of care pathways, and communication skills. Not all individuals with substantive expertise can make accurate probabilistic assessments in that field; indeed there are several examples of studies where clinicians, with substantive skills, made inaccurate probabilistic assessments (Hoffmann and Del Mar, 2017).

Tetlock et al. (2016) studied the ability of lay people to assign probabilities to future (unobserved) outcomes of events across a range of topics (including politics, finance, entertainment and sports) and scored their performance using Brier's probability score (see Equation 2.4 in section 2.2.2). The authors found that some individuals were consistently better at predicting future outcomes than professionals who worked in those fields. These individuals, named 'superforecasters' by the authors, based their predictions on publicly available resources and were 30% more accurate than intelligence officers with access to classified information, suggesting that substantive expertise may not be crucial for making accurate probabilistic assessments; it is instead about how the available information is utilised in generating priors.

If some experts in the sample are believed to be less susceptible to cognitive bias and more accurate at making probabilistic assessments of parameter values (and their uncertainty around them), then differential weighting could potentially minimise the bias in the aggregate prior by giving more say to those experts.

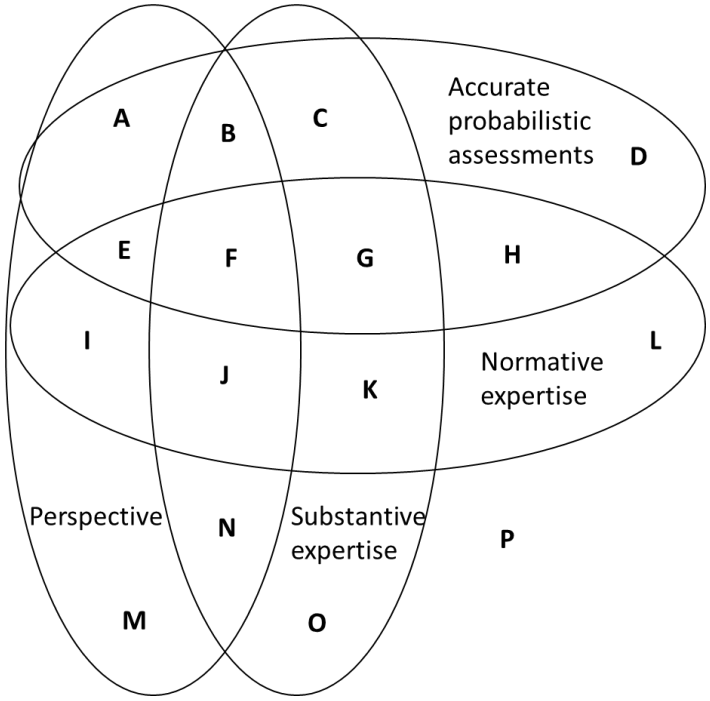
2.3.5. Which factors should the weights reflect?

The last section identified four factors that can affect uncertainty and bias in experts' priors. The different factors are likely to vary between experts, as shown in Figure 2.3. For example, experts in sections A, E, I and M are not considered substantive experts but they

are familiar with the parameter of interest – this could include junior doctors working with the patient population targeted by a new treatment for which elicitation is being conducted. In contrast, experts in sections C, G, K and O are substantive but their perspective can bias their prior on the target parameter – they could be experienced clinicians who work with a specific patient subgroup. Experts in areas B, F, J and N are both substantive and have a complete perspective of the target parameter.

The importance of each of the four factors is likely to depend on the expert and the elicitation process. If an expert has a biased perspective (i.e. they only see one specific subgroup of patients) then their contribution can depend on their ability to extrapolate their knowledge (from their perspective) to make accurate probabilistic assessments. The role of their perspective can also depend on the background information provided by the investigator. If experts are presented with information on how their experience compares to that of other experts, then their priors are less likely to be biased by their perspective, and experts in areas F and G in Figure 2.3 could be considered to be 'equally good'. In contrast, if they are unaware that their patient population is different to the general population, then perspective may affect their priors more, and so priors elicited from experts in area F in Figure 2.3 could be less biased than those elicited from experts in area G.

Figure 2.3. Interaction between factors that affect experts' priors.



The next two sections discuss each of the two approaches to deriving weights, highlighting which of the four factors they capture and by what means, and how this affects their role in the elicitation process.

2.4. Evaluation and comparison methods for deriving weights

In what has been covered so far, it is clear that the methods for deriving weights differ in the way they measure experts' contribution. Weights can be based on observed characteristics or on the accuracy of their assessment of seed parameters, and multiple methods exist within each approach. Methods for deriving weights from experts' characteristics vary in the characteristics used as a basis for deriving weights and how those are scored. Methods for deriving weights from experts' elicitation performance vary in the choice of seed, the method used to score experts' performance, and the methods used to derive weights from those scores.

This section aims to evaluate and compare the existing methods. First, section 2.4.1 describes the existing literature; sections 2.4.2 and 2.4.3 further explore how the method for deriving weights can affect their role in elicitation.

2.4.1. What the literature says

This section describes the literature that aims to evaluate and compare methods for deriving weights, identified in section 2.2.1 (row 4 in Table 2.).

Characteristic-based weights have been evaluated by comparing the accuracy of characteristic-weighted aggregate priors on a seed parameter to equally weighted priors on the same seed (Cooke, ElSaadany and Huang, 2008).

Performance-based weights have been evaluated using two approaches: 1) an assessment of internal validity, and 2) an assessment of external validity.

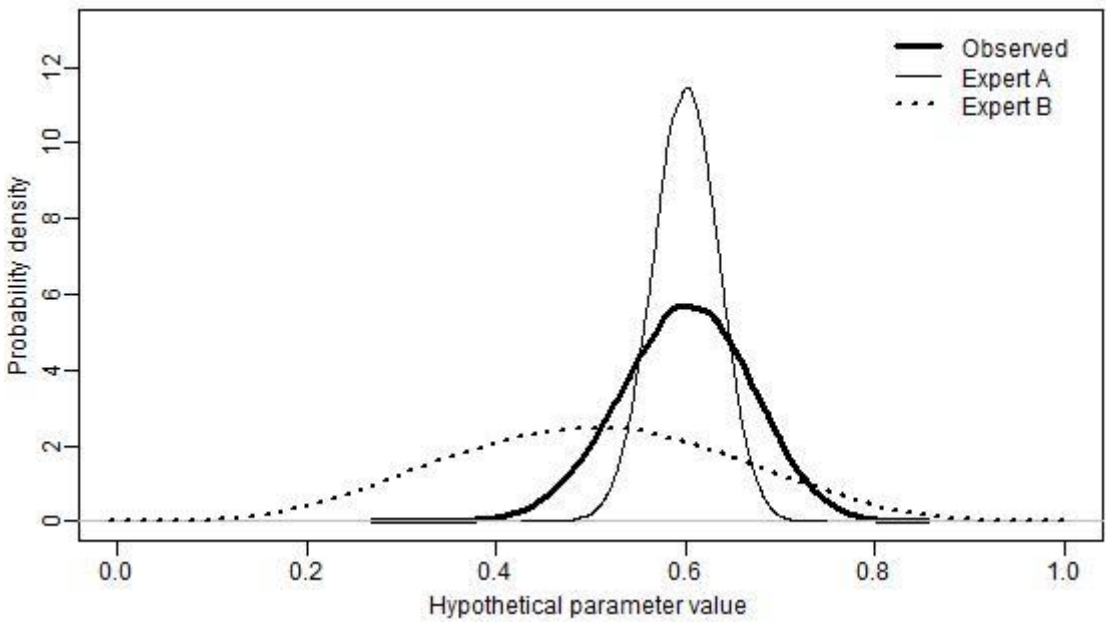
Internal validity refers to the ability of performance-based weights to improve the accuracy of the aggregate prior on the seed that was used to derive the weights. The purpose of assessing internal validity is to determine whether the derived weights effectively capture experts' contribution to the aggregate prior, as demonstrated in Figure 2.4. The figure compares priors elicited from two hypothetical experts to the observed probability

distribution. Intuitively, Expert A is more knowledgeable than Expert B. Expert A is also, overconfident and so placing a very low weight on Experts B could underestimate uncertainty of the aggregate prior, making it less accurate than an equally weighted prior (Goossens, 2008a).

External validity refers to the ability of a weighting method to improve the accuracy of the aggregate priors on the target parameter (Colson and Cooke, 2017). Since the value of the target parameter is rarely known (that is why elicitation is conducted), various methods for assessing external validity have been developed.

Clemen (2008) proposed the remove-one-at-a-time (ROAT) method as a measure for out of sample validity of the Classical Model. ROAT involves the use of datasets with priors elicited from multiple experts on multiple seeds. One seed question is removed at a time and the score is recalculated for the rest of the sample. Then, the derived scores are used to weight experts' priors on the one seed that was removed. The accuracy of the aggregate priors is then compared to the accuracy of the prior derived by equal weighting.

Figure 2.4. Beliefs elicited from two experts about the same seed and its observed probability distribution. Experts A is more accurate than expert B but they are overconfident so the aggregate priors should take both into account.



Cooke (2008) built on ROAT analysis by splitting each set of seeds into two halves, where one half served as the 'training' set (i.e. the seeds) and the other half as the 'test' set (i.e. as the target parameters). They then assessed whether weights derived from the training set led to more accurate aggregate priors on the test set than unweighted aggregate priors.

Sections 2.4.1.1 to 2.4.1.3 describe the findings of each approach to evaluating methods for deriving weights.

2.4.1.1. Evaluation of characteristic-based weights

Only one study that evaluated the characteristic weighted priors was identified. Cooke et al. (2008) used experts' citation count to derive weights, and compared citation-weighted aggregate priors to unweighted and performance-weighted ones. The authors found that the citation weighted priors were more accurate than unweighted, but less accurate than the performance-weighted priors.

2.4.1.2. Internal validity

Goossens and Cooke (2008a) analysed the internal validity of seeds for every study in the TU Delft database[5] published before 2006 (N=45). The authors compared the accuracy of aggregate priors derived using equal weights and performance-based weights (derived using the Classical Model), and the prior elicited from the most accurate expert (the expert who achieved the highest score in the Classical Model). The authors found that in 15 out of 45 studies the best experts and the performance-based model were identical (i.e. all weight was assigned to one expert), in 27 out of 45 studies the performance-weighted aggregate priors were the most accurate, in two cases the best expert was the most accurate, and in one case it was the equally weighted prior.

Colson and Cooke (2017) analysed the internal validity of seeds used in every study in the TU Delft database published between 2006 and 2015. The authors identified 33 studies in total where all studies contained between 7 and 17 seeds, and the majority contained 10. The authors found that the equally weighted aggregate priors outperformed the performance-weighted priors (using the Classical Model) in 30% of the cases, while the best expert was more accurate than the performance weighted prior in 3 out of the 33 studies.

The findings suggest that, in general, the Classical Model effectively weighs priors, although not always.

2.4.1.3. External validity

Lin and Cheng performed ROAT analysis on 28 (2008) and then 40 (2009b) studies from the TU Delft database. In the first study (Shi-Woei Lin and Chih-Hsing Cheng, 2008) the authors

[5] TU Delft database is a database of professionally conducted Classical Model studies.

found that the performance-weighted priors were significantly more accurate than the equally weighted priors, while in the second study (Shi-Woei Lin and Cheng, 2009b) the difference was not statistically significant.

Cooke (2008) identified studies in the TU Delft database with 16 or more calibration questions. The author split each set of seeds into two halves, where one half served as the 'training' set (i.e. the seeds) and the other half as the 'test' set (i.e. as the target parameters). They then assessed whether weights derived from the training set led to more accurate aggregate priors on the test set than equally weighted aggregate priors. It is not clear how the authors selected which seeds were training and which were test sets. The performance-weighted aggregate priors outperformed the equally weighted priors in 20 out of 26 comparisons.

Flandoli et al. (2011) performed similar analysis on further five data sets, where the size of the test set was either 8 seed questions or 30% of the total seed questions – whichever was larger. The author sampled all possible combinations of training and test sets and reported that performance-weighted priors generally outperformed equally weighted priors.

Eggstaff et al. (2014) performed cross-validation on the 62 studies in the TU Delft database that were available at the time. The authors included every possible training/test set combination and found that performance-weighted aggregate priors were significantly more accurate than equally weighted priors.

More recently, Colson and Cooke (2017) performed cross-validation analysis on all studies in the TU Delft database published between 2006 and 2015. The authors identified 33 studies. All studies contained between 7 and 17 seeds, while the majority contained 10. The authors found that the overall score was higher for the performance-weighted priors, but that this was largely driven by the information score; the calibration (accuracy) score was in fact higher in the equally weighted priors.

Given the inconclusive findings from the existing literature, the next section explores factors that affect the ability of weights to improve the accuracy of the aggregate prior. Section 2.4.2 discusses methods for deriving weights from experts' characteristics, then 2.4.3 discusses methods for deriving weights from experts' elicitation performance.

2.4.2. Further exploration of the methods for deriving weights using experts' characteristics

Section 2.3.2 identified two elicitation exercises in HTA where experts' characteristics were used to weight their priors (Shabaruddin *et al.*, 2010; Haakma *et al.*, 2014); both based their weights on substantive expertise. Haakma et al. (2014) derived scores from experts' length of experience, frequency of use of the technology of interest, and experience of using the technology of interest in other areas, while Shabaruddin et al. (2010) scored experts according to the number of patients they prescribe the treatment under consideration.

In both studies, the characteristics used to define experts' experience and the weights assigned to different levels of experience were determined by the investigator, and it is not clear whether the relationship between characteristics and weights was based on evidence or chosen arbitrarily.

Identifying the characteristics that can be used in weighting is indeed difficult – Chapter 1 highlighted the uncertainty in what makes experts substantive. In a recently published book chapter, Bolger (2017) discussed expert selection in elicitation and reviewed indicators of substantive expertise. The author listed many characteristics as indicators of expertise, including job title, role, formal qualifications, proof of completion of training courses, years of on-the-job experience, awards, citations, and published papers. However, the author cautioned that all stated indicators are associated with limitations. Indicators of rank – such as job title, role and awards – can be acquired through means other than skills and knowledge; peer-recommendation can be misleading as the way colleagues/service users perceive experts may not be a good indicator of that expert's skill set; qualifications and completion of training courses are likely to demonstrate a basic level of knowledge but may not always guarantee on-the-job experience, which is believed to be required to develop expertise (Shanteau, 1992); and more years of experience do not always lead to more expertise (Ericsson, 2006).

Furthermore, weighting experts on the basis of their substantive experience only assumes they are equally normative and accurate when making probabilistic assessments. The role of experts' characteristics in capturing other factors believed to affect their priors is not clear.

2.4.3. Further exploration of the role of methods for deriving weights from experts' elicitation performance

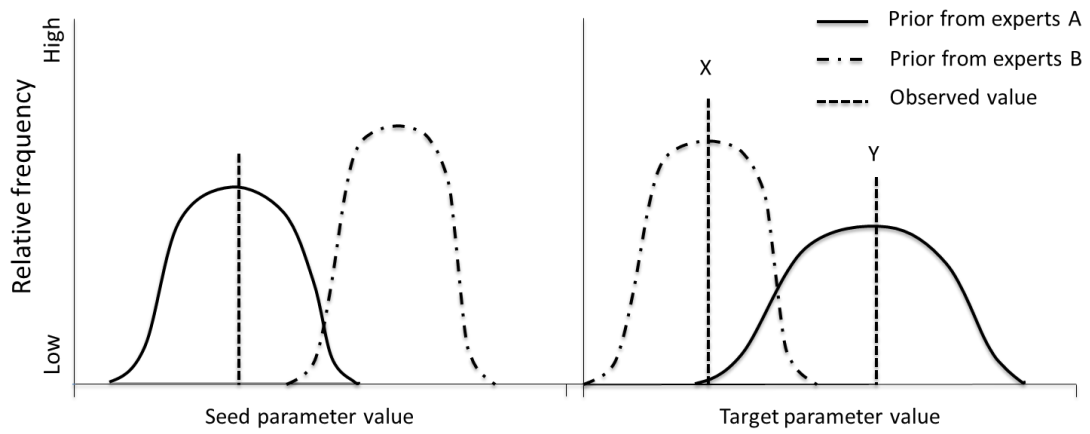
Section 2.2.2 suggested that deriving weights from experts' elicitation performance involves three steps: 1) choosing one or more seed parameters; 2) assigning scores to experts' priors on those seeds; and 3) deriving weights from experts' scores. Sections 2.4.3.1 to 2.4.3.3 discuss the role of each of these three steps in achieving internal and external validity.

2.4.3.1. Choosing the seed parameter

The choice of seed can affect the external validity of weights. Weighting experts by their elicitation performance assumes that the accuracy of their prior on the seed is representative of the accuracy of their prior on the target parameter. This is illustrated in Figure 2.5, showing hypothetical priors on one seed and one target parameter elicited from two experts. Expert A is assigned a higher score because their prior on the seed parameter is closer to the observed value of the seed. Experts A would thus be assigned a greater weight than expert B. If the true value of the target parameter is Y then Expert A's prior on the target parameter will be more accurate than the prior from Expert B, and assigning higher weight to A will improve the accuracy of the aggregate prior. The score can be said to be generalisable. If the observed value of the target parameter is X the opposite is the case, and assigning a higher weight to Expert A will bias the aggregate prior. Generalisability of the score thus determines whether weighting experts by their measured performance improves the accuracy of the aggregate prior.

The choice of seed can affect the generalisability of the score. Section 2.3 proposed that experts' priors are affected by their substantive expertise, perspective, normative expertise and accuracy of assessments (i.e. their susceptibility to bias). If this is the case, their performance-based weights will generalise if the extent to which each of these factors affect the accuracy of experts' priors is similar for the seed and the target parameter. For example, if an expert has a biased perspective on the seed (because of the patient population they see) their score may be low. If their knowledge of the target parameter is then less biased, their score will not generalise. In fact assigning lower weight to such experts could reduce heterogeneity of the expert sample, which, as discussed in Chapter 1, can bias the aggregate prior.

Figure 2.5. Hypothetical priors on one seed and one target parameter elicited from two experts: A and B.

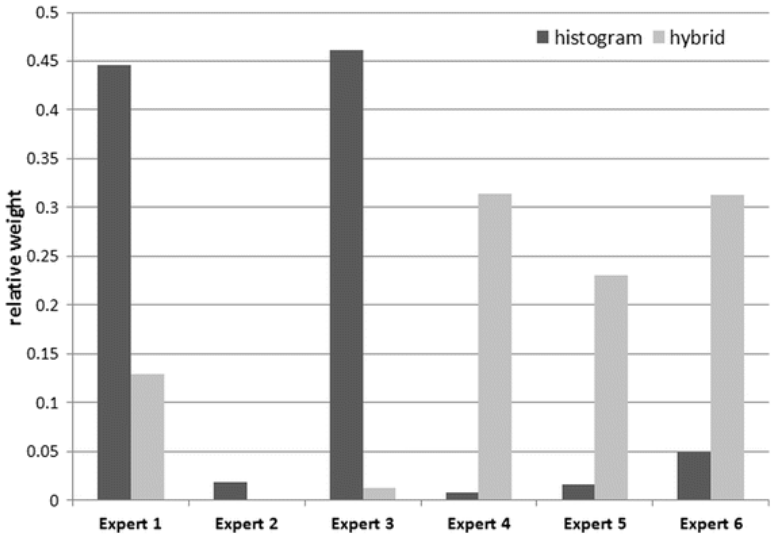


Quigley et al. (2017) highlighted that there is no empirical evidence on what affects generalisability of the seed. The authors argued that seeds should be domain specific, or at least adjacent, because experts tend to be more accurate within their domain of expertise. Section 2.2.2 identified five applied elicitation exercises in HTA that used performance-based weights (shown in Table 2.2) and found that all were related to the target parameter. However, their generalisability is not certain. Out of the five studies, two reported sufficient information to discuss the generalisability of experts' scores (Soares *et al.*, 2011; Grigore *et al.*, 2016).

Soares et al. (Soares *et al.*, 2011) asked four seed questions and found that different seeds lead to disparate weights, suggesting the scores were not generalizable. The authors chose to weight experts equally when aggregating the priors.

Grigore et al. (2016) elicited one seed and one target parameter, using two different methods for each: the roulette and the hybrid methods. The authors reported the weights generated using each method (shown in Figure 2.6), and experts' priors on the target parameter using each method (shown in Figure 2.7).

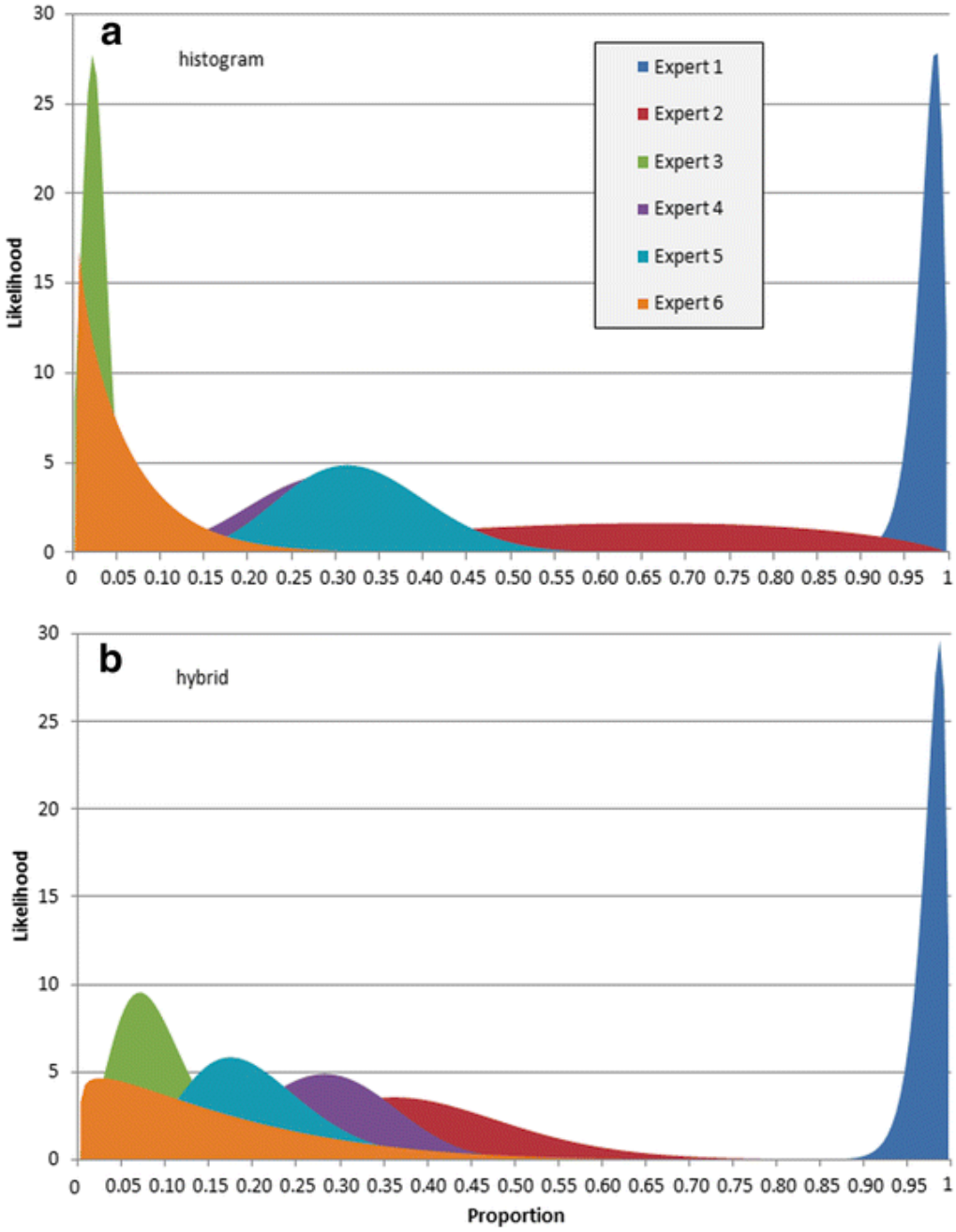
Figure 2.6. Weights derived from experts prior on the seed parameter using two elicitation techniques: histogram and hybrid. Cited from Grigore et al. (2016)



The histogram method placed almost all weight (around 90%) on Experts 1 and 3. Both Experts 1 and 3 were assigned similar weights suggesting the accuracy of their priors was similar, and different to the remaining three experts. Yet, their priors on the target parameter were opposite and extreme, suggesting they could not both be equally accurate and better than the remaining three experts. Therefore, the weights are unlikely to correlate with their performance on the target parameter. When priors on the same seeds were elicited using the hybrid method, the majority of weights were assigned to Experts 1, 4, 5 and 6, suggesting their performance was similar (in particularly for Experts 4,5 and 6 who had similar scores). However, experts' priors on the target parameter suggested that Experts 1 had the opposite view to everyone else, while Expert 2, who was assigned a very low weight, had a very similar prior to Experts 4, 5, and 6. The weights based on the priors elicited using the hybrid method are therefore also unlikely to be correlated with experts' performance on the target parameter.

The two analysed case studies suggest that experts' scores in elicitation in HTA do not generalise. Colson and Cooke (2017) found that the external validity of performance-based weights improved with the addition of new seeds and that over 10 seed questions are desirable; however, identifying over 10 seed questions that capture experts' substantive expertise and perspective in HTA can be difficult – the maximum identified in this literature review was 8 (Fischer, Lewandowski and Janssen, 2013).

Figure 2.7. Priors elicited from six experts, on one parameter, using two techniques: a)histogram, b)hybrid. Cited from Grigore et al. (2016)



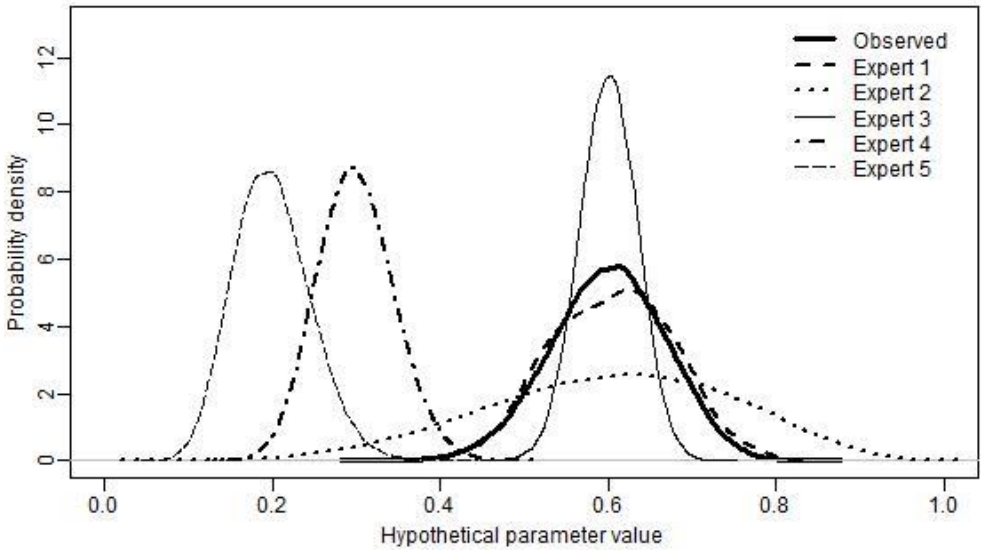
It may be possible to choose seeds that isolate specific characteristics. For example, a seed can be a parameter that all experts are known to have equal knowledge and experience on, so that any variation in performance can be attributed to their normative expertise and ability to make probabilistic assessments. This approach assumes that substantive expertise and perspective are comparable between experts. Furthermore, it is not clear how such

seeds can be identified – the literature review described in this chapter did not identify any applied examples of such seeds.

2.4.3.2. Choosing the scoring method

Experts' priors on seeds can vary in bias and uncertainty (Cooke, 1991). Bias defines how close experts are to the 'true value' of the parameter, whereas precision is the degree of agreement for a series of measurements (i.e. uncertainty). Figure 2.8 demonstrates five examples of how priors can vary in bias and precision. Expert 1 (whose prior is identical to the probability distribution of the seed) is unbiased and precise. Expert 2 is unbiased, but imprecise (uncerconfident) compared to Expert 1. Expert 3 is unbiased but overconfident, as they underestimate uncertainty around the parameter. Experts 4 and 5 are both biased and precise. They place equal probability on the parameter probability distribution, although Expert 4 is less biased than Expert 5.

Figure 2.8. Beliefs of five experts about a seed parameter and its 'true' probability distribution used to demonstrate different levels of bias and uncertainty.



Discrepancy between experts' priors and the true probability distribution (i.e. experts' accuracy) is defined by both bias and precision, but a scoring method may only wish to only capture one of the factors. For example, if the seed question does not aim to capture experts' knowledge, but only their ability to assess their own uncertainty accurately, it may be desirable to penalise experts' overconfidence only, rather than bias and imprecision. Conversely, if the seed question aims to capture experts' substantive expertise, it may be

desirable to assign better scores to more knowledgeable experts by penalising underconfidence and bias.

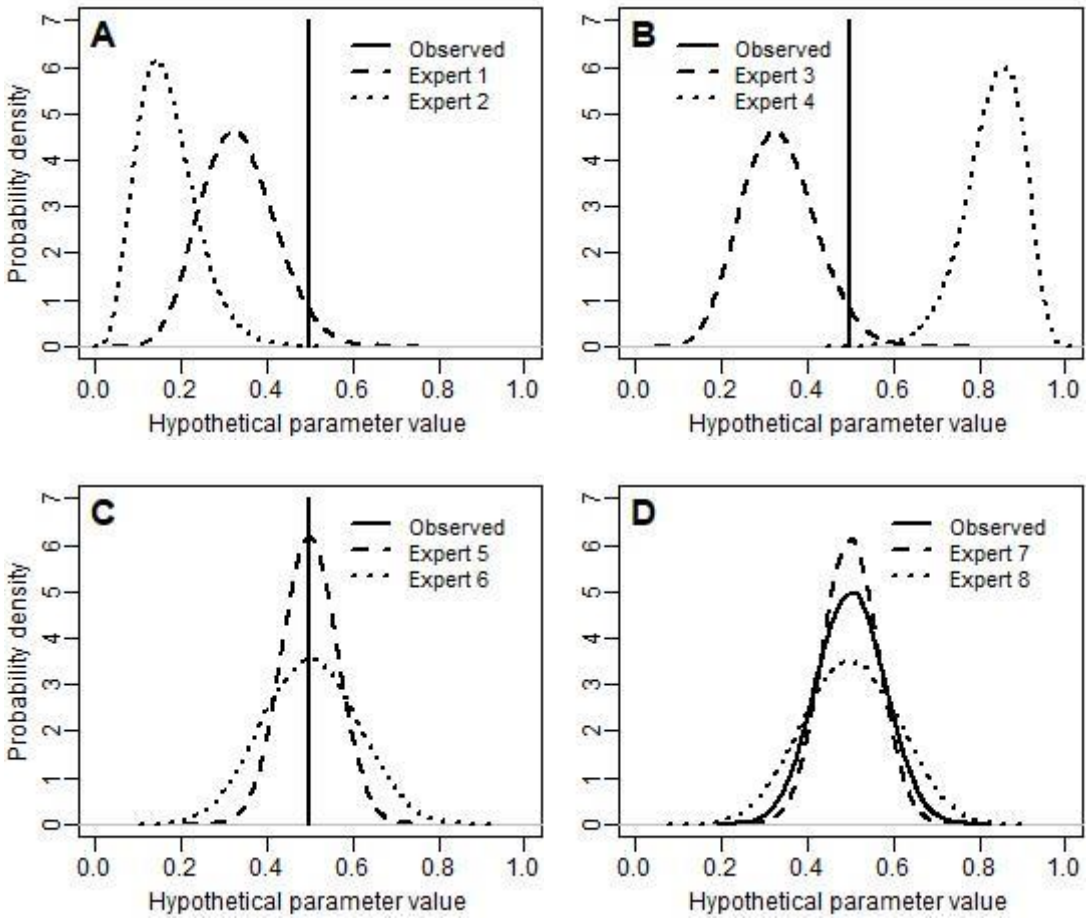
The way scoring methods value bias and uncertainty can affect both internal and external validity of the resulting weights. Assuming that an accurate prior is unbiased, and as precise as the observed probability distribution, the performance-based weights can only improve the accuracy of the aggregate prior on the seed (i.e. to be internally valid) if the score takes into account the direction of bias in experts' priors and their uncertainty relative to the uncertainty with which the value of the seed is known. This is demonstrated in Figure 2.9, showing four samples of two experts.

In Figure 2.9.A both priors are biased in the same direction (they both underestimate the value of the parameter) and the aggregate prior is the least biased if Experts 1 and 2 are assigned weights 1 and 0 respectively. In Figure 2.9.B the prior elicited from Expert 3 is identical to that from Expert 1 in Figure 2.9.A, while the prior elicited from Expert 4 is as biased as that from Expert 2 (the distance between the prior and the observed value are identical) but in the opposite direction (they overestimate the parameter value). The optimal aggregate prior for Experts 3 and 4 is one where both experts are assigned non-zero weight as the observed value of the seed is between them. If a scoring method captures the size of bias but not the direction, then priors from Expert 1 and 3 will be assigned identical weight, as will those elicited from Experts 2 and 4. In addition, Experts 1 and 3 will be assigned greater weights because they are less biased than 2 and 4 (their priors are closer to the observed value). However, unless the weight for Experts 1 and 3 is 1, and that of Experts 2 and 4 is 0 (so that both aggregate priors are identical), the aggregate prior derived from Experts 3 and 4 will be more accurate than that derived from 1 and 2 because any distribution that combines 1 and 2 is further away from the observed value of the seed than any distribution between 3 and 4.

In Figure 2.9.C the value of the seed is known with certainty, and priors elicited from Experts 5 and 6 are both unbiased, but Expert 5 is more certain, or precise, than Expert 6.The aggregate prior will thus be unbiased, and it will be most informative (precise) if Experts 5 and 6 are assigned weights 1 and 0, respectively. In Figure 2.9.D, priors elicited from Experts 7 and 8 are identical to those from Experts 5 and 6 but the value of the seed is uncertain, so Expert 7 is judged to be overconfident. Their aggregate prior will also be unbiased, but assigning weights 1 and 0 (which led to the most informative aggregate prior

in Figure 2.9.C) would lead to an overconfident prior. The overconfidence can be prevented by assigning a non-zero weight to Expert 8.

Figure 2.9. The role of scoring methods in achieving internal validity demonstrated through hypothetical priors (elicited from four samples of experts).



External validity is also affected by the scoring method. If the seed aims to capture only experts' ability to assess their own uncertainty and express it in the required format, then it may be desirable to penalise bias and overconfidence, but not imprecision. If the value of the seed is uncertain due to a lack of evidence and experts are judged to be reasonably able to know the value with more certainty then it may be desirable not to penalise their overconfidence, only bias and imprecision. If the value of the seed is known with certainty, and the score aims to capture their substantive expertise, then it may be desirable to penalise bias, overconfidence and imprecision.

Section 2.2.2 identified six methods for scoring experts' priors based on their accuracy (Kullbeck-Liebler discrepancy combined with Shannon's relative information, Brier's Probability Score, absolute difference CICP), yet no studies were identified that analyse the

differences between them. This sections starts with an analysis of how each of the methods assigns scores to experts' priors, then discusses the implications of the findings for their role in deriving performance-based weights.

Classical model

Section 2.2.2 described that the Classical Model is implemented by eliciting 5th, 25th, 50th, 75th and 95th percentiles for multiple seed parameters, then comparing the proportion of seeds that are observed to be within each elicited interval to the probabilities of each elicited quantile (0.05, 0.2, 0.25, 0.25, 0.2, 0.05) (Cooke, 1991).

Experts' accuracy is assessed using the Kullbeck-Liebler discrepancy (see Equation 2.1 for details on how to derive the score). The KL score compares experts' probabilities to observed relative frequencies. Table 2.7 shows that the method penalises bias, uncertainty and overconfidence, as all three lead to discrepancies between the probabilities placed on each outcome (or value) of the seed and the frequencies with which they occur. However, the scores do not reflect the direction of bias, nor whether experts are overconfident or uncertain.

Table 2.7. KL discrepancy scores derived from four experts with different degrees of bias and uncertainty.

Expert	Proportion of observed outcomes (or values) of the seed that fall between the stated percentiles						KL score
	0-5th P=0.05	5-25th P=0.2	25-50th P=0.25	50th-75th P=0.25	75th-95th P=0.2	95th-100th P=0.05	
Perfect	0.05	0.2	0.25	0.25	0.2	0.05	0
Imprecise	0.01	0.16	0.3	0.3	0.16	0.01	0.0058
Overconfident	0.08	0.22	0.2	0.2	0.22	0.08	0.0279
Biased	0.2	0.35	0.2	0.15	0.1	0	0.3519

The information score (derived using Shannon's entropy) in the Classical Model is an additional measure of uncertainty introduced to place higher weights on experts who are more certain, making the aggregate priors more certain (Cooke, 1991). If two experts have identical KL scores (because of the proportion of observed seed values that were in each quartile of experts' priors) but one of them is consistently more certain (i.e. their quantiles

are closer together) then that expert will achieve a better information score and be assigned a higher weight overall. While there may be value in assigning higher weights to the more precise expert when two equally accurate experts are compared, it is not clear what the value of the information scores is when experts' priors are not identical, as the less certain prior will have already been penalised for its uncertainty in the KL score.

Absolute difference /best estimate fraction

As highlighted in section 2.2.2 absolute difference represents the absolute difference between random draws from experts' priors and the observed value of the seed (taking into account its parametric uncertainty) (Bojke *et al.*, 2010). Experts are penalised for bias and uncertainty, as both increase the distance between the priors and the observed probability distributions, decreasing the probability that the experts will be 'the best' in the sample. Experts are not penalised for overconfidence, as narrow priors decrease the maximum distance between priors and the observed probability distribution achieved when samples are drawn from opposite ends of the distribution, as demonstrated in Table 2.8.

Absolute difference scores have been used to derive the best estimate fraction (Bojke *et al.*, 2010) that represents the proportion of random samples from an expert's prior that minimises the absolute difference from the observed value of the seed in comparison to other experts in the sample.

Table 2.8. Experts' priors on a seed with observed mean value of 0.5 and range 0.3-0.7, and the resulting maximum distance between random samples of experts' priors and the observed probability distribution.

Expert	Expert's mean (range)	Maximum distance between random samples (achieved if values are sampled from opposite ends of the range)
Perfect	0.5 (0.3-0.7)	0.4 (0.7-0.3)
Imprecise	0.5 (0.2-0.8)	0.5 (0.8-0.3 or 0.7-0.2)
Overconfident	0.5 (0.4-0.6)	0.3 (0.6-0.3 or 0.7-0.4)
Biased	0.2 (0-0.4)	0.6 (0.6-0)

Neither absolute difference nor BEF scores reflect what drives the score (bias, imprecision or overconfidence), nor the direction of bias in the prior.

Brier's probability score

Brier's score (Brier, 1950) compares experts' probabilities assigned to specific outcomes of events to scores of 0 or 1 depending on whether the outcome occurs or not (see Equation 2.4 in section 2.2.2 for details). Brier's score assesses probabilities assigned to binary variables, and experts are penalised for assigning probabilities that are too high or too low. For example, if an outcome occurs with a relative frequency of 0.5, then assigning a probability of 0.5 to that outcome will lead to a score of 0.25[6], and assigning any higher or lower probabilities increases (worsens) the scores (e.g. if expert's probability is 0.7, the score is 0.29[7]).

Decomposition of Brier's probability score

Section 2.2.2 highlighted that Brier's scores can be decomposed to determine what determines experts' performance (for example, calibration, variance or resolution). The final scores derived using these methods are mathematically equivalent to Brier's PS, and so, when scores are used to weight experts, they lead to identical weights. This method is thus not considered in this section in further detail.

Ranked Probability Score

Section 2.2.2 discussed that the ranked probability scores were derived specifically for the applied example in meteorology, developed by Epstein (1969); applying the scoring method in HTA would require deriving a new equation specific to that example, taking into account the number of possible outcomes and the consequences of each of those outcomes.

Taking into account the cost of wrong decisions would require modelling of each possible outcome, which adds complexity and is unlikely to be feasible in HTA, particularly if elicitation is conducted to inform the structure of the model. This method is thus not considered in further detail in this chapter.

[6] d_j=1 in 50% of observations, and d_j=0 in the remaining 50% (see Equation 2.4), and so score = 0.5^2*0.5+0.5^2*0.5=0.25

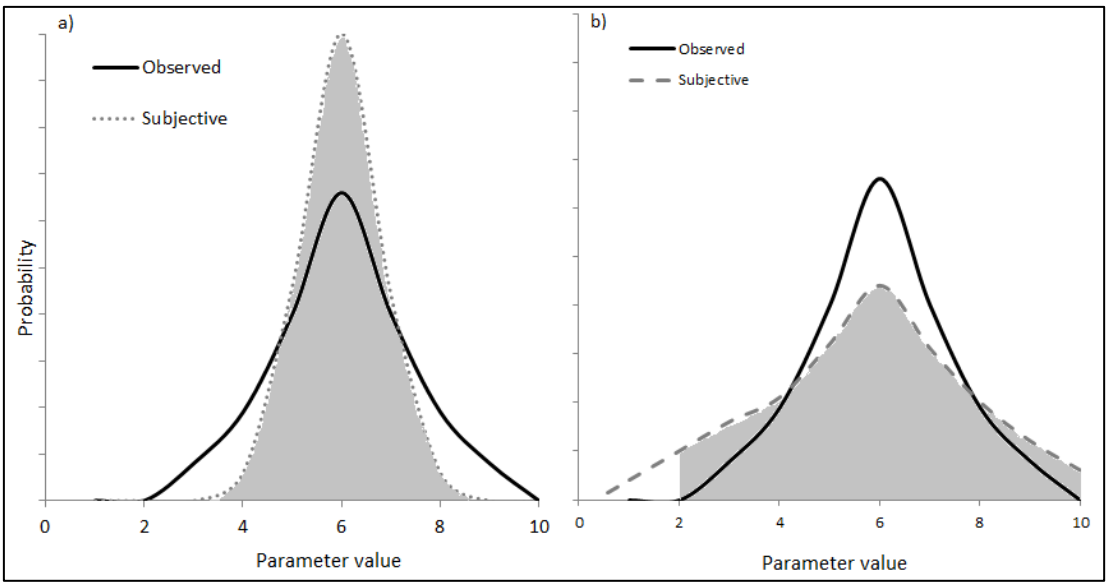
[7] 0.7^2*0.5+0.3^2*0.5=0.29

Confidence Interval Coverage Probability

CICP scores capture the proportion of (point estimate) observed values to experts' confidence intervals. (Murphy and Winkler, 1977) Scores are only affected if experts are overconfident, and so the observed values fall outside their range.

However, it is possible to adjust the methods used to derive CICP to capture uncertainty. If the seed is measured as a continuous variable and its parametric uncertainty is taken into account, the CICP score can be calculated as the proportion of values sampled from experts' priors that fall within the observed probability distribution. The score derived in this way reflects the probability that the experts place on the observed value (see the highlighted area in Figure 2.10), i.e. it penalises imprecision.

Figure 2.10. Probability distribution derived from an observed sample and elicited from an expert. The CICP is the highlighted proportion of the subjective probabilities.



Comparison of different scoring methods

The analysis of different scoring methods has highlighted that these methods vary in how they value bias and precision, and Table 2.9 summarises the findings. The table includes different variations of methods discussed in this section – the KL scores in isolation (as opposed to in combination with the information score as part of the Classical Model), and the CICP score that penalises imprecision rather than overconfidence.

Table 2.9. The ability of different scoring methods to capture different aspects of experts' beliefs.

Method	Penalises bias	Captures direction of bias	Penalise over-confidence	Penalise imprecision	Captures uncertainty relative to the seed
Classical Model (combined score)	✓	×	✓	✓	×
KL discrepancy	✓	×	✓	✓	×
Absolute difference	✓	×	×	✓	×
Brier's PS	NA	NA	✓	✓	×
CICP	×	×	✓	×	✓*
CICP precision score	×	×	×	✓	✓**

***if overconfident**

****if imprecise**

The way different methods value bias and precision can affect their role in deriving weights.

None of the scoring methods capture the direction of bias, and so their internal validity is uncertain. However, the role of assessing internal validity is not clear; if those experts who underestimate the seed also underestimate the target parameter and vice versa, then internal validity can be required for the performance-weighted aggregate priors to be more accurate than unweighted ones. Conversely, if experts' priors on seed parameters are consistent in accuracy, but exhibit no specific patterns of belief, then performance-based weights can potentially improve the accuracy of the aggregate prior on the target parameter even if they are not internally valid, by giving more say to those experts who tend to be more accurate when making probabilistic assessments.

CICP does not reflect bias so long as the expert places positive probability on all possible outcomes. It is the only method that penalises overconfidence but does not penalise imprecision, so it is a suitable method if the scores aim to capture experts' susceptibility to overconfidence and not precision.

All the remaining methods penalise imprecision.

Absolute difference and CICP precision scores do not penalise overconfidence. They could be suitable if the value of the seed is uncertain due to a lack of evidence and experts are judged to be reasonably able to know the value with more certainty. BEF may be preferred to CICP precision scores as they also capture the extent of bias, while CICP precision scores only depend on the probability placed on the observed value(s) of the seed.

The Classical Model, KL discrepancy and absolute difference all penalise bias, imprecision and overconfidence and so they are suitable scoring methods when both precision and accuracy (lack of bias) are considered to be desirable, and experts are judged not to be able to know the value of the seed with more certainty that can be concluded from the data used to measure the seed. When KL scores for different experts are not identical, the Classical Model additionally penalises the imprecise expert.

It is also important to note that the methods vary in the format of the seed and the units in which the seed is measured. Brier's score utilises binary variables as seeds; it compares experts' probabilistic assessments (as point estimates) to whether the event has occurred or not. The remaining methods use continuous or ordinal variables as seeds. Furthermore, the Classical Model and CICP compare the observed value of the seed (the point estimate) to experts' priors, while the Best Estimate Fractions (utilising absolute difference) compare the probability distribution of the parameter (taking into account the parametric uncertainty) to experts' priors.

Methods that consider the point estimates of the seed require elicitation of more seeds to take into account the uncertainty with which it is measured. This is likely to be why BEF is the most common method for deriving performance-based weights in HTA where the mean number of seeds used is relatively small (3.75 as shown in section 2.2.2) (Bojke *et al.*, 2010; Soares *et al.*, 2011; D Sperber *et al.*, 2013).

Brier's score, the Classical Model and CICP can in theory all be adapted for use in HTA where fewer seeds are used but uncertainty around them is taken into account, although no applied examples where this has been done were identified in this literature review.

Brier's score can be adapted for use with seeds that are continuous variables by using the probability distribution of the parameter to sample possible outcomes. The random samples can then be compared to the probability that experts placed on that particular outcome (or interval of values).

KL discrepancy scores can be adapted to take parametric uncertainty into account (and so require fewer seeds to be elicited) by sampling from the probability distribution of the parameter and treating every observation as a new seed. Alternatively, it is also possible to compare experts' priors and observed probability distributions directly as the KL discrepancy score is a measure of discrepancy between probability distributions (Kullback and Leibler, 1951).

It is not clear how the derived KL scores can be used in the Classical Model, as deriving scores in the Classical Model requires the sample size (i.e. the number of seeds) while the method doesn't use multiple seeds, only multiple iterations of one seed.

2.4.3.3. Choosing the method for deriving weights

Section 2.2.2 described different approaches to deriving weights from experts' scores. The Classical Model combines the calibration and information into a single score, and uses it as the weight (while scaling them so they add up to 1 first). Bojke et al. (2010) use the best estimate fraction scores directly to derive weights. In both cases the weights are proportional to experts' scores.

The Contribution-Weighted Model use experts' contribution to the score of the aggregate priors to derive weights. If there are only two experts in the sample, the Contribution-Weighted Model leads to the same scores as the Classical Model, as removing one expert from the sample will give the score of the second expert.

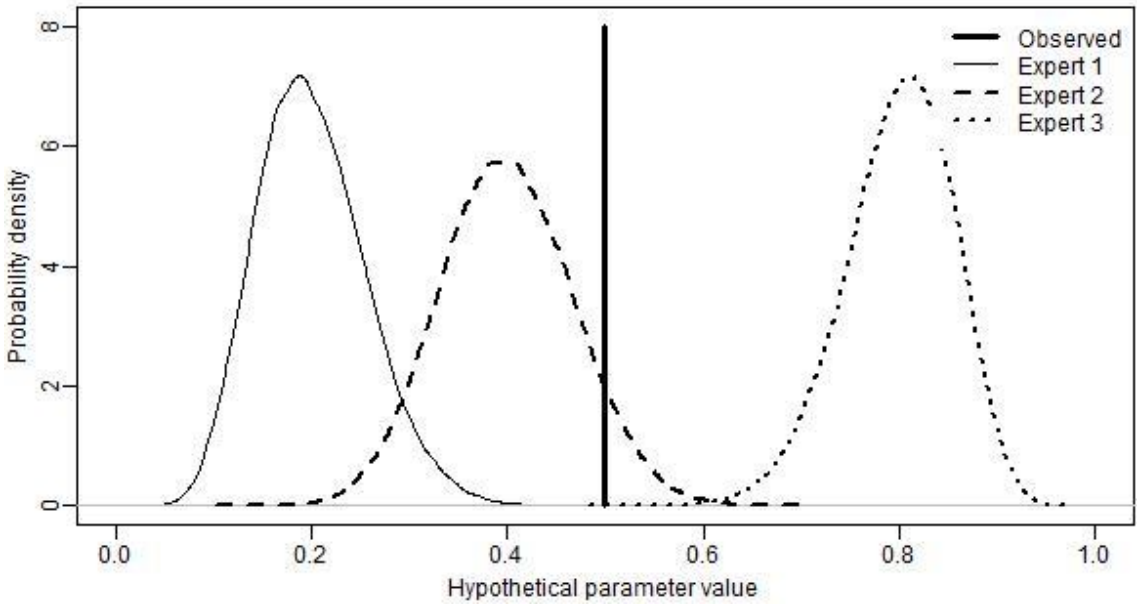
Both approaches can in theory be applied to all scoring methods – Brier's score can be converted into weights directly, and both the Classical Model and Best Estimate Fraction scores can be used to derive weights based on how the score of each individual expert affects the score of their unweighted aggregate prior.

It is not clear which methods is preferred.

The Contribution-Weighted Model is more likely to lead to internal consistency. This can be demonstrated using the examples in Figure 2.11, where hypothetical priors elicited from three experts are shown. The observed value of the seed is 0.5. Expert 1 and Expert 3 are equally biased but in opposite directions, while Expert 2 is less biased, and their bias is in the same direction as that of Expert 1. The classical Model would assign the same score to Expert 1 and Expert 3 because they are equally biased, and a higher score to Expert 2 who is less biased in comparison. The Contribution-Weighted Model would assign weights based

on the accuracy of the aggregate priors derived by pooling two experts at a time. Pooling priors from Experts 1 and 3 would lead to an aggregate prior with mean 0.5, as they are symmetrical around this value. Pooled priors from Experts 2 and 3 would be less accurate, as their mean would be higher than the observed value (0.6). Pooled priors from Experts 1 and 2 would be the least accurate, as their mean would be 0.3. Removing Expert 2 from the sample would lead to the highest score, and so they would be assigned the lowest weight.

Figure 2.11. Beliefs of three experts about a seed parameter and its observed probability distribution used to demonstrate the effect of different methods for deriving weights.



However, as highlighted earlier in this section, internal validity is not required to achieve external validity unless experts' patterns of belief are the same for the seed and the target. If this is not the case, then the Contribution-Weighted Model can decrease external validity by assigning the lowest weight to the most accurate expert, as shown in the example in Figure 2.11 where Expert 2 is assigned the lowest weight despite being the most accurate of the three.

2.5. Summary of findings

This chapter aimed to derive guiding principles for deriving weights in elicitation. To do so, three specific objectives were set: 1) to identify existing methods for deriving weights, 2) to discuss the role of weighting in elicitation, and 3) to evaluate and compare methods for deriving weights.

Section 2.2 conducted a literature review to address the first objective, and identified two approaches to deriving weights, and multiple methods within each approach (objective 1).

The literature review was a non-systematic BCSC. A potential caveat of the BCSC method is the reliance on authors' referencing to identify relevant publications (Hinde and Spackman, 2015). If a publication is insufficiently referenced, and it has not been cited by other publications on the topic of interest, it can lead to a 'citation island'. Any such citation islands could have been missed from the literature search. Furthermore the search strategy was focused on applied examples in HTA. It is possible that additional weighting methods that have been applied in other fields were missed. The methods highlight the difficulty in carrying out systematic searches in elicitation due to its widespread application and varied terminology.

In order to derive guiding principles for choosing between the various options for deriving weights, section 2.3 revisited the aims of an elicitation exercise highlighted in Chapter 1, and then discusses factors that could affect experts' contributions towards achieving those aims (objective 2). Section 2.4 then discussed advantages and limitations of different methods for deriving weights in opinion pooling (objective 3).

The chapter proposed that weighting can potentially compensate for methodological challenges in elicitation by giving 'more say' to experts who are believed to be less affected.

Four factors were identified that could affect experts' contribution: substantive expertise, perspective, normative expertise and ability to make accurate probabilistic assessments. Variation in the four factors could provide a basis for differential weighting.

The importance of each factor is likely to depend on the elicitation process – expert recruitment, the provision of background information and an opportunity for discussion with other experts are thought to improve substantive expertise and minimise bias due to perspective, whereas elicitation process design, training, and evaluation and feedback are thought to reduce cognitive biases in assessing quantities and expressing uncertainty.

Different weighting methods can be used to capture different factors, and understanding where the process is lacking can inform which weighting method to use.

Performance-based weights are affected by experts' normative expertise and their ability to make accurate probabilistic assessments. Seed parameters tend to be domain specific, and so are likely to be affected by substantive expertise and perspective as well. It may be possible to ask non-domain seeds that all experts have equal knowledge on, in order to

capture normative expertise and the ability to make accurate probabilistic assessments only, but this has not been used in applied elicitation exercises.

Weights derived from experts' characteristics can be used to capture their substantive expertise independently of their normative expertise, perspective and ability to make accurate probabilistic assessments. It is not clear whether characteristics can be used as proxies for other factors that can affect experts' priors – no applied examples were identified in the review.

The challenge in implementing the outlined principles arises from the lack of understanding of how to determine what the challenges in the elicitation process are. For example, it is not clear how to demonstrate that training and planning were optimal and that the only basis for differential weighting is substantive expertise.

Furthermore, there are many methodological challenges in deriving weights that make it unclear whether they successfully achieve their objective.

The characteristics that have been used as proxies for substantive expertise and the derived weights tend to be chosen arbitrarily. It is not clear whether they successfully minimise bias and uncertainty in the weighted aggregate priors.

When seeds used to derive performance-based weights are domain-specific, their score on the seed will represent their performance on the target parameter only if their substantive expertise and perspective equally affect the seed and the target parameters. When this is not the case, there is a risk that lower weights will be assigned to experts with a unique but important perspective, reducing the heterogeneity of the expert sample. Several applied elicitation exercises have reported this challenge (Fischer, Lewandowski and Janssen, 2013; Grigore *et al.*, 2016).

The rest of the thesis compares and evaluates different methods for deriving weights in an elicitation exercise applied in HTA. Chapter 3 describes the design of the study, while Chapters 4, 5 and 6 analyse the results. Specifically, Chapter 4 analyses the results of the elicitation exercise. Chapter 5 then explores whether experts' characteristics can predict their elicitation performance, and Chapter 6 observes the effect of different weighting methods on the accuracy of the aggregate priors and the results of the cost-effectiveness decision model.

Chapter 3. Comparison of methods for weighting experts' priors: REFORM elicitation study protocol

3.1. Introduction

Chapter 2 identified two general approaches for deriving weights: based on experts' observed characteristics and their measured performance in elicitation. Both aim to improve accuracy of the aggregate prior but they differ in the way in which they capture experts' 'contributions'.

Chapter 2 concluded that it is not clear which method for deriving weights is optimal.

This chapter aims to design a study that will compare different weighting methods in an elicitation exercise applied in CEDM. The study has two specific objectives:

1) To explore factors that affect experts' priors.
2) To compare the impact of different weighting methods in an applied case study in CEDM.

Understanding factors that affect experts' priors will inform what is captured by different weighting methods and how they should be applied. For example, if priors are predominantly affected by experience (substantive expertise and perspective) then weights should be based on characteristics (in particular substantive expertise). If, however, they are predominantly affected by expert' ability to make accurate probabilistic assessments, then performance-weighting may be preferred, and seeds may not have to be domain specific. Furthermore, understanding factors that affect experts' priors can resolve methodological uncertainties in other steps of the elicitation process, such as how to define experts for elicitation (what type of experience improves their accuracy).

Chapter 2 highlighted that only including the most accurate experts in a sample may not lead to the most accurate aggregate prior- in fact it can reduce the heterogeneity of the sample. The second objective is thus to observe the effect of weighting methods on the weighted aggregate prior.

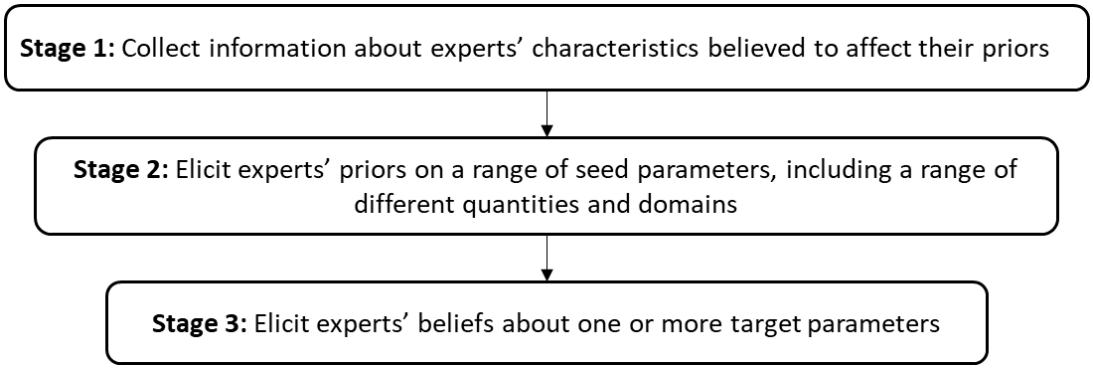
Section 3.2 describes the methods employed to design a study that achieves the two objectives, while section 3.3- 3.5 describe the study design.

3.2. Methods

The study consisted of three stages shown in Figure 3. Stage 1 involved recruiting a relatively large sample of experts and capturing characteristics believed to affect their priors. Stage 2 elicited a range of seed parameters from experts. Stage 3 elicited experts' priors on at least one target parameter that informed a cost-effectiveness model.

The results from Stages 1 and 2 were used to achieve the first objective of the study - to explore factors that affect experts' priors – by measuring the effect of the captured characteristics on their priors on the seed parameters. The second objective - to compare the impact of different weighting methods – was based on the results from all three stages where experts' characteristics and priors on the seed parameters were used to derive weights, and their priors on the target parameter were used to observe the effect of different weighting methods on the results of cost-effectiveness analysis.

Figure 3.1. Structure of the elicitation study.



The study was designed in three steps:

1) Identify a suitable case study.
2) Derive measures of characteristics believed to affect experts' priors.
3) Compile a protocol for the elicitation exercise.

Sections 3.2.1-3.2.3 describe methods for each step, respectively.

3.2.1. Identifying a suitable case study

RCTs are considered to be the gold standard evidence in HTA, and so the optimum measure of the 'true' value of domain-specific seed parameters. RCTs were therefore identified as

suitable case studies for achieving the objectives of the study. Trial outcomes are also suggested as appropriate seeds by Cooke (2017).

The trial used in this case study was required to fulfil the following criteria**.**

1) **RCT outcomes can be used as seeds.** The intervention evaluated in the trial must be a novel intervention the effect of which had not been observed by experts.
2) **Timely availability of final outcomes.** Results of the trial needed to be published after the elicitation exercise so that the observed value of the seeds were not known to the experts. The trial also needed to report early enough to allow analysis before submission of this thesis.
3) **Early access to results.** In order to gain access to early results (before publication) the study needed to be conducted at the University of York.
4) **Construction of a cost-effectiveness decision model.** In order to observe the effect of different weighting methods on the cost-effectiveness model outputs, the applied elicitation exercise needed to include elicitation of cost-effectiveness decision model parameters. Thus the trial results needed to inform a decision model where data was not available for one or more parameters.

The available trials were identified on consultation with the director of the York Trials Unit in the Department of Health Sciences, University of York.

3.2.2. Methods for deriving measures of characteristics believed to affect experts' priors

Chapter 2 proposed a set of factors believed to affect experts' priors and discussed how those could be measured. The findings from Chapter 2 were thus used to decide on the characteristics that were captured in this study, and to derive measures of each identified characteristic.

3.2.3. Methods for designing an elicitation exercise

Chapter 1 highlighted the importance of careful and evidence based design of an elicitation exercise. The chapter also emphasised numerous methodological uncertainties, requiring investigators to decide on the most appropriate methods when designing an elicitation exercise. Given that the choice of methods can have a significant impact on the outcomes

of an elicitation exercise, a protocol was derived for this case study to ensure transparency and accountability.

The development of the protocol was informed by the health economics elicitation literature cited in Chapters 1 and 2, and aimed to follow the reporting guidelines for the use of expert judgement derived by Iglesias et al. (2016)

The protocol development was guided by two clinicians specialising in the field for which the elicitation was conducted to gain understanding of the therapeutic area in the trial. They were consulted on potential target parameters to determine whether these can be reasonably assessed by the experts.

The exercise was piloted in the following three stages.

Pilot 1 was used to test different methods for eliciting uncertainty. Chapter 1 highlighted that there are multiple techniques for eliciting priors, and that it is not clear which method is the best. The aim of Pilot 1 was to determine which method was most likely to lead to statistically coherent priors that can be used in the analysis. The methods (described in further detail in section 3.5.3) were piloted on a lay participant who was a university-educated health professional (pharmacist) with quantitative skills comparable to the experts recruited in the study. While they did not have domain specific expertise, this was not expected to impact the findings from the pilot because the elicited parameter was not field-specific- the participant was asked to express their uncertainty around the number of days it rained in their locality every November. The pilot was conducted remotely using publicly available software MATCH (Morris, Oakley and Crowe, 2014); while the investigator guided the participant over the phone. The elicitation piloted techniques and the results of Pilot 1 are provided in section 3.5.3.

Pilot 2 was conducted to test different approaches to eliciting the target parameter. As suggested in Chapter 1, a single model parameter can be informed by eliciting different quantities and there is no guidance on which is best – it tends to be based on what experts find most intuitive. The second pilot was conducted to test different approaches to eliciting parameters in this case study. The methods were piloted on two highly quantitative lay participants to minimise the burden of training. The participants were both postdoctoral Research Fellows in health economics at the University of York. The pilot was delivered in the same format as the main study, as described in section 3.5. The piloted quantities and the results of Pilot 2 are also discussed in section 3.5.2.

Pilot 3 involved testing the final version of the exercise on a sample of seven experts to identify any practical challenges with undertaking the exercise. The sample of experts, the pilot delivery method are discussed in further detail in section 3.5.7.

3.3. Case study: REFORM trial

The chosen case study was the REFORM trial, conducted to measure the clinical and cost-effectiveness of a multifaceted podiatry intervention designed to prevent falls in the elderly. The trial compared the outcomes and the cost of care in people who receive the podiatry intervention and in those who received standard care. The outcome measures in the trial included fall related behaviour (the rate of falls, the proportion of fallers, the time to first fall, the fracture rate), general HRQoL measures (in particular, EQ5D (Klarman, Francis and Rosenthal, 1968)), Geriatric Depression Scale (Yesavage, Brink and Rose, 1982)and two fall-specific outcome measures (Short Falls Efficacy Scale (Kempen *et al.*, 2007), Fear of falling and Activity of Daily Living (Lachman *et al.*, 1998)). The REFORM trial was chosen as it satisfied all criteria set out in section 3.2.1, sections 3.3.1 – 2.3.4 discuss why.

3.3.1. The trial outcomes can be used as seeds

REFORM trial evaluated the costs and effects of a novel intervention and so its effects had not been observed by clinicians. Furthermore, the trial included multiple outcome measures allowing elicitation of multiple seeds.

3.3.2. Construction of a cost-effectiveness decision model.

Trial analysis required a decision model. Trial results provided information on the cost-effectiveness of the intervention after one year (the length of trial follow-up). Chapter 1 highlighted that analysis time horizon for an economic evaluation should be the duration over which any differences in costs and benefits between competing interventions are apparent (Philips *et al.*, 2006; Sculpher *et al.*, 2006). The podiatry intervention in the REFORM trial was designed to be received indefinitely, with a potential effect on mortality through reduced risk in falls, and so the appropriate analysis time horizon was lifetime. (Sculpher *et al.*, 2006) Costs and effects after one year thus needed to be modelled. The structure of the model was influenced by the elicited parameters, and is presented in Chapter 6.

3.3.3. Timely availability of final outcomes.

The trial was due to report 18 months into this thesis, leaving sufficient time both to conduct the elicitation exercise and to analyse the results after the delivery of the exercise.

3.3.4. Early access to results

The trial was conducted at the University of York and access to the trial results was granted.

Further details of the trial are provided in Appendix 3.1, while the protocol has been published in the British Medical Journal (Cockayne *et al.*, 2014).

3.4. Identifying and measuring characteristics believed to affect experts' priors

Chapter 2 suggested that experts' priors are affected by the following four factors:

- Substantive expertise,
- Perspective,
- Normative expertise, and
- Their ability to make accurate probabilistic assessments.

This section describes the measures used to reflect each of the factors in the REFORM elicitation study.

3.4.1. Substantive expertise

As discussed in Chapter 2, a recent review of methods for identifying experts for elicitation (Bolger, 2017) found that there are many potential indicators of substantive expertise, including job title, role, formal qualifications, proof of completion of training courses, years of on-the-job experience, awards, citations, and published papers.

The author found that all specified measures had potential limitations and so information was collected on multiple characteristics, and the effect of each was then explored. Specifically, the following information was collected from experts:

- Their role. Experts are asked to describe their role in free text in as much detail as possible.
- Years of experience in current role, entered as number of years.

- Research experience, aiming to use experts' contribution to the field as a potential indicator of expertise. Research experience was captured in two questions: the number of publications (0-3, 4-20 or 20-50 or over 50 publications) and the number of successful research grants co-written (0, 1-5 or more than 5) to explore the effect of different levels of research activity.
- Proportion of working time spent with patients who are at increased risk of falling, either helping prevent falls or treating fall related injuries. Categories were 0-10%, 11-30%, 31%-50% and 50%-100% to reflect different levels of patient contact.
- Awareness of any research into podiatry interventions designed to reduce the risk of falls (yes/no).

The role, years of experience and research experience were included, as they were suggested as indicators of expertise by Bolger (2017). Patient contact was included as it was used as basis for weighting by Soares et al. (2011). Research awareness was used to identify experts who specialise in fall prevention but may not spend a significant proportion of their time with patients due to additional activities, such as work for professional bodies or committees. The question may also identify experts with more in-depth knowledge in the field, even for those experts who do spend the majority of their time with the target patient population.

The questions used to collect information about experts' substantive experience are provided in Appendix 3.2.

3.4.2. Normative expertise

Normative expertise refers to experts' ability to complete the elicitation exercise. As discussed in Chapters 1 and 2, it involves statistical/quantitative skills required to give coherent estimates, and to ensure that the priors accurately represent uncertainty. An assessment of the extent to which experts' priors represent their beliefs is typically based on statistical coherence of their priors, and feedback and validation (where the investigator describes what experts' priors imply about the value of the parameter and the expert expresses whether they agree or disagree with those descriptions) (P Garthwaite, J Kadane and O'Hagan, 2005). In the REFORM elicitation exercise experts were not fed back their priors (details on feedback and evaluation are provided in section 3.5.8) and so only statistical coherence was used as a measure of normative expertise.

The definition of coherence is provided in Chapter 5, when the results of the REFORM elicitation study are analysed.

3.4.3. Experts' ability to make accurate probabilistic assessments

The aim was to capture experts' ability to extrapolate their knowledge to assess their own uncertainty without capturing their knowledge of the subject matter. Experts' ability to make accurate probabilistic assessments was measured using a non-domain seed. Experts were asked to assess the number of days it rained in York every September. Expert' beliefs were elicited using the same question format, and the same elicitation technique as the remaining seed and target parameters in the study, details are provided in section 3.5 and Appendix 3.3.

The seed was chosen as it was, arguably, a quantity all experts are familiar with. An unbiased expert would give a range that includes the 'true' value, even if they have never been to York - they would assess rainfall in a region more familiar to them and how it compares to York and if neither are familiar to them they could simply indicate a wider range of plausible values.

Experts' priors on the non-domain seeds were scored using Confidence Interval Coverage Probability (CICP), as it is the only method identified in Chapter 2 that does not reward precision and penalise uncertainty. CICP accuracy scores were derived by comparing experts' priors to the average number of rainy days in September recorded by the Met Office between 1980 and 2010. Methods for deriving the CICP score are detailed in Chapter 4.

Chapter 2 suggested experts' ability to make judgements depends on their ability to interpret information. Experts' ability to draw inference was thus measured using four questions from the Watson-Glaser adaptive thinking test (Watson and Glaser, 2010) designed to measure individual's ability to draw inference. The test involves presenting experts with the summary of a study that explored risk factors for cardiovascular disease, then presenting them with four possible conclusions of the study. The experts are then required to express the extent to which the summary of the study supports each of the four conclusions. The text and the questions are provided in Appendix 3.4.

3.5. REFORM elicitation exercise design

The REFORM elicitation exercise aimed to recruit a relatively large sample of experts to allow comparison of priors elicited from experts with different characteristics. The resource constraints of the exercise meant that experts were not rewarded financially for taking part and so, to encourage participation, the exercise was designed to be completed in up to one hour. The remainder of section 3.5 details the exercise design.

The process is divided into the following nine sections:

1) What to elicit;
2) How to elicit;
3) Elicitation technique;
4) Expert recruitment: who is an expert, target sample size and recruitment strategy;
5) Background information;
6) Training;
7) Delivery;
8) Fitting, evaluation and feedback;
9) Aggregation.

3.5.1. What to elicit?

The parameters elicited were chosen on the basis of what was required in the analysis and what was reasonable to ask of experts, as recommended in the literature described in Chapter 1. The elicitation exercise had two objectives: to elicit experts' priors on the trial outcomes (so they can be used to explore whether experts' characteristics affect their accuracy) and to inform uncertainty around the cost-effectiveness of the intervention after the trial end point.

All parameters that measured the effect of the intervention were considered as candidates for elicitation. This meant that the observed effects during the trial could be used as seeds and the change in the treatment effect after the trial end point could be used as the target variable. The parameters considered for elicitation thus included trial outcomes that could affect the cost-effectiveness of the intervention (changes in the risk and rate of falls, the risk of having a fracture, HRQoL or costs of care after receiving the intervention for more than one year).

HRQoL and costs of care were excluded on the basis that they were unlikely to be fully observed by healthcare staff, and it is generally accepted that experts should not be asked about unobservable quantities, (Kadane and Wolfson, 1998) as highlighted in Chapter 1.

This narrowed down the list of parameters to the effect of the intervention on the proportion of fallers, the time to first fall, the rate of falls, and the risk of fractures.

As discussed in section 3.2.3, two physiotherapists specialising in fall prevention were consulted on which of the remaining parameters would be most appropriate for capturing any changes in falls behaviour in individuals that receive the intervention. After discussion with the physiotherapists it was concluded that interventions designed to prevent falls can have a number of effects. They can:

- reduce the probability of falling in all patients equally. This would be detected by measuring the time to first fall, the proportion of fallers or the rate of falls.
- They can reduce the frequency of falls in those who fall the most. This would reduce the rate of falls but not the proportion of fallers.
- They can reduce the severity of falls. If the intervention increased participants' confidence and mobility, their rate of falls could remain unaffected but the nature of falls could change with fewer falls resulting in fractures. Such change in the severity of falls would only be detected in the rate of fracture.

In order to capture any of the described effects of the intervention, the parameters chosen to be elicited were the proportion of fallers, the rate of falls and risk of fractures. Time to first fall was not included on the advice of one of the physiotherapists, who advised that exercise (one of the components of the intervention) can take time to work, and so the benefit may not be observed until after the patient has had their first fall, making it less sensitive to the treatment effect.

Table 3.1 summarises all parameters that were considered for elicitation and the basis for inclusion/exclusion.

Table 3.1. Criteria for choosing the elicitation parameter applied to the trial outcome measures.

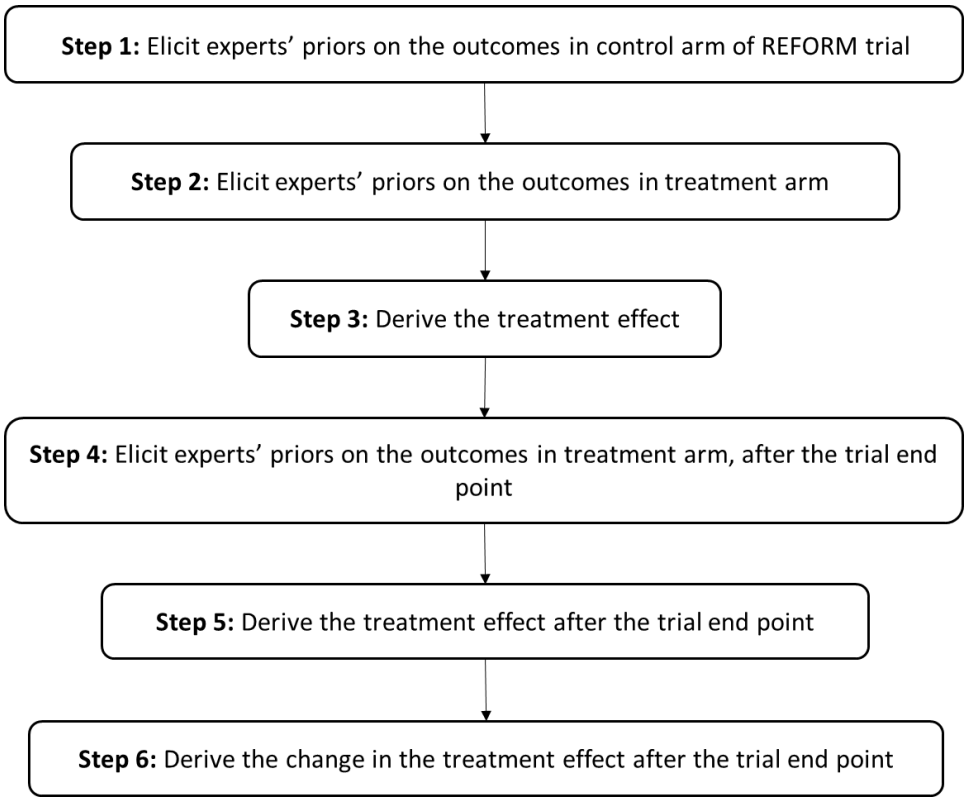
	Observable	Used in the CEA	Captures the treatment effect
Rate of falls	✓	✓	✓
Proportion of fallers	✓	✓	✓
Time to first fall	✓	✓	✓
HRQoL (EQ-5D)		✓	✓
Short Falls Efficacy Scale			✓
Fear of falling			
Activity of Daily Living			
Fracture rate	✓	✓	✓
Health service utilisation		✓	✓
Geriatric Depression Scale			

3.5.2. How to elicit?

Chapter 1 highlighted that there are multiple ways to elicit the same parameter. This section discusses how the seed and target parameters were elicited in this study.

The aim of the REFORM elicitation exercise was to elicit experts' priors on the treatment effect during the trial, and any change in the treatment effect after the trial end point. Treatment effect is not directly observable, only outcomes in those who have received treatment and those who have not. Therefore experts' beliefs were elicited for outcomes in patients who had received the intervention and in those who had not, and the resulting priors were used to derive the treatment effect. Similarly, change in treatment effect over time is not observable, and so experts' priors were elicited on the outcomes in patients who had received the intervention and continued to receive it after the trial end point, generating the treatment effect after the trial. The process is summarised in Figure 3.2.

Figure 3.2. Methods for eliciting the treatment effect and changes in the treatment effect.



Sections 3.5.2.1 - 3.5.2.4 describe each of the steps in Figure 3.2 in detail.

3.5.2.1. Step 1: Methods for eliciting outcomes in the control arm

Section 3.5.1 described three parameters that were chosen for elicitation: the rate of falls, the proportion of fallers and the risk of fracture. Methods for eliciting each parameter are described here, in turn.

Eliciting the rate of falls

The rate of falls is derived from the frequency distribution of falls in a population. The frequency distribution is skewed - approximately one in three (33%) people over the age of 65 have been reported to have a fall every year (Tinetti, Speechley and Ginter, 1988), around half of those (16.5% of over 65s) have been reported to fall more than once, (Nevitt, 1989; Tinetti and Speechley, 1989), and the probability of falling more than twice is even lower. Estimating the rate of falls requires weighting each possible outcome, in this case the number of falls, by its probability. Peterson and Miller (1964) studied experts' ability to determine the expected value of parameters with a skewed frequency distribution and found that their mean tends to be biased towards the median, i.e. they do not adjust for outliers sufficiently when calculating the expected value. Peterson and Miller's (1964)

findings raise concerns that estimating the expected rate of falls may be cognitively challenging, potentially leading to biased priors.

An informal, non-systematic scoping search of the elicitation literature was conducted to identify methods for eliciting probability distributions of rates – the methods used in the search are described in Appendix 3.5. However, no relevant studies were identified at the time of the search (January, 2016).

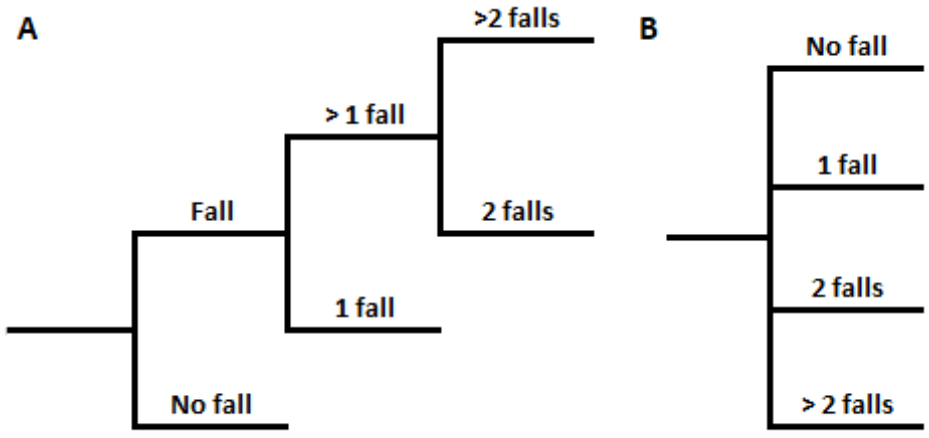
In the absence of evidence from the elicitation literature, a novel method for indirect elicitation of rates was designed. The indirect method involved deriving the rate of falls from elicited multinomial distributions of the number of falls.

Indirect elicitation of the rate of falls: multinomial distribution of the number of falls

For n independent trials where each trial results in one of k mutually exclusive outcomes, the multinomial distribution gives the probability of any particular combination of frequencies of outcomes (Evans, Hastings and Peacock, 2000). Applied to this example, the distribution will give the probability that, for example, of 100 patients exactly 90 do not fall, 7 suffer one fall, and 3 suffer two falls, or any other combination of outcomes.

Alternatively, multinomial distributions can be presented as a series of binomial distributions. Each pair of events (fall/no fall, 1 fall/>1 fall, 2 falls/>2 falls) is conditional on the preceding event occurring and their probabilities add up to 1. The two different presentations are shown in Figure 3.3.

Figure 3.3. Multinomial distribution presented as A) a sequence of binomial distributions, conditional on previous events, and B) four outcomes.



When eliciting multinomial distributions, assumptions around the relationship between different outcomes must be taken into account. (Soares *et al.*, 2013) The possible assumptions are the following.

- No correlation between outcomes, implying that the probability of having more than one fall P(x>1) is independent of the probability of falling P(x>0), where x is the number of falls. This assumption can lead to statistically incoherent priors where the probability of falling more than once is higher than the probability of falling, and so it was not used in the exercise.
- Conditional independence means that the probability of falling more than once, conditional on the patient having had at least one fall P(x>1| x>0), is independent of the probability of falling P(x>0).
- Conditional dependence means that the probability of falling more than once, conditional on the patient having had at least one fall P(x>1| x>0), depends on the probability of falling P(x>0).

Eliciting correlation between the conditional probabilities would require training experts in the concept of correlation. It would also require additional steps in the exercise. The additional training and elicitation steps would add an unfeasible workload to the exercise and so, for simplicity, conditional independence was assumed.

Eliciting multinomial distributions requires consideration of which outcomes the probability distributions should be elicited. In this example the outcomes are the numbers of falls (a patient can have no falls, one fall, two falls, etc). Studies on fall prevention often report the highest number of falls per participant per year to be more than 10 (Spink *et al.*, 2011). Eliciting the probability of over ten outcomes would have been cognitively and time intensive and so to reduce the burden on experts, possible outcomes (numbers of falls) were grouped into three categories. The first category was the probability of falling at least once. The probability of falling was also measured in the REFORM trial, and so it provided an additional seed. A second category aimed to capture changes in the frequency of falls in those participants who fall the most. A trial for a similar intervention was conducted in Australia, on a different population group (Spink, Menz and Lord, 2008; Spink *et al.*, 2011). The trial found that in the intervention arm no participant fell more than 6 times, while in the control arm the highest number of falls was 12. The second category was thus the probability of falling more than ten times. A third category was the probability of falling more than five times. Five was chosen as the mid-point of the first two categories. These

three categories were confirmed as reasonable by a physiotherapist specialising in falls prevention who advised on the structure of the exercise.

The elicited parameters were thus:

1) P(x >0),
2) P(x >5| x >0),
3) P(x >10| x >5).

Where x is the number of falls.

Experts' priors on each category were then used to derive the rate of falls. Ten thousand samples were drawn from each prior, and for each iteration joint probabilities P(x >5) and P(x >10) were derived using Equation 3.1.

$$P(x) = P(x|y) * P(y) \qquad \textbf{Equation 3.1}$$

The probability of observing 1, 2, 3, 4, 6, 7, 8, 9, 11 or more falls was predicted from the regression model in Equation 3.2. The probability of each number of falls was then calculated using Equation 3.3. The rate of falls was calculated by multiplying each number of falls by its corresponding probability.

$$\text{logit}(P(x > X)) = \alpha + \beta x \qquad \textbf{Equation 3.2}$$

Where logit (x) = log(x/(1-x); the logit transformation was used to ensure the predicted probabilities do not exceed the limits of the parameter (0-1).

$$P(x = X) = P(x > X - 1) - P(x > X) \qquad \textbf{Equation 3.3}$$

The methods for eliciting the rate of falls directly, and indirectly were compared in a pilot. The results are described in the next section.

Pilot to determine whether to elicit rates directly or indirectly

It was not clear which of the two methods for deriving rates was better: eliciting them directly or eliciting P(x>0), P(x>5|x>0) and P(x>10|x>5) and deriving the rate of falls using Equation 3.1-Equation 3.3. Kleinmuntz (1996) found that decomposing a problem and eliciting conditional probabilities leads to more accurate probabilistic predictions,

suggesting that deriving rates from priors on conditional probabilities could lead to more accurate estimates of the rate of falls than if they are elicited directly. However, the cited study only takes into account accuracy of point estimate forecasts; it is not clear how representation of uncertainty would differ between the two methods. This was explored by piloting both methods on two Research Fellows in health economics at the University of York. The pilot is described in detail in section 3.2.3 (Pilot 2). The summary statistics derived from the two participants are presented in Table 3.2.

Table 3.2. Summary statistics of experts' priors on the rate of falls elicited directly, and those derived from elicited multinomial distributions.

		Mean	Median	Min	Max
Expert 1	Rate (direct elicitation)	1.19	1.10	0.3	2.10
	Rate (derived from probabilities)	0.76	0.73	0.09	1.76
Expert 2	Rate (direct elicitation)	0.28	0.28	0.125	0.475
	Rate (derived from probabilities)	0.80	0.75	0.12	2.22

Directly and indirectly elicited rates were more consistent for Expert 1 than Expert 2.

There was no constancy in effect from using two different methods – for Expert 1 the directly elicited rates were higher than the indirectly elicited rates and their uncertainty was comparable, while for Expert 2 the directly elicited rates were much lower and more certain.

Rates derived using the indirect method appear to be similar for the two experts, whereas directly elicited rates were very different. Furthermore, both experts expressed that they found the indirect method more intuitive and so this was chosen for the exercise.

The described method for deriving probabilities of each number of falls had the potential to result in probabilities that deviated from experts' expressed beliefs. For example, adding the derived probabilities for falling 1-5 times could indicate different probabilities to those expressed by the expert. Sensitivity analysis was conducted to evaluate whether the

regression model accurately represents the elicited priors. The results are reported in the analysis, in Chapter 4.

The next section describes the quantities used to formulate the question.

Quantities used to elicit the rate of falls

As concluded in the last section, the chosen parameters to elicit were three sets of probabilities: P(x>0), P(x>5|x>0) and P(x>10|x>5). Chapter 1 highlighted that priors on probabilities can be elicited as different quantities (probabilities, proportions, percentage, relative frequency, odds and natural frequency) and that using different quantities can lead to different priors being elicited (O'Hagan *et al.*, 2006).

Chapter 1 also highlighted that experts tend to find frequencies the most intuitive when representing interactions between different quantities. Rate of falls was derived from elicited probabilities and conditional probabilities in two arms and at two time points (as will be described later in this section). In order to keep the questions clear, the problem was presented using relative frequencies.

When frequencies are used in elicitation, a choice must be made about the value for the denominator –i.e. the sample size for which they have to estimate the number of patients who will suffer at least one fall, more than five falls, etc. In this exercise the denominator was chosen to be a multiple of 10, for simplicity, and greater than 100 to allow experts to express probabilities less than 1%. The denominator was chosen to be 1000.

The resulting question format is presented in Box 3.1.

Box 3.1. Example question for eliciting the rate of falls in control arm. [L] and [M] are the modes in experts' priors in previous questions.

Question *1*

'Consider 1000 patients over the age of 70, randomly selected in the UK, who participate in the trial but DO NOT RECEIVE the intervention.'

'Out of 1000 people who participate in the trial and DO NOT receive the intervention, how many do you think will have a fall in one year, during the trial?'

Question *2*

'Now let's assume that out of 1000 patients exactly [L] fall at least once.' (Note that this is the number you stated to be the most likely in the grid above.)

'How many out of these [L] individuals do you think will fall MORE THAN FIVE TIMES in one year?'

Question 3

'Now let's assume that out of 1000 patients exactly [L] fall at least once, and [M] fall more than five times.'

'How many out of these [M] individuals do you think will fall MORE THAN TEN TIMES in one year?'

Eliciting the proportion of fallers

As described in the previous section, the proportion of fallers was elicited in the form of frequencies, and used to generate the rate of falls.

Eliciting the probability of fracture after a fall

The probabilities of having a fracture after a fall was elicited in form of odds. Odds were chosen for two reasons: 1) to compare experts' priors when assessing uncertainty around different types of quantities, and 2) because the risk of fracture after a fall in the literature tends to be reported as odds and so this quantity was thought to be more intuitive for experts. The resulting question format is presented in Box 3.2.

Box 3.2. Example question for eliciting the odds that a fall will result in a fracture in control arm.

'This section aims to find out about the severity of falls.'

'For patients who do not receive the intervention, one in how many falls, on average, do you think will result in a fracture?'

3.5.2.2. Steps 2 and 3: Methods for eliciting outcomes in the treatment arm of the REFORM trial and deriving the treatment effect

As described at the beginning of this section (3.5.2) the proportion of fallers, rate of falls and the odds of fracture were elicited in the control arm and the treatment arms separately. The treatment effect was derived by calculating the relative risk for the elicited probabilities[8], the rate ratio[9] for the rate of falls derived from experts' priors, and odds ratios[10] for the odds of having a fracture after a fall.

When eliciting probabilities in the treatment arm, assumptions about its relationship with the probability of falling in control arm can affect the how they are elicited. As discussed in 'Step1, three different assumptions can be made: independence (P(x) independent of P(y)), conditional independence (P(x|y) independent of P(y)) and conditional dependence (P(x|y) depends on P(y)).

As discussed earlier in the section, eliciting correlation between conditional probabilities requires teaching experts about correlation. This is not feasible in this exercise and so conditional independence was assumed. Potential implications of this assumption are further discussed in Chapter 4.

The resulting question format used to elicit the rate of falls in control arm is shown in Box 3.3.

[8] $Relative\ risk = \frac{P_T(x>0)}{P_C(x>0)}$

[9] $Rate\ ratio = \frac{Rate\ of\ falls_T}{Rate\ of\ falls_C}$

[10] $Odds\ ratio = \frac{Odds\ of\ fracture_T}{Odds\ of\ fracture_C}$

T subscript indicates the probability of falling and the rate of falls in treatment arm.
C subscript indicates the probability of falling and the rate of falls in control arm.

Box 3.3. Example question for eliciting the rate of falls in treatment arm. [L] and [M] are the modes in experts' priors in previous questions.

Question *1*

'Consider 1000 patients over the age of 70, randomly selected in the UK, who participate in the trial for a year and are offered to continue with the intervention for [T] years after the trial.'

'After this period, how many of them do you think will have a fall in one year?'

Question *2*

'Now let's assume that out of 1000 patients exactly [L] fall at least once.'

'How many out of these NA individuals do you think will fall MORE THAN FIVE TIMES in one year?'

Question 3

'Now let's assume that out of 1000 patients exactly [L] fall at least once, and [M] fall more than five times.'

'How many out of these NA individuals do you think will fall MORE THAN TEN TIMES in one year?'

3.5.2.3. Step 4: Methods for eliciting outcomes in treatment arm after the trial end point

The treatment effect after the trial end point was elicited by asking experts to assume that the rate of falls in the control arm does not change over time, and then asking them what they believe would happen in the treatment arm over time. The information they were presented with is shown in Box 3.4.

In order to elicit experts' beliefs about the treatment effect after the trial end point, first, a series of multiple choice questions (MCQs) were asked to determine whether experts believed the treatment effect could change (see Figure 3.4). At the point at which they believed that the treatment effect could change over time, a second set of probability distributions was elicited (a second time point).

Box 3.4. Assumptions about the rate of falls and the treatment effect after the REFORM trial end point provided to experts.

'This section is about what happens after the trial has finished, given that patients remain enrolled in the intervention.'

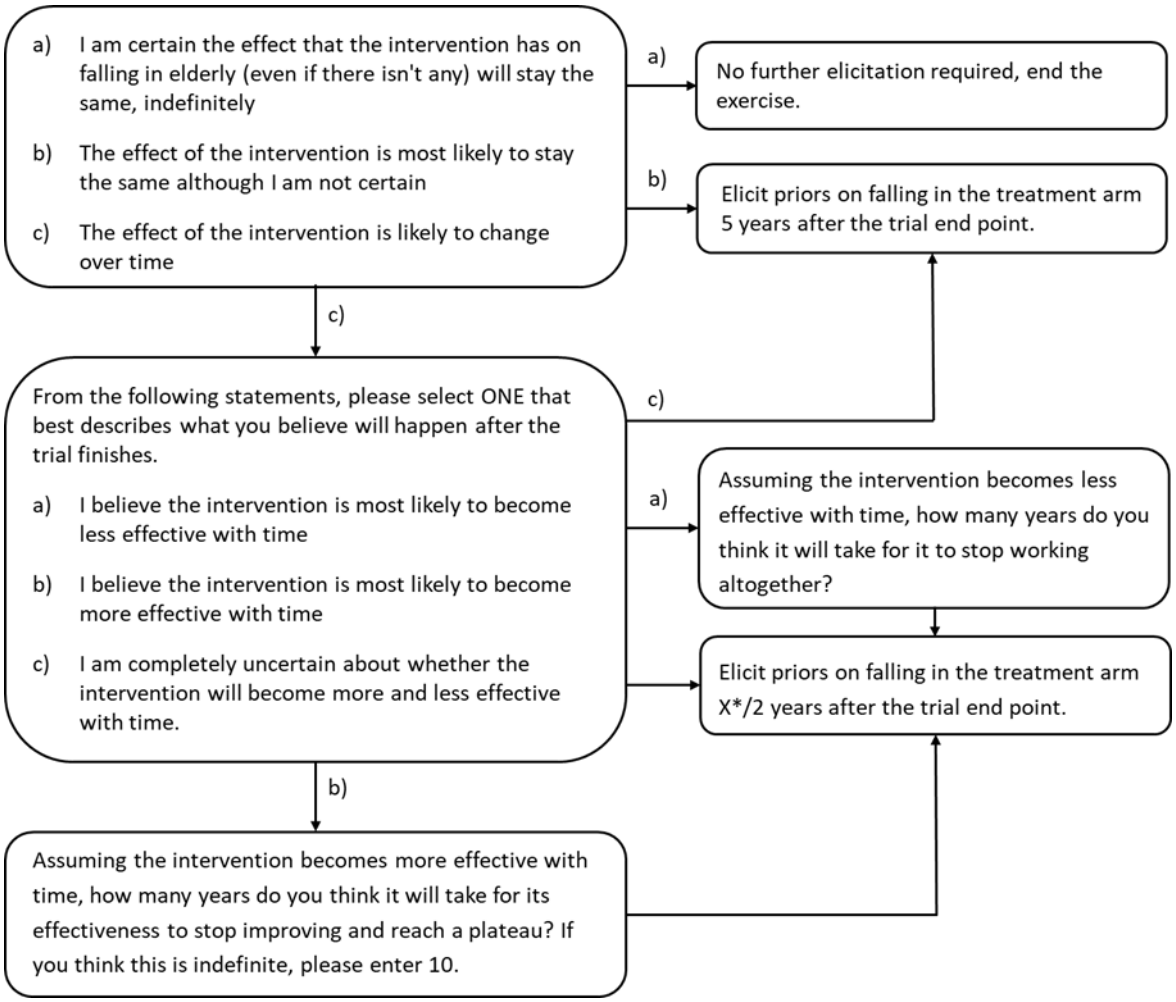
'Let's assume that all patients who were in the trial were offered to continue with the intervention in the same way: they continue to wear the appropriate footwear, foot orthoses (if required) and do self-directed exercise. Let's also assume that they no longer have to report their falling behaviour and they only see their podiatrist if required. We would like to know whether you believe that the intervention would have the same effect (if any) on the probability of falling as it did in the trial.'

'Please take into account any factors that can influence the effect of the intervention on the risk of falling. This could be patient compliance after the trial ends, possibility that effectiveness of treatment will wear off, the possibility that the intervention will become more effective with time, or anything else that you think may be relevant.'

'From the following statements, please select ONE that best describes what you believe will happen after the trial finishes.'

The experts were then presented with all their previous answers, as shown in Box 3.5, and their priors on $P(x>0)$, $P(x>5|x>0)$ and $P(x>10|x>5)$, and the odds of fracture at the second time point were elicited.

Figure 3.4. Algorithms used to determine the second time point for which probabilities would be elicited.



* X = time point indicated by the expert.

Box 3.5. Assumptions about the rate of falls observed in the REFORM trial provided to experts. [T] is the second time point in years.

'Questions in this section refer to what you think will happen [T] years after the trial finishes, provided all patients who DID RECEIVE the intervention in the trial (the treatment arm) are offered to continue with the intervention.'

'We will assume that the risk of falling remains the same for those people who do not receive the intervention. We would like to know what you think will happen to those individuals who DO RECEIVE the intervention (the treatment arm) in the trial and are offered to continue with it after the trial has finished.'

'In your responses, assume the following about trial outcomes:'

'Out of 1000 patients who DO NOT receive the intervention [X] *fall more than once,* [Y] *fall more than five times and* [Z] *fall more than ten times.'*

'Out of 1000 patients who DO receive the intervention [A] *fall more than once* [B] *fall more than five times and* [C] *fall more than ten times.'*

'You can look back at these numbers at any point while answering the next question by clicking back on the 'Treatment after the trial' tab on the side panel.'

'You can also change your responses to previous questions by clicking back on relevant tabs, but please note that these are not 'correct' answers. These numbers were obtained from your previous responses and are different for every expert who completes this exercise. If you chose to change your answers, please make sure you save the new answers.'

3.5.2.4. Steps 5 and 6: Deriving the treatment effect after the trial end point, and the change in treatment effect over time

The treatment effect after the trial end point was derived in the same way as during the trial, as described in Steps 2 and 3. The change was derived using Equation 3.4.

$$\Delta TE = \frac{TE_{t2} - TE_{t1}}{TE_{t1}}$$ **Equation 3.4**

Where ΔTE = change in the treatment effect,

TE_{t1} indicates treatment effect during the trial,

TE_{t2} indicates treatment effect after the trial end point.

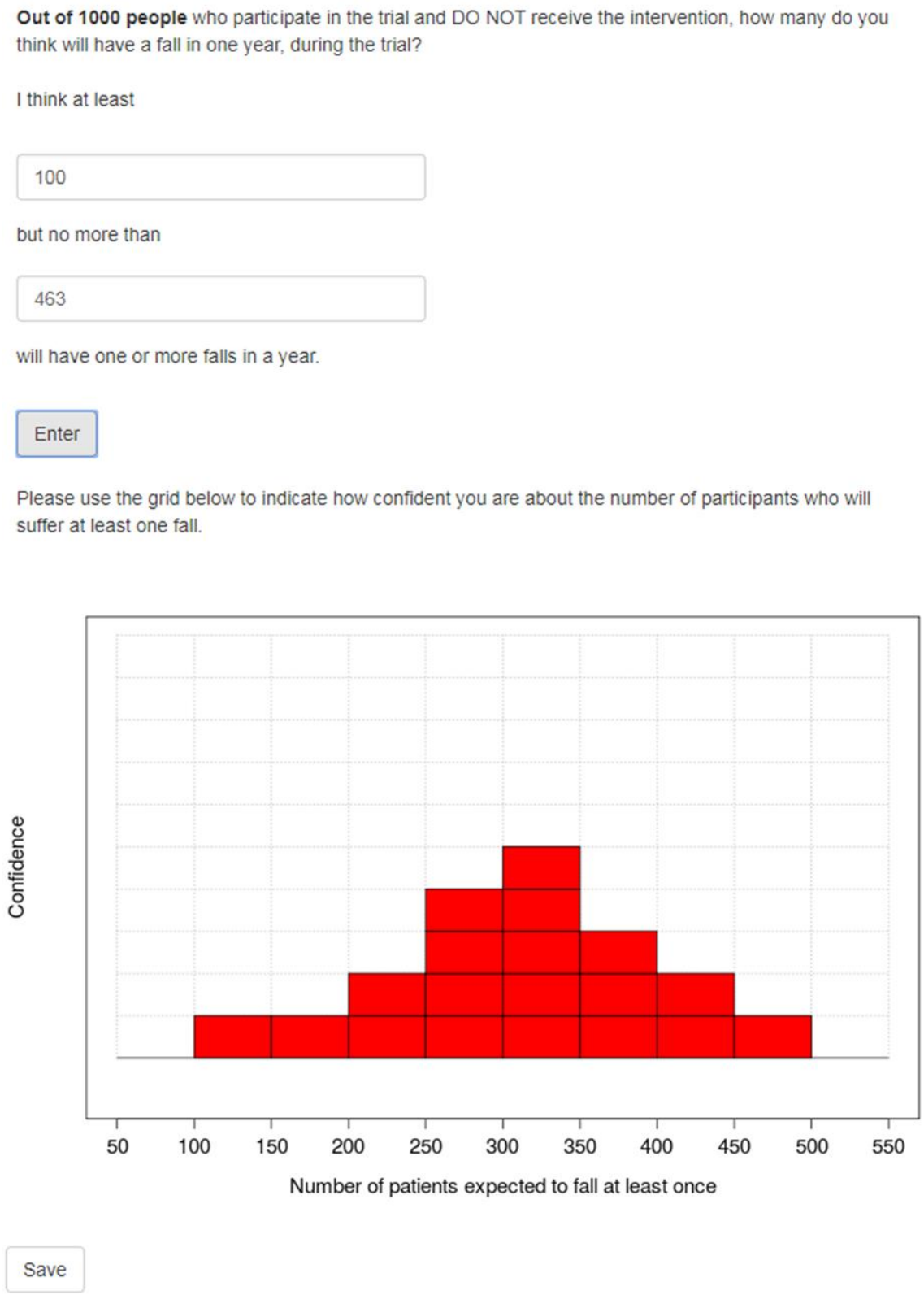
Detailed methods for deriving and analysing each elicited parameter are provided in Chapter 4.

3.5.3. Elicitation technique

Given the questions specified above, the next step is to determine how to elicit the probability distributions for these. The methods for eliciting distributions are described in Chapter 1. The fixed interval method and the histogram technique are the most widely used methods in HTA (Grigore *et al.*, 2013). Both methods were considered for use in this study. The methods were piloted on one lay participant, as described in section 3.2.3 (Pilot 1). When testing the fixed interval method the elicited prior followed a U-shaped distribution (the probability density was the highest at the edges or the range) that was judged by the investigator to be implausible. The participant verbally expressed confusion with the method. They found the histogram technique to be more intuitive and the distributions they provided were more plausible (bell-shaped). The histogram technique was thus adopted in the exercise.

For each elicited parameter, experts were first asked to give the minimum and maximum plausible values to avoid anchoring, (Kadane and Wolfson, 1998) and to narrow down the range of values on the chips and bins grid. The x-axis on the grid had a range of parameter values determined by the minimum and the maximum suggested by the expert. The bin width was always 1, 2, 5, 10, 20, 50 or 100, whichever of these resulted in as close to 10 bins as possible. Two extra bins outside the experts' range were added, unless they were outside the limits of the parameter. The grid was always 10 bins high, and there was no limit on how many chips could be added to each grid. An example question is shown in Figure 3.5.

Figure 3.5. An example question used to elicit experts' uncertainty in the REFORM elicitation study.



Experts were trained on the scale of the y-axis - entering more chips indicates increasing confidence and the units are relative to other bins (see Appendix 3.7 for details).

3.5.4. Who is an expert?

The following section considers the profile of experts who are considered to have substantive expertise for assessing the effects of a podiatry intervention on the risk of falls. The exercise required an understanding of falls behaviour in the elderly and how much foot and ankle health contributes to the risk of falls. The intervention was delivered by podiatrists; however, they are unlikely to observe the effects of the intervention. Additional professions were thus considered. The aim was to recruit clinicians who satisfy the following criteria:

- have a good knowledge of foot and ankle physiology,
- understand the risk factors for falling,
- have a knowledge of fall prevention interventions and evidence behind them,
- have experience in delivering behavioural interventions (and knowledge of how patients respond to them).

The professions were first identified by reading literature on fall prevention and observing professions of authors who published in the field. This led to identifying physiotherapists and geriatricians as potential experts. After seeking advice from physiotherapists who work in the field of fall prevention (described in section 3.2.3) it was discovered that each hospital in the UK has a fall prevention team. The structure of teams varies (often depending on the size of the Trust) but tends to be operated by geriatricians, physiotherapists, nurses and occupational therapists (OTs). Experts were thus defined as clinicians who work in one of these professions and specialise in preventing falls or treating patients who have suffered fall related injuries.

It was noted that experts in these four professions are most likely to see patients with history of falling and fall related injuries. In order to capture beliefs of experts who see a broader population of patients at risk of falling, general practitioners (GPs) were also targeted.

Finally, health researchers whose research focuses on fall prevention were included to explore the effect of having less clinical experience but thorough knowledge of relevant literature on experts' priors.

Participants were identified via the following four avenues.

- Contacting clinicians who have published research in the field of fall prevention, in particular any studies that evaluate the effect of exercise and foot and ankle health on falling behaviour.
- Contacting members of the appropriate professional bodies including, but not restricted to, the Chartered Society of Physiotherapy, ASPIRE (a special interest group for physiotherapists working with elderly patients) and the British Geriatrics Society.
- Contacting individual fall clinics/departments in NHS trusts in England. Trusts will be chosen based on recommendation or geographical location and regional patient characteristics to give a heterogeneous sample of experts, as recommended in the literature (O'Hagan et al., 2006).
- On recommendation by contacts gained through these bodies and by other research staff at the University of York.

All experts were contacted via e-mail or phone, using publicly available details or those provided by those contacts who recommend them.

The target sample size was 30 to 50 experts. The sample size was decided to include a representative sample of experts from each profession. The upper limit of 50 was based on feasibility.

3.5.5. Background information

Experts' were asked to express their uncertainty about the expected trial outcomes. They were provided with background information on why the elicitation exercise was conducted (shown in Appendix 3.3) and information about the trial. The latter included information about the intervention, outcome measures, data collection and inclusion and exclusion criteria in the trial (see Appendix 3.6 for details).

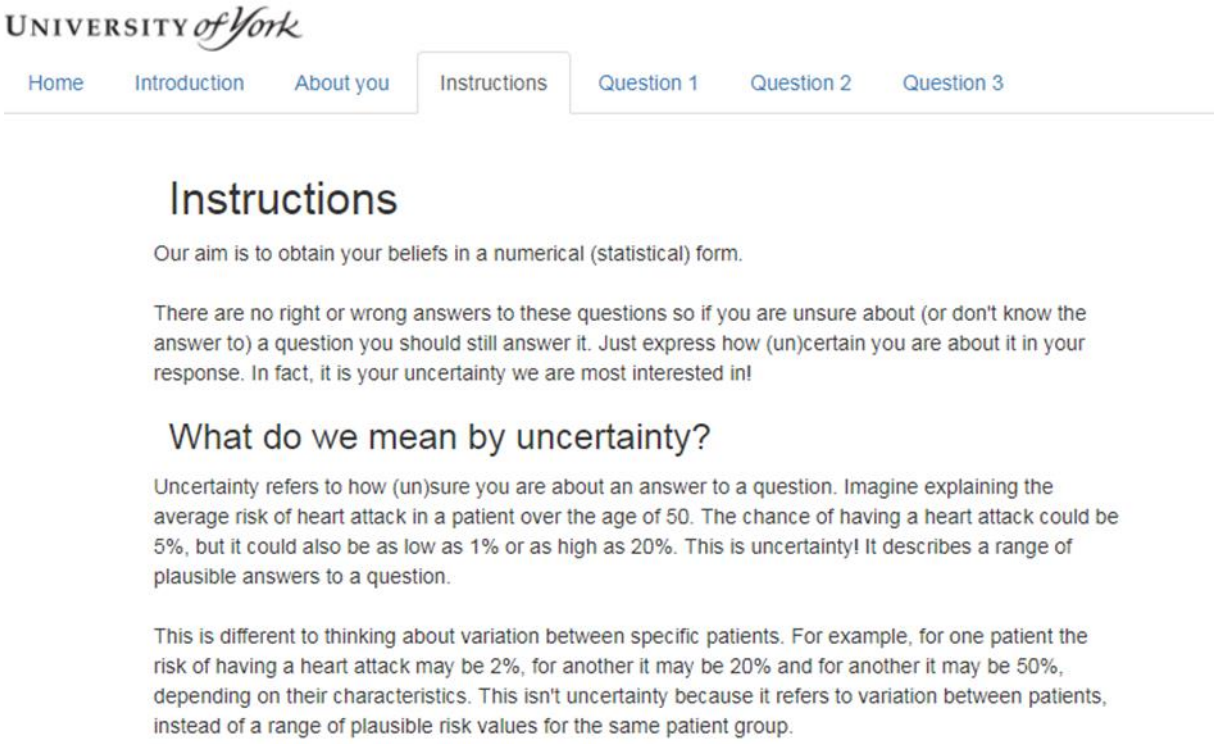
3.5.6. Training

Training was provided at the beginning of the exercise and consisted of two components: explaining uncertainty and teaching experts to use the histogram technique. Sections 3.5.6.1 and 3.5.6.2 describe the two components, respectively.

3.5.6.1. Explaining uncertainty

The aim of elicitation is to elicit uncertainty around parameter values, rather than heterogeneity or variability, as recommended by Bojke et al. (2017). While this is to some extent affected by the question format, training included a discussion about the difference between uncertainty and variability to help experts understand the difference. The information provided is shown in Figure 3.6.

Figure 3.6. Training on the difference between uncertainty and variability.



3.5.6.2. Training in elicitation technique

Experts were shown how to use the histogram technique, and were then provided with examples of a uniform distribution, a normal distribution and complete certainty with explanations of what they imply about experts' beliefs. The experts were then asked to complete a practice example as many times as they required. The experts could not move onto the next section without completing the training. The training material is included in Appendix 3.7.

3.5.7. Delivery

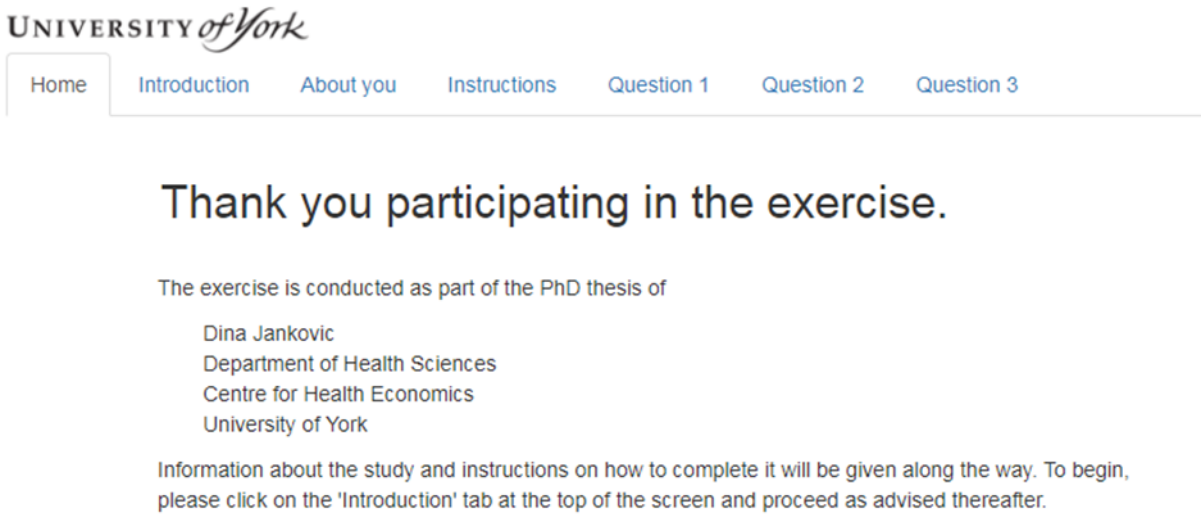
The exercise was conducted using an R-based, web app. The tool was an extension of the MATCH code developed by Morris et al. (2014) modified specifically for the REFORM elicitation exercise.

The tool was designed for independent completion. All experts were given the choice of completing the exercise on their own or with the help of the investigator, although the latter was encouraged. If they chose to complete the exercise with the investigator, the exercise was completed at pre-set times and help was available on the phone or in person. When more than one participant was available at the same time, in the same region, the exercise was completed with all participants in one room, where the investigator introduced the exercise and went through examples with everyone as a group. The experts then provided their opinion separately, without consulting each other. To do this, the exercise was either conducted in a computer cluster, or on laptops and touchscreen devices provided by the investigator.

The tool guided experts through seven tabs: 1) information about the investigator, 2) information about the project and why they had been invited to participate, 3) a questionnaire about their professional experience, 4) training, 5) a non-domain seed question, 6) information about the trial and questions about trial outcomes, 7) questions about the effectiveness of the intervention after the trial finishes.

The homepage is shown in Figure 3.7.

Figure 3.7. Homepage of the REFORM elicitation tool.

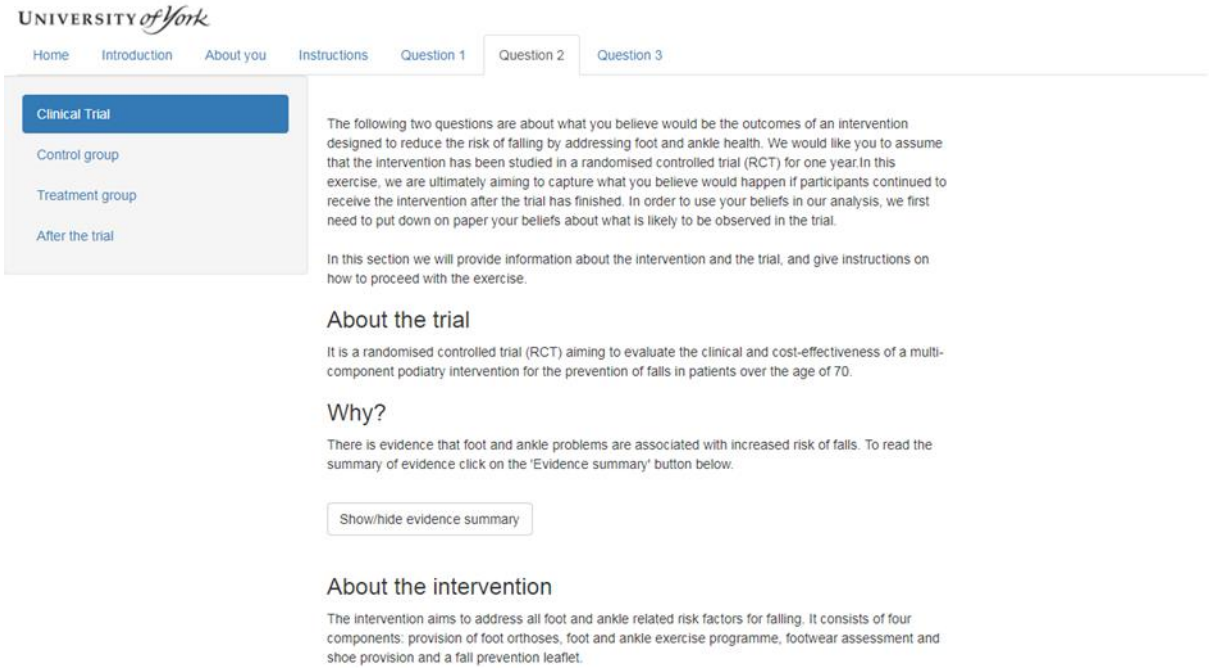


The 'Introduction' tab provided information about the investigator and about the project. It included an explanation of what elicitation is, why the study was being conducted and why the experts were included in the study. The print out of the section is provided in Appendix 3.8. Questions about experts' professional experience in the 'About you' section were discussed in section 3.4, while the contents of the 'Instructions' tab were provided in section 3.5.6.

Details about the elicited parameters were discussed in sections 3.5.1 and 3.5.2. In summary, experts' priors were elicited on 9 to 13 parameters (depending on their beliefs about the change in treatment effect after the trial end point). In order to make the workload manageable, the questions were separated into three sections: the non-domain seed about the number of rainy days in York (Question 1 in Figure 3.7), those about REFORM trial outcomes (Question 2 in Figure 3.7) and those about the probabilities of falling and the odds of fracture after the trial end point (Question 3 in Figure 3.7).

The layout of Question 2 is shown in Figure 3.8. The tab contained three side panels: background information about the trial (provided in Appendix 3.6), one page with questions about the probabilities of falling in control arm and one in treatment arm, and one page with the multiple choice questions about long term effects of the intervention, after the trial finishes.

Figure 3.8. Screenshot of the question about trial outcomes.



It was not possible for experts to start the exercise without completing the training (Question 1 was enabled when experts saved their practice answer), and each subsequent question was enabled when their answer to the previous question was saved. Because the exercise was on occasions completed without the investigator, it was designed to prevent statistically incoherent priors. For example, experts could not enter figures outside the given range.

The tool was piloted on seven podiatrists who delivered the intervention in the REFORM trial (Pilot 3). The participants were chosen based on the ease of recruitment (the participants were attending a presentation at the University of York), quick delivery (as they did not require background information on REFORM trial), and the assumption that their quantitative skills were comparable to those of experts in the REFORM elicitation study. The participants were delivered a group training session in a presentation that covered the contents of the 'Instructions' tab (described in section 3.5.6). The experts used the practice example in the exercise and completed the exercise independently. The pilot was conducted to test the exercise for sense, clarity and ease of use. One technical challenge arose in the pilot: several experts entered the minimum and maximum values in the wrong field (they entered the minimum in the 'maximum' field and vice versa), precipitating an error message. As result, the tool was updated to automatically select the smaller value as the minimum and the larger value as the maximum.

3.5.8. Fitting, evaluation and feedback

Chapter 1 discussed that experts' priors can be fitted using parametric probability distributions, prior to use in the cost-effectiveness model. If distributions are fitted, the investigator must decide whether to fit during the elicitation exercise (so that the fit can be validated by the expert) or after the exercise. Furthermore, if distributions are fitted after the exercise, they can be fitted before or after aggregating experts' priors. The benefits and limitations of each approach were discussed in Chapter 1 (section 1.3.4.5). Distributions were not fitted during the exercise because training experts on interpreting probability distributions is time consuming, and so was not feasible during this exercise. The priors were only fitted to probability distributions after aggregation - the rationale and detailed methods are provided in the analysis, in Chapters 5 and 6.

3.5.9. Aggregation

As discussed in section 3.1, the objectives of this study were to explore the extent to which experts' characteristics explain their priors, and to compare different methods for mathematical aggregation. In order to achieve the study objectives priors were elicited individually, then aggregated mathematically. Several aggregation methods were compared in the analysis – these are described in detail in Chapter 6.

3.6. Summary of Chapter 3

The REFORM elicitation study was designed to evaluate and compare different weighting methods. The study has two specific objectives: 1) to explore factors that affect experts' priors, and 2) to compare the impact of different weighting methods in an applied case study in HTA.

Several studies have assessed characteristics that affect experts' priors, or compared performance-weighted priors to unweighted ones – these studies tend to analyse results of elicitation exercises reported in databases, retrospectively. For example, Nemet et al. (2017) measured the effect of experts' characteristics and elicitation process design on the width of experts' 80% confidence interval in their judgments about future energy technologies. The studies evaluating and comparing different weighting methods are generally based on the applied exercises in the TU Delft. (Goossens, 2008b; S Lin and Cheng, 2009; Flandoli *et al.*, 2011; Colson and Cooke, 2017) Using findings from databases can limit the characteristics and weighting methods that can be compared.

The REFORM elicitation study is the first study prospectively designed to evaluate and compare weighting methods in CEDM.

In order to fulfil the first objective - to explore factors that affect experts' priors – the study needed to recruit a relatively large sample of experts (30-50, compared to the 8.3 recruited for elicitation in CEDM on average) and so ease of completion was an important factor in the study design.

The elicitation exercise was designed to be delivered remotely, in no longer than one hour.

The aim of the elicitation exercise was to elicit experts' beliefs on the treatment effect of the podiatry intervention evaluated in the REFORM trial, and the temporal change in the treatment effect. Chapter 1 highlighted that there is a lack of understanding of how best to elicit different types of quantities (see section 1.3.4.1 for details). A decision was made to elicit both quantities indirectly, based on practice in previous exercises that elicited the treatment effect (Bojke *et al.*, 2010; Soares *et al.*, 2011). The treatment effect was assumed to be independent of the rate of falls and risk of fractures (as described in section 3.5.2).

The elicitation methods required experts' beliefs on the rate of falls to be elicited in those patients who receive the intervention and those who do not. The skewed distribution of the frequency of falls made it a difficult parameter for experts to assess. There were no identified studies for deriving rates and so a novel method for eliciting rates indirectly was derived where a series of binomial distributions was elicited and conditional probabilities of different outcomes (number of falls) were assumed to be independent. It is unclear which of the two methods (direct or indirect elicitation of rates) is better and so the decision to use the indirect method was based on the results of a pilot where the participants expressed the indirect method to be more intuitive, and led to more comparable results between them. It is not clear whether this also makes it a better method.

Alternative methods for eliciting multinomial distributions exist – for example eliciting the multinomial distribution for the frequency of falls and correlation between conditional probabilities of different number of falls,(Clemen, Fischer and Winkler, 2000) but the methods require extensive training and active guidance by the investigator (Bojke *et al.*, 2017) and so were not feasible in this study.

The plausibility and implications of the assumptions made in the elicitation process are discussed in further detail in Chapters 4, 5 and 6 when the results of the elicitation exercise are analysed and discussed.

Chapter 4. Interpreting the results from the REFORM elicitation exercise

4.1. Introduction

Chapter 3 described the REFORM elicitation study conducted to compare different methods for deriving weights in elicitation. The study elicited experts' beliefs on the following quantities (9-13 depending on experts' beliefs about the trajectory of the treatment effect after the trial endpoint):

- One non-domain seed regarding rainfall in York;
- Eight domain seeds (outcomes measured in the REFORM trial);
- Four target parameters about the treatment effect after the trial end point, provided experts believed that the treatment effect could change over time.

This chapter presents an overview of the results of the elicitation exercise. Section 4.2 describes how experts' elicited quantities were used in the analysis, section 4.3 describes the sample of experts who took part, section 4.4 gives an overview of what experts' priors suggest about their beliefs, and section 4.5 evaluates the elicited priors, and the methods used to analyse them. Section 4.6 then provides a summary of the findings.

4.2. Methods to decode experts' priors

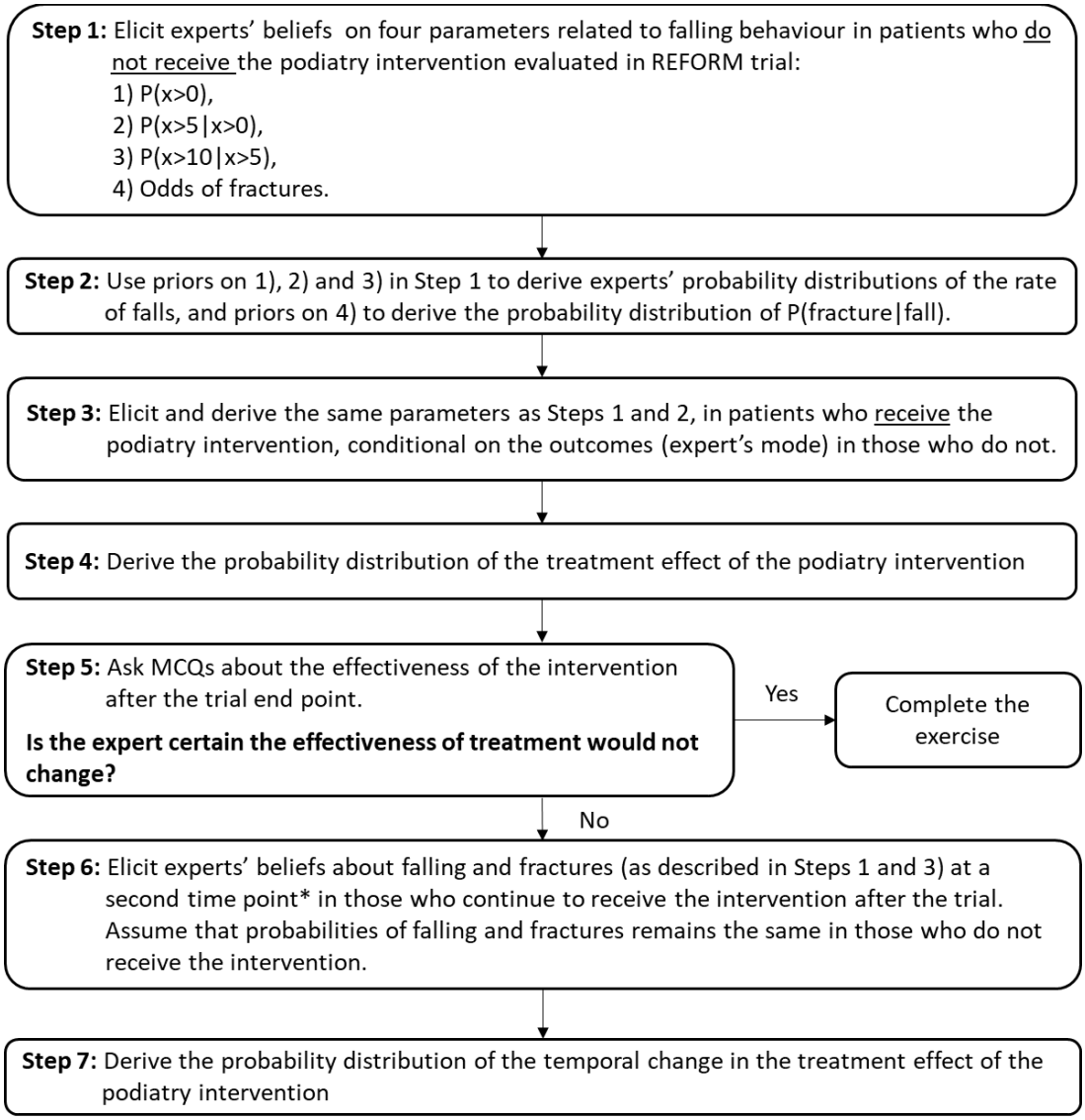
The REFORM elicitation study elicited one non-domain quantity, and 12 domain-specific quantities regarding falls and fractures in participants in the REFORM trial.

The non-domain seed was a question about the average number of rainy days in York every September. The seed was elicited and analysed in terms of frequencies (number of days out of 30). As discussed in Chapter 3, experts' accuracy in assessing rainfall in York was scored using CICP, and the scores are used:

- To explore whether experts' characteristics can predict their scores in non-domain seeds (Chapter 5);
- To explore whether experts' accuracy in assessing non-domain seeds predicts experts' accuracy on the domain seeds (Chapter 5);
- To derive weights (Chapter 6).

The 12 quantities elicited about falls and fractures in participants in the REFORM trial are summarised in Figure 4.1. These parameters were also observed in the REFORM trial, and so they are used as seeds in Chapter 5 where the effect of experts' characteristics on their priors is assessed, and in Chapter 6, where they are used to derive weights for experts' priors on the target parameter. The parameters elicited and derived in Steps 5-7 are used to derive the change in the treatment effect after the trial end point; their value was not observed in the REFORM trial and so they represent the target parameters used in the CEA in Chapter 6.

Figure 4.1. Elicited quantities regarding the treatment effect of the intervention evaluated in the REFORM trial.



*The second time point was determined by experts, as described in Chater 3.

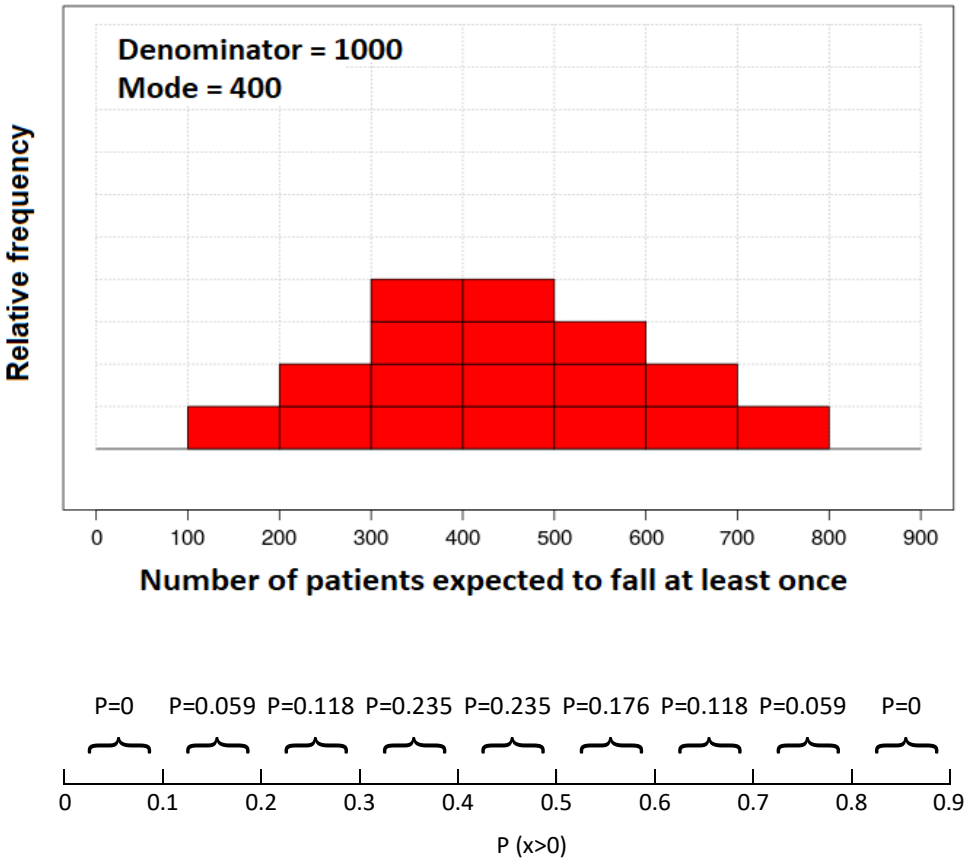
Figure 4.1 also shows that there were three different quantities elicited: probabilities, conditional probabilities and odds. These were used to derive the rate of falls, the treatment effect, and the temporal change in the treatment effect. Sections 4.2.1- 4.2.5 describes how each of the said quantities were derived from the elicited priors.

Finally, Chapter 1 discussed that evaluation is an integral part of the elicitation process, aiming to determine 'how well' the elicitation has been done and so section 4.2.6 describes the methods used to evaluate this exercise.

4.2.1. Priors on the probability of falling

Experts' priors on falling were elicited as relative frequencies; these were then converted into probabilities. The proportion of fallers was derived by dividing experts' relative frequencies by 1000 (the denominator for their frequencies), as shown in Figure 4.2.

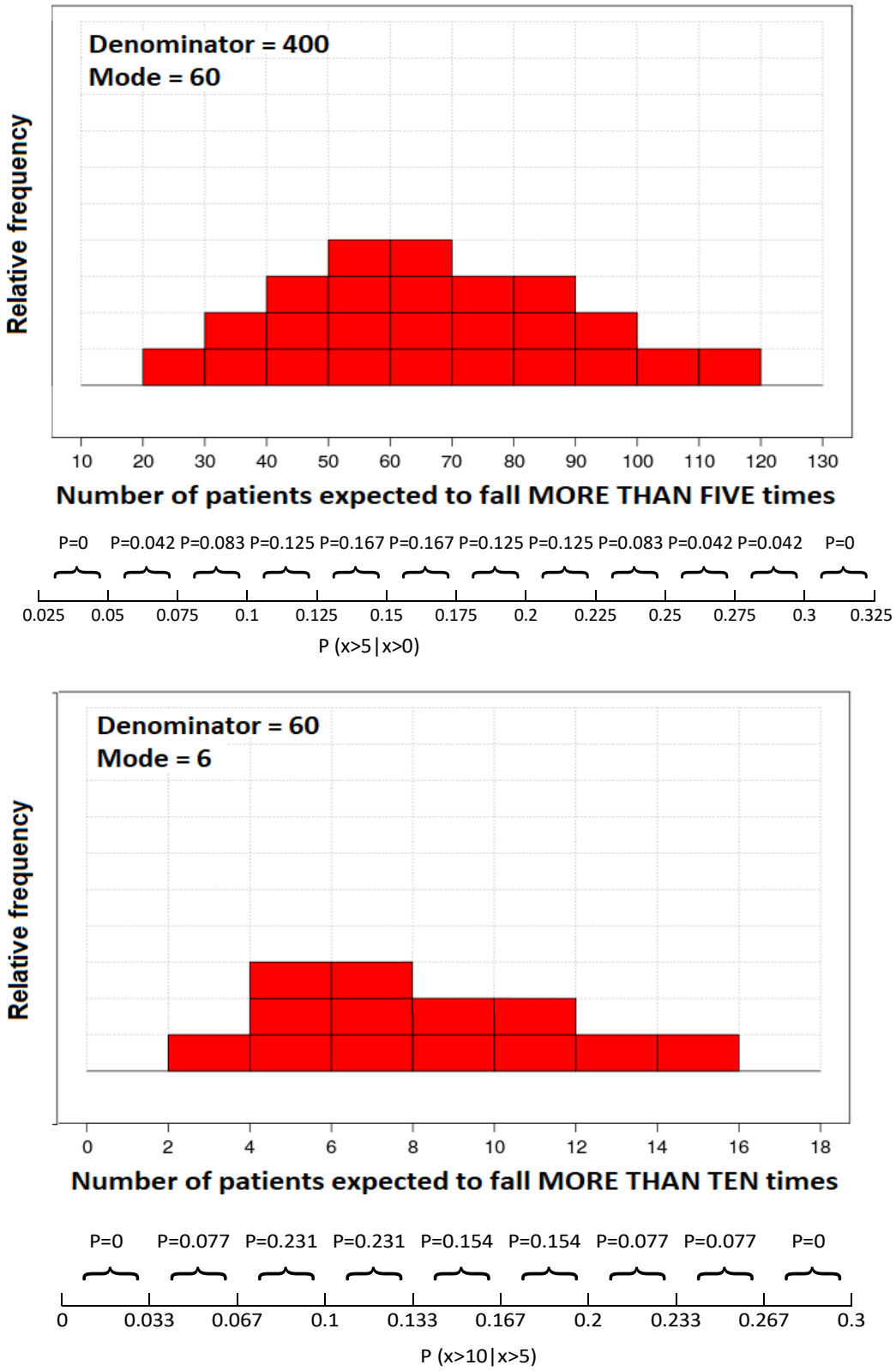
Figure 4.2. Elicited priors on the frequency of falling and the resulting probability summaries used in the analysis.



Experts' priors on the number of patients expected to fall more than five times were conditional on their mode number of patients expected to fall at least once and so P(x>5|x>0) was derived by dividing the frequencies elicited from experts' by their mode number of patients expected to fall at least once. Similarly, P(x>10|x>5) was derived by

dividing experts' frequencies by their mode number of patients who would suffer more than five falls. The methods are demonstrated in a hypothetical example in Figure 4.3.

Figure 4.3. Experts' priors on the frequency of multiple falls and the resulting probability summaries used in the analysis.



4.2.2. Deriving the rate of falls

Methods for deriving rates were described in Chapter 3. In summary, 10,000 random samples were drawn from experts' priors on P(x>0), P(x>5|x>0) and P(x>10|x>5). Probability distributions were not fitted to priors elicited from individual experts and so when sampling, the probability distribution in each bin was assumed to be uniform. For each iteration joint probabilities P(x>5) and P(x>10) were derived using Equation 3.1 in Chapter 3, and the probabilities of observing more than 1, 2, 3, 4, 6, 7, 8, 9, 11, etc. falls were predicted from the regression model in Equation 3.2. The probability of each number of falls was then calculated using Equation 3.3. The rate of falls was calculated by multiplying each number of falls by its corresponding probability.

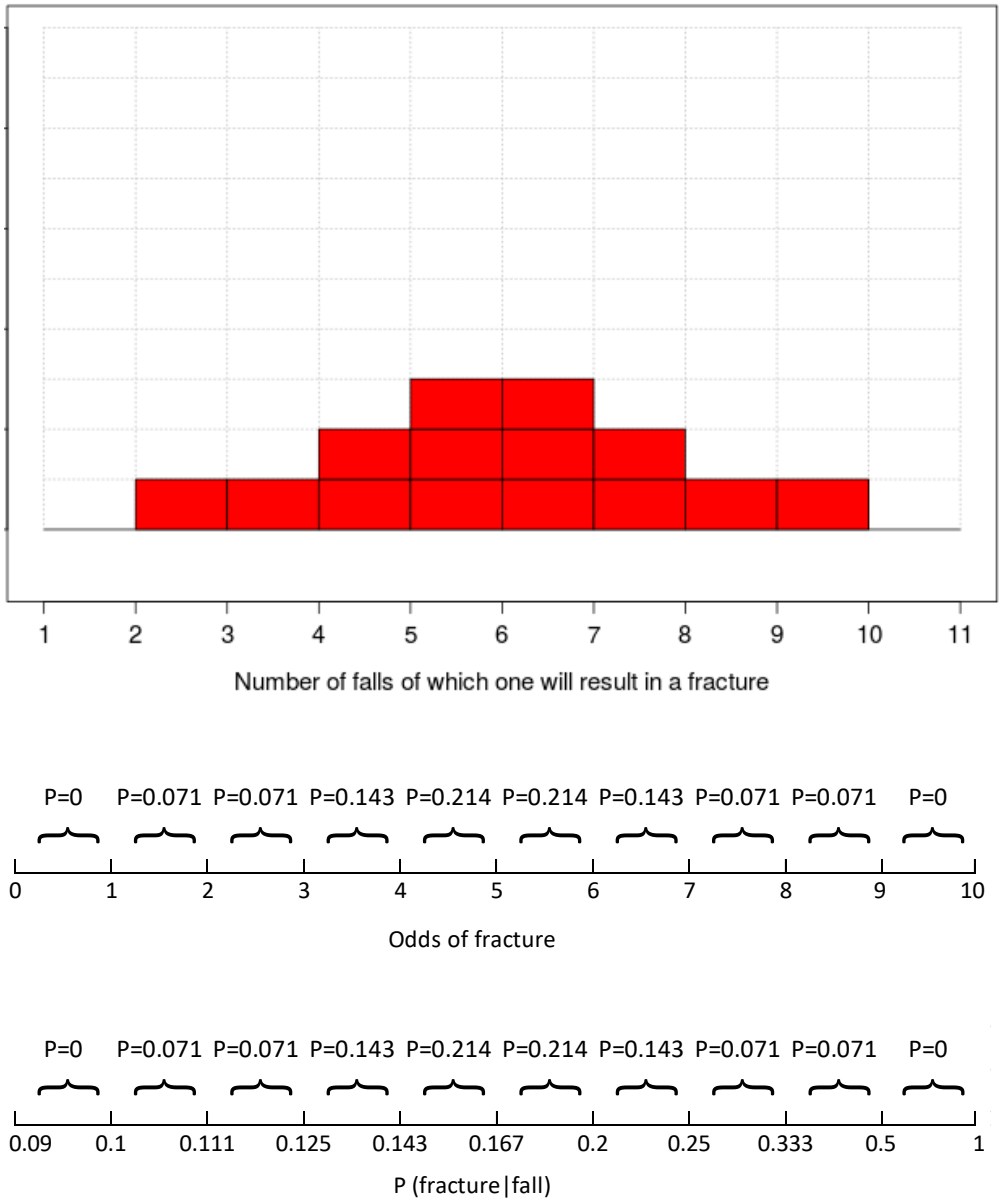
4.2.3. Odds and risk of fracture

Experts' priors on the odds of fractures were elicited by asking - one in how many falls will result in a fracture. The elicited quantities represent odd+1. For example, if an expert believed that 1 in 2 falls would result in a fracture, the odds of fracture were 1 (Number of falls that do not result in a fracture/ Number of falls that result in a fracture). The elicited quantities were thus converted into odds by subtracting 1.

Experts' priors on the odds of fracture were analysed both as odds and as probabilities of fracture conditional on falling P(fracture|fall) for comparison; both quantities were analysed because odds were the quantity that was directly elicited, while probabilities were the quantity used in the cost-effectiveness model in the case study. The probabilities were derived by inverting the elicited quantities.

The methods are demonstrated in a hypothetical example in Figure 4.4. Note that the direction of the axis is different for odds and probabilities – in Figure 4.4 the odds of 10 (the upper limit in the grid) represent a probability of 0.09 (the lower end of the probability range).

Figure 4.4. Experts' priors on odds and probabilities of fracture and the resulting probability summaries used in the analysis.



4.2.4. Deriving the treatment effect

Chapter 3 explained that in the REFORM elicitation study experts' priors on falling and fractures in those who do and do not receive the intervention were elicited separately, then used to derive the treatment effect of the intervention. The treatment effect on the probabilities, rates and odds was measured as relative risk (RR), rate ratios (RtR) and odds ratios, respectively. Sections 4.2.4.1 to 4.2.4.3 describes the methods used to derive each parameter in turn.

4.2.4.1. Deriving the relative risk of falls

Expert's frequencies were converted into probabilities as described in Figure 4.3 in section 4.2.1. The relative risk was then derived using Equation 4.1.

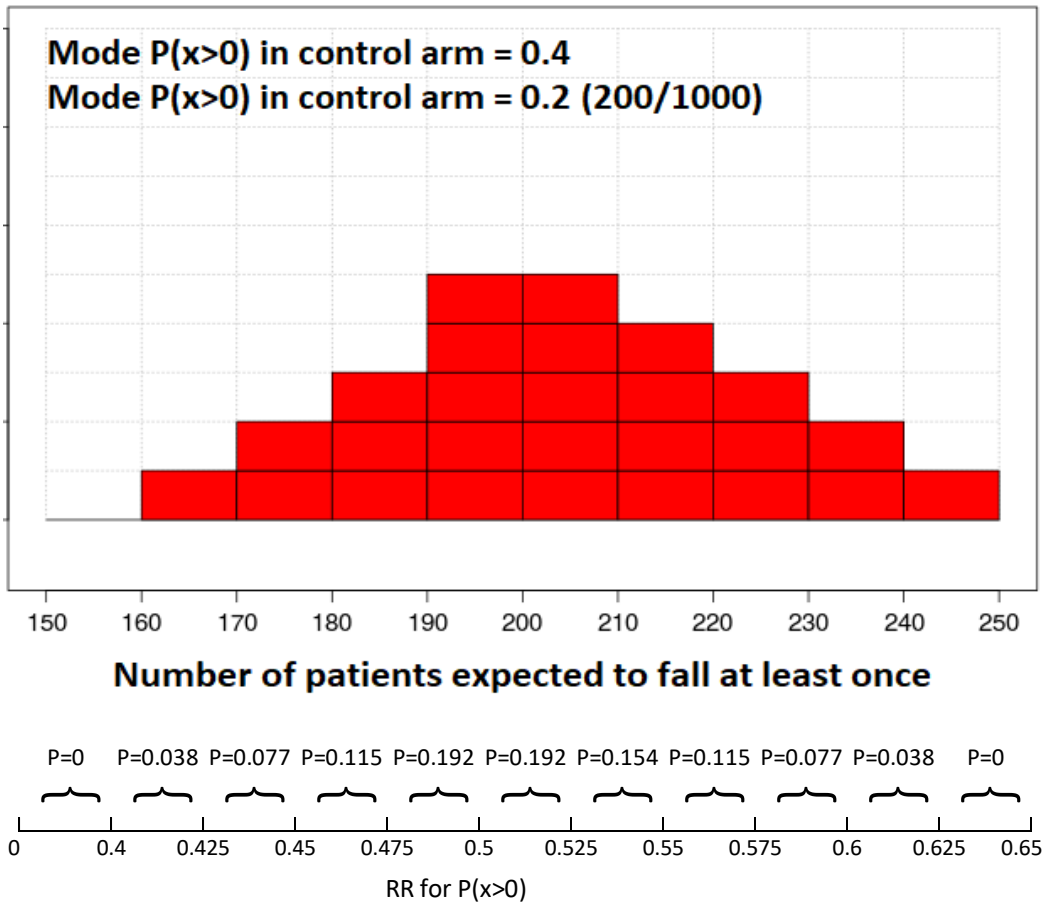
$$RR = \frac{P_T}{P_C} \qquad \textbf{Equation 4.1}$$

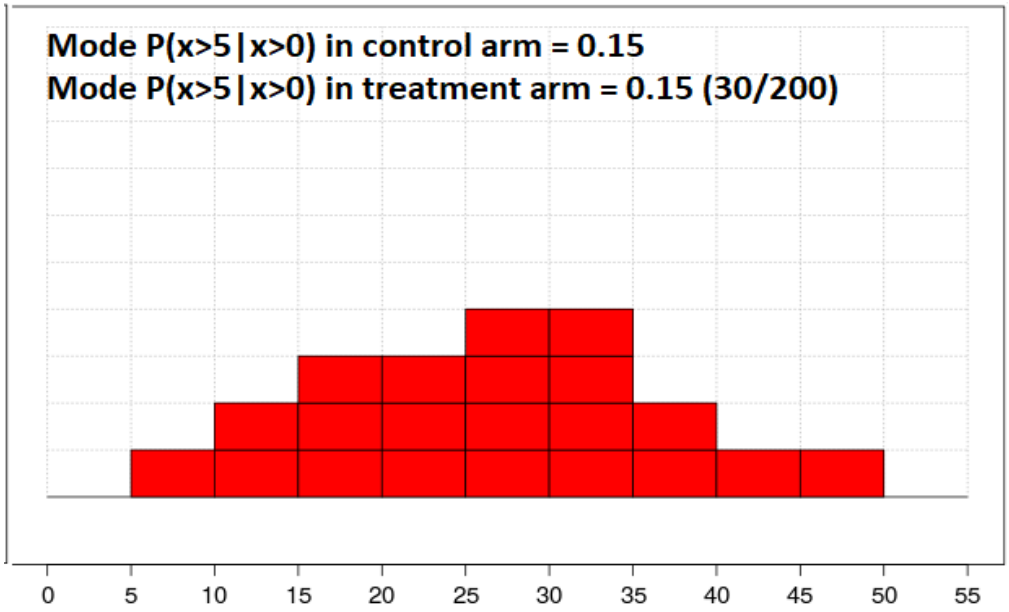
Where P_T is an experts' probability of falling in treatment arm,

P_C is the experts' mode probability of falling in the control arm.

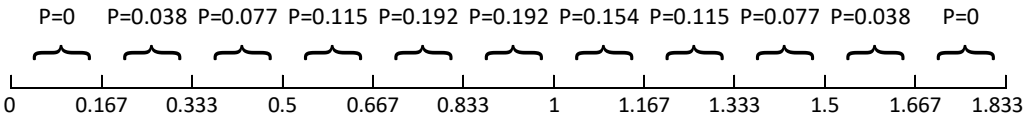
The methods for deriving the treatment effect are demonstrated in Figure 4.5.

Figure 4.5. Experts' priors on the frequency of falling in the treatment arm and the resulting treatment effect summaries used in the analysis.

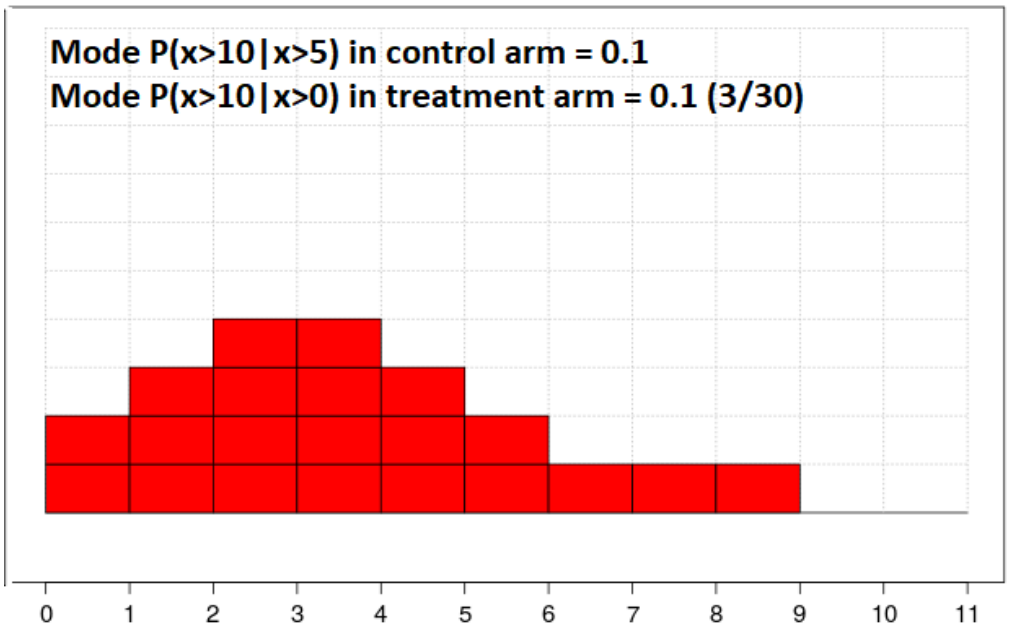




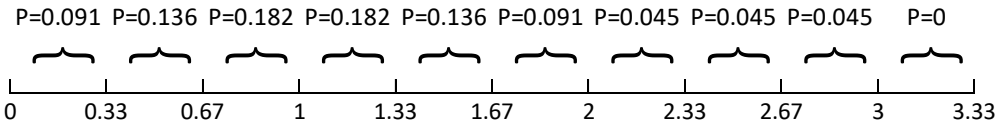
Number of patients expected to fall MORE THAN FIVE times



RR for P(x>5|x>10)



Number of patients expected to fall MORE THAN TEN times



RR for P(x>10|x>5)

4.2.4.2. Deriving odds ratios and relative risk of fractures

Section 4.2.3 highlighted that experts' priors on the odds of fracture were analysed both as odds and as probabilities of fracture conditional on falling, for comparison. The treatment effect of the intervention was thus analysed both as the odds ratio (OR) and the RR.

Odds of fracture were not elicited conditional on the mode and so the OR was derived by sampling from both priors, assuming a uniform probability distribution in each bin. The odds ratio was then derived using Equation 4.2.

$$OR = \frac{Odd_T}{Odd_C} \qquad \textbf{Equation 4.2}$$

Where Odd_T is a random sample of the odd of having a fracture in treatment arm,

Odd_c is a random sample of the odd of having a fracture in control arm.

The relative risk of fractures was calculated using Equation 4.3

$$RR = \frac{P(fracture|fall)_T}{P(fracture|fall)_C} \qquad \textbf{Equation 4.3}$$

Where $P(fracture|fall)_T$ is a random sample of the risk of fracture in treatment arm,

$P(fracture|fall)_C$ is a random sample of the risk of fracture in control arm,

4.2.4.3. Deriving rate ratios

The rate of falls in patients who received treatment was derived in the same way as for those who did not - the methods are described in section 4.2.2. The mode rate of falls in control arm was then derived from mode probabilities of falling in control arm. The rate ratio was derived using Equation 4.4.

$$RtR = \frac{Rate\ of\ falls_T}{Rate\ of\ falls_C} \qquad \textbf{Equation 4.4}$$

Where $Rate\ of\ falls_T$ is a random sample of the rate of falls in treatment arm,

$Rate\ of\ falls_C$ is the point estimate of the rate of falls derived from experts' mode probabilities of P(x>0), P(x>5|P>0) and P(x>10|P>5).

4.2.5. Deriving the temporal change in the treatment effect

The treatment effect in the case study was measured as the rate ratio for falls and the relative risk of fractures (details are provided in Chapter 6), and so the elicited priors were used to derive the temporal change in these two parameters, derived from experts' priors.

Step 6 in Figure 4.1 in section 4.1 shows that experts expressed their beliefs about falling and fractures in patients who continued to receive the intervention after the REFORM trial ended, assuming that falling and fractures in those who do not receive the intervention remained the same. The time point ($t2$) was determined by experts. Their priors were used to derive the rate of falls and the risk of fractures (see section 4.2.2 and 4.2.3 for details), and the treatment effect (RtR and RR) at the second time point (see section 4.2.4 for details).

The temporal change in the treatment effect, ΔTE, was then derived using Equation 4.5 (copy of Equation 3.4 in Chapter 3).

$$\Delta TE = \frac{TE_{t2} - TE_{t1}}{TE_{t1}} \qquad \textbf{Equation 4.5}$$

Where TE could be RtR or RR,

TE_{t1} is the treatment effect in REFORM trial,

TE_{t2} is the treatment effect after the trial.

TE can take any value between 0 and infinity, and so ΔTE could take any value between -1 and infinity, where negative values indicate the treatment effect would decrease, 0 indicates no change in the treatment effect, and positive values indicate that the treatment effect would increase. An increase in the treatment effect means that any harmful effect of the treatment is potentiating, or that the beneficial effect is diminishing, while a decrease in the treatment effect indicates that harmful effect is diminishing or that any beneficial effect is potentiating.

The second time point at which the treatment effect was elicited $t2$ varied between experts. In order to make experts' priors on the change in the treatment effect comparable, ΔTE was used to derive the annual change in treatment effect, ΔATE.

The change in the treatment effect was assumed to be linear, and so ΔATE was derived using Equation 4.6.

$$\boldsymbol{\Delta ATE = \frac{\Delta TE}{t2}} \qquad \textbf{Equation 4.6}$$

Where $t2$ is the second time point at which the treatment effect was elicited, measured in years.

In addition, sensitivity analysis was conducted, where the temporal change in the treatment effect was assumed to be log-linear, and derived using Equation 4.7.

$$\boldsymbol{\Delta ATE = e^{\frac{ln\Delta TE}{t2}}} \qquad \textbf{Equation 4.7}$$

To get the probability distribution around ΔATE, the random samples were drawn from experts' priors on ΔTE, and ΔATE was calculated for each.

4.2.6. Evaluating the elicitation exercise

Chapter 1 discussed that evaluation is an integral part of the elicitation process, aiming to determine 'how well' the elicitation has been done.

In this study, two aspects of the elicitation exercise were explored: 1) the plausibility and implications of the assumptions imposed in the exercise, and 2) how well the elicited priors represent their beliefs. The methods used for both of these are described in sections 4.2.6.1 and 4.2.6.2.

4.2.6.1. Exploring the plausibility and implications of the assumptions imposed in the exercise

Chapter 3 described the elicitation methods in detail and the discussion in section 3.6 highlighted that the methods employed were based on a series of assumptions. In summary, the following assumptions were made:

1. Conditional probabilities of different outcomes (1-5, 6-10 and >10 falls) are independent;
2. The rate of falls derived from experts' priors accurately represents their' beliefs;
3. The treatment effect of the podiatry intervention is independent of the baseline rate of falls and risk of fractures;
4. The change in the treatment effect is linear.

These assumptions may affect the validity of the results – if the assumptions do not hold, then the derived probability distributions of the rate of falls, the treatment effect, and the

change in the treatment effect don't represent experts' uncertainty, their score's don't represent their accuracy, and any effect of experts' characteristics on their scores is invalid.

The plausibility of each assumption was thus explored as follows.

Assumption 1

The assumption could not be tested directly post-hoc as it requires input from experts to indicate whether they agree with the assumption or not. Instead, correlation coefficients between expert' mode P(x>0) and P(x>5|Px>0), P(x>0) and P(x>10|Px>5), and P(x>5|Px>0) and P(x>10|Px>5) were derived to explore whether there was any correlation between the conditional probabilities.

The correlation coefficient measures the strength of relationship between two variables. (Sedgwick, 2012) The coefficient can take any value between -1 and 1, where positive values indicate positive correlation, and negative values indicate negative correlation. The higher the absolute value the stronger the correlation.

Assumption 2

Section 4.2.2 described that the rate of falls was derived by predicting the number of having 1, 2, 3, etc. falls from experts' elicited probabilities on P(x>0), P(x>5) and P(x>10).

In order to test whether the derived rates represented experts' beliefs, first, the probabilities of having 1-5, 6-10 and >11 falls, predicted by the model used to derive rates (see Chapter 3, section 3.5.2 for details), were compared to those derived directly from experts' priors, using Equation 3.1 in Chapter 3 to derive P(x>5) and P(x>10), and subsequently Equation 4.8 to derive probabilities P(0<x≤5), P(5<x≤10), and P(10<x≤30).

$$P(a < x \leq b) = P(x > 0) - P(x > b) \qquad \textbf{Equation 4.8}$$

The probabilities were compared for five randomly chosen experts from the sample selected using the random number generator in R.

Furthermore, the plausibility of the assumption was explored by comparing the rates derived using the method described in Chapter 3 to those derived using two alternative methods.

Both methods involved deriving P(x>5) and P(x>10) using Equation 3.1. Then, probabilities P(0<x≤5), P(5<x≤10), and P(10<x≤30) were derived using Equation 4.8. The maximum

number of falls was assumed to be 30 (P(x>30)=0) as the model for deriving rates predicted that the probability of having more than 30 falls was negligible. The first alternative method assumed that the probabilities of all outcomes within each category (1-5 falls, 6-10 falls and 11-30 falls) were equal. For example, if an expert believed that the probability of 1-5 falls is 0.5, then the probability of 1, 2, 3, 4 and 5 falls was 0.1 (0.5/5).

The second method assumed that the probability of each category (1-5, 6-10 and 11-30 falls) was that stated by the expert, but the probability of each additional fall within that category decreases at a constant rate. The probability of each number of falls was calculated using Equation 4.9.

$$P(x = f) = P(x = f + 1) + \frac{P(c_{min} \leq x \leq c_{max})}{\sum C_f} \qquad \text{Equation 4.9}$$

Where f is one of 30 falls,

C_f is the number of falls in category of fall f, so if f = 4 then $C_f = (1, 2, 3, 4, 5)$, and $\sum C_f$ = 15,

$c_{min}$is the smallest number of falls in category $C(f)$, so if f = 4 then c_{min}=1

c_{max} is the largest number of falls in category $C(f)$, so if f = 4 then c_{min}=5.

For example, if an expert believed that the probability of falling more than 10 times was 0.1, and the highest possible number of falls was 30 (an assumption imposed in the analysis), then the model assumed that the probability of 31 falls was 0 and distributed the 0.1 probability across the remaining outcomes (11-30). This probability constantly decreases by 0.000476 (generated by Equation 4.9) for each fall between 11 and 30. This results in the sum of probabilities for having 11-30 falls of 0.1. The probability of having 11 falls was thus estimated to be 0.010 (20*0.000476). Similarly, if the same expert believed that the probability of having 6-10 falls was 0.3, then the probability of each additional fall (for 6-11 falls) was assumed to decrease at a constant rate. Hence, the probability of exactly six falls was estimated to be 0.106 and the probability of having each additional fall was 0.0194 lower so that the sum of probabilities of having 6-10 falls was 0.3.

Assumption 3

Like Assumption 1, Assumption 3 could not be tested post-hoc, but the correlation coefficient between experts' mode probabilities of falling, odds of falling, and rate of falls and the treatment effect on these were derived to explore whether there was any

consistency in the way experts' beliefs about falling and fractures affected their beliefs about the treatment effect.

Assumption 4

In the REFORM elicitation study, the experts expressed when they expected the treatment effect to diminish or plateau (time point *t3*), their priors were then elicited at a second time point t2. When experts believed the treatment effect would diminish, *t2* = *t3*/2, whereas when they believed the treatment effect would potentiate, *t3* = *t2*. In order to test whether the estimated annual change in the treatment effect reflected experts' beliefs, it was applied to the mode of experts' treatment effect after one year to derive the treatment effect at the time point *t3*. When the change in the treatment effect was assumed to be linear, the treatment effect at the time point *t3* was derived using Equation 4.10 (derived from Equation 4.5 and Equation 4.6).

$$TE_{t3} = TE_{t1} * (1 + \Delta ATE * t3) \qquad \textbf{Equation 4.10}$$

Where TE_{t3} is the predicted treatment effect at time t3,

TE_{t1} is expert's mode within trial treatment effect,

ΔATE is the annual change in the treatment effect derived from expert's priors.

When the change in the treatment effect was assumed to be log linear, the treatment effect at the time point *t3* was derived using Equation 4.11[11].

$$TE_{t3} = TE_{t1} * \Delta ATE^{t3} \qquad \textbf{Equation 4.11}$$

For experts who believed the treatment effect would diminish, the predicted TE_{t3} should be close to 1 indicating no treatment effect. In this study 'close to 1' was defined as 0.8 - 1.2. Predicted values outside this range would indicate that the change in the treatment effect is not linear (or log linear).

[11] Since $lnTE_{t3} = lnTE_{t1} + ln\Delta ATE * t3$

4.2.6.2. Evaluating experts' priors

Methods for evaluating the elicited priors were discussed in Chapter 1 (section 1.3.4, *Fitting, Evaluation and Feedback*) and include internal consistency and seed-calibration. (Wallsten and Budescu, 1983)

Seed calibration was not used to evaluate the elicitation exercise in this chapter, as detailed analysis of experts' calibration is provided in Chapter 5. While the elicitation exercise was designed to prevent any statistically incoherent responses – for example, experts could not exceed the parameter limits when expressing their range – experts' priors on the treatment effect after the trial could be inconsistent with their verbal responses (to MCQs) about the same. Experts' internal consistency was thus used to assess whether the elicited priors represented experts' beliefs.

In their answers to the MCQs, experts could express five beliefs about the effectiveness of the intervention after the trial end point. Of these five, four led to further elicitation and so were used to assess internal consistency, as shown in Table 4.1.

Table 4.1. Indicators of internal inconsistency in experts' priors.

Beliefs expressed in MCQs	**Indication of internal inconsistency**
The intervention will be as effective, indefinitely	N/A
The effect of the intervention is most likely to stay the same although the expert is not certain.	Expert's confidence interval of ΔTE does not include 0, suggesting the expert is certain the treatment effect would change over time
The intervention is most likely to become less effective over time.	Expert's median ΔTE is positive indicating TE is most likely to potentiate (regardless of whether the expert thought the intervention was harmful or beneficial)
The intervention is most likely to become more effective over time.	Expert's median ΔTE is negative indicating TE is most likely to (regardless of whether the expert thought the intervention was harmful or beneficial)
The intervention effect is likely to change over time, but the expert is not sure whether it will get better or worse.	Expert's confidence interval of ΔTE does not include 0, suggesting the expert is certain about the direction of change in the treatment effect (i.e. it is either positive or negative)

Furthermore, exercise was evaluated by visually inspecting experts' priors to identify any features that could indicate that experts did not understand the task.

4.3. Sample description

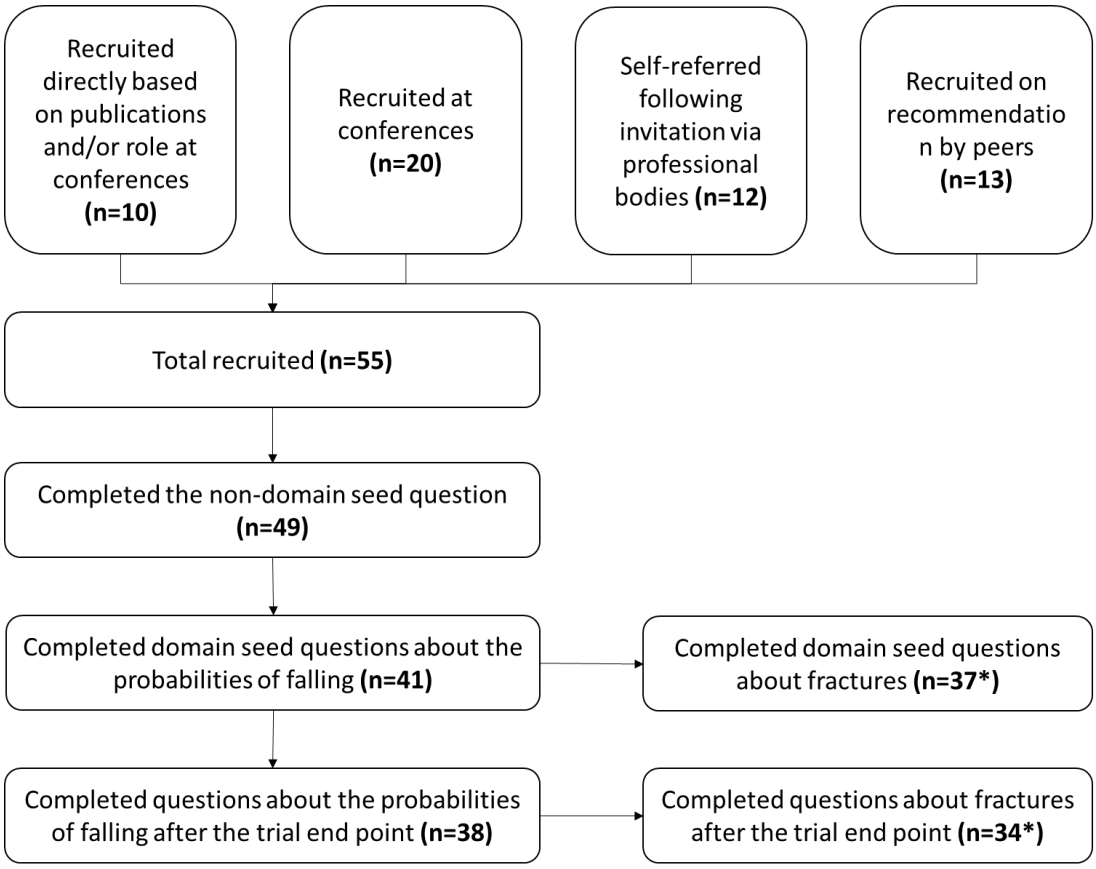
The number of experts who completed the REFORM elicitation study was 39. This included physiotherapists, geriatricians, academics, nurses, occupational therapists (OTs), and general practitioners (GPs). Section 4.3.1 describes the recruitment and completion rate, section 4.3.2 shows the composition of the sample, and section 4.3.3 reports on the mode of delivery.

4.3.1. Recruitment and completion rate

Experts were recruited via four different avenues: 1) approached directly based on publications and/or their role at conferences; 2) recruited at conferences; 3) self-referred through a colleague or an email sent from a society; 4) contacted on recommendation from experts recruited through the first three routes. The number of experts approached and recruited through each of these routes is presented in Figure 4.6. Initially, 55 experts consented to taking part, of which 16 did not complete the exercise. Fifteen experts did not complete the exercise because they encountered technical problems. These were almost exclusively due to poor internet connection – once experts were disconnected they had to commence the exercise from the beginning. One more expert terminated the exercise due to time constraints.

Forty-one experts expressed their priors about falling during the trial, and 39 answered MCQs about the effectiveness of the intervention after the trial end point. Two experts dropped out because their internet connection failed. Thirty-eight experts completed the entire exercise while one additional expert dropped out before completion due to a poor internet connection. Of the 38 experts who completed the exercise, 36 expressed their priors on the treatment effect at the second time point, after the trial. One experts did not express their priors at the second time point because they believed that the intervention treatment effect would remain the same, while one expert believed that the effect of the intervention would diminish within three months and so instead of eliciting their priors about the effect at the second time point it was assumed that it would diminish immediately in the analysis in Chapter 6.

Figure 4.6. Recruitment and completion rate of the REFORM elicitation study.



*Three additional exerts provided the range but did not express their uncertainty as the chips and bins grid failed to show.

An additional four experts who completed the exercise could not express their priors on the odds of fractures – they could input the minimum and maximum but the chips and bins grid failed to show.

4.3.2. Composition of the sample

Table 4.2 shows the profession and method of recruitment for the 55 participants who provided information about their professional experience. The method of recruitment differed according to the profession. This variation to some extent reflects the differences in operations between different professional bodies. The British Geriatrics Society (BGS) has no branch or sub group specialising in fall prevention so participants in this profession were recruited by attending regional branch meetings and conferences. There are a large number of such meetings and so this was the most efficient method of recruitment.

Table 4.2. Method of recruitment by profession.

	N	Identified by investigator	Recommended by peers	Self-referred	Recruited at conference
Physio.	20	4	4	10	2
Geriatricians	21	3	2	0	16
Academics	1	1	0	0	0
Nurses and OTs	7	0	4	1	2
GPs	4	1	2	0	1
Other	2	1	1	0	0
Total	55	10	13	11	21

The Chartered Society of Physiotherapy (CSP) has a professional network for physiotherapists (and other health professionals) who work with older people, AGILE. AGILE sent an email invitation to its network to invite participation in the exercise to all members in England, which is likely to be the reason they have the highest proportion of self-referred participants.

The majority of publications in fall prevention are written by physiotherapists, geriatricians and health researchers in fall prevention (academics), which is likely to be why they have the highest number of experts who were directly approached for the study.

Nurses and occupational therapists were recruited at BGS meetings, through AGILE and on recommendation from physiotherapy colleagues who work in fall prevention clinics. The participation rate from these groups is low as relatively few nurses are members of these societies.

4.3.3. Mode of delivery

Mode of completion, time taken to complete the exercise and proportion of experts who amended their answers are presented in Table 4.3 (classified by profession).

Table 4.3. Summary of experts' elicitation, classified by profession. Time taken to complete refers to the time experts took to complete the question about the outcomes of the REFORM trial.

	Mode of completion (proportions)				Time taken to complete (min)	Proportion who ammended their answers	N
	No assistance	Assistance over the phone	Assistance in person, one-on-one	Assistance in person, in groups			
Physiotherapists	0.15	0.15	0.62	0.08	18.8	0.46	13
Geriatricians	0.73	0.07	0.2	0	13.7	0.2	15
Academics	0	0	1	0	16	0	1
Nurses and OTs	0.57	0.14	0.29	0	18.7	0.43	7
GPs	0.25	0	0.5	0.25	18	0.25	4
Other	0	0	1	0	15	0	1

The mode of delivery was self-reported so some inconsistencies exist. For example, all geriatricians who completed the exercise at regional meetings and conferences declared that they completed the exercise independently, while one physiotherapist and one GP who completed it at a BGS regional meeting along with other geriatricians felt that they were guided by the investigator in a group.

The remainder of this chapter provides an overview of the results of the elicitation exercise.

4.4. Results: Overview of experts' priors

This section describes experts' priors on each elicited parameter. All priors elicited from experts who completed the domain seed on falling are reported including:

- 41 priors on the non-domain seed,
- 41 priors on the probabilities of falling and the treatment effect on falls during the REFORM trial,
- 37 priors on the odds of having a fracture and effect of the intervention on fractures during the trial,
- Beliefs of 38 experts about the probabilities of falling after the trial (36 priors), and
- Beliefs of 34 experts about the odds of having a fracture after the trial (32 priors).

The number of priors is not identical to the number of experts whose beliefs were elicited about the treatment effect after the trial, because two experts expressed beliefs that did not warrant further elicitation – further details are provided in section 4.4.4.

4.4.1. Non-domain seed

Experts' priors on the non-domain seed (the number of rainy days in York in September) are shown in Figure 4.7 and summarised in Table 4.4 in comparison to the average number of rainy days in September recorded by the Met Office, as discussed in Chapter 3 (section 3.4.3).

Figure 4.7. Experts' priors on the number of rainy days in September in York. Box = median and interquartile range, whiskers = range. Horizontal lines = mean and 95% confidence interval of the observed values.

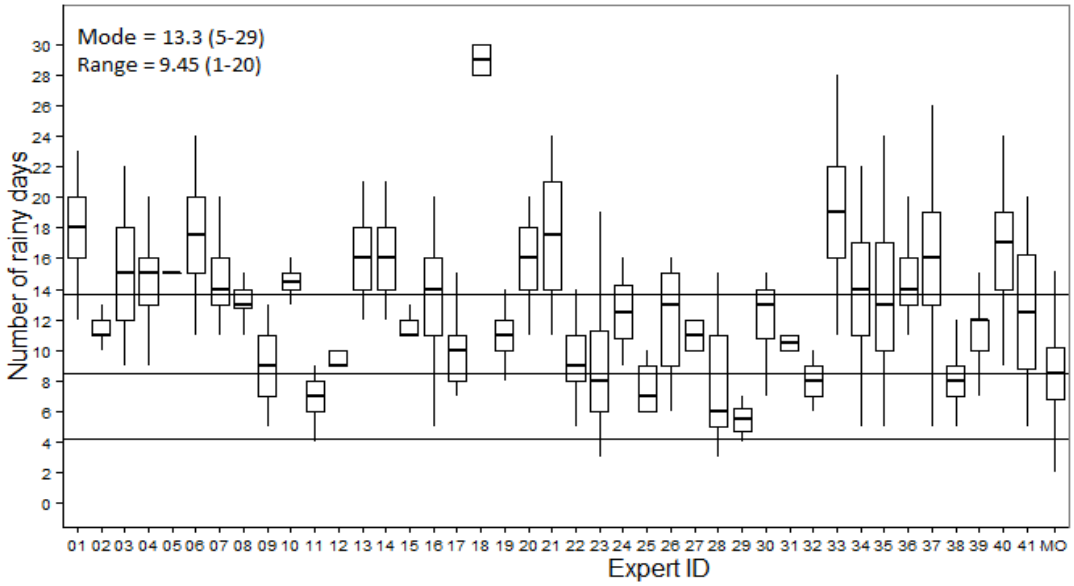


Table 4.4. Summary of experts' priors on the number of rainy days in September in York (the non-domain seed).

	Mean of experts' modes (range of modes)	Range indicated to be plausible by experts	Observed number of rainy days (95% CI)
Days of rainfall - out of 30 (n=41)	13.3 (5-29)	2-31	8.6 (4.3-13.7)

On average, experts overestimated the total number of rainy days – their priors indicated that the most likely number of rainy days in York was 13.3 compared to the 8.6 days recorded by the Met Office. There was considerable variation in experts' beliefs – their modes varied between 5 and 29, and the range varied across almost the entire range of possible values (2-31).

4.4.2. Falling and fractures without treatment

The summary of experts' priors about the probabilities of falling in those patients who do not receive the intervention are presented in Figure 4.8 and in Table 4.5.

Figure 4.8. Experts' priors on the probabilities of falling in patients who did not receive the intervention and the values observed in the REFORM trial (RCT). Box = median and interquartile range, whiskers = range. Horizontal lines = mean and 95% confidence interval observed in the trial.

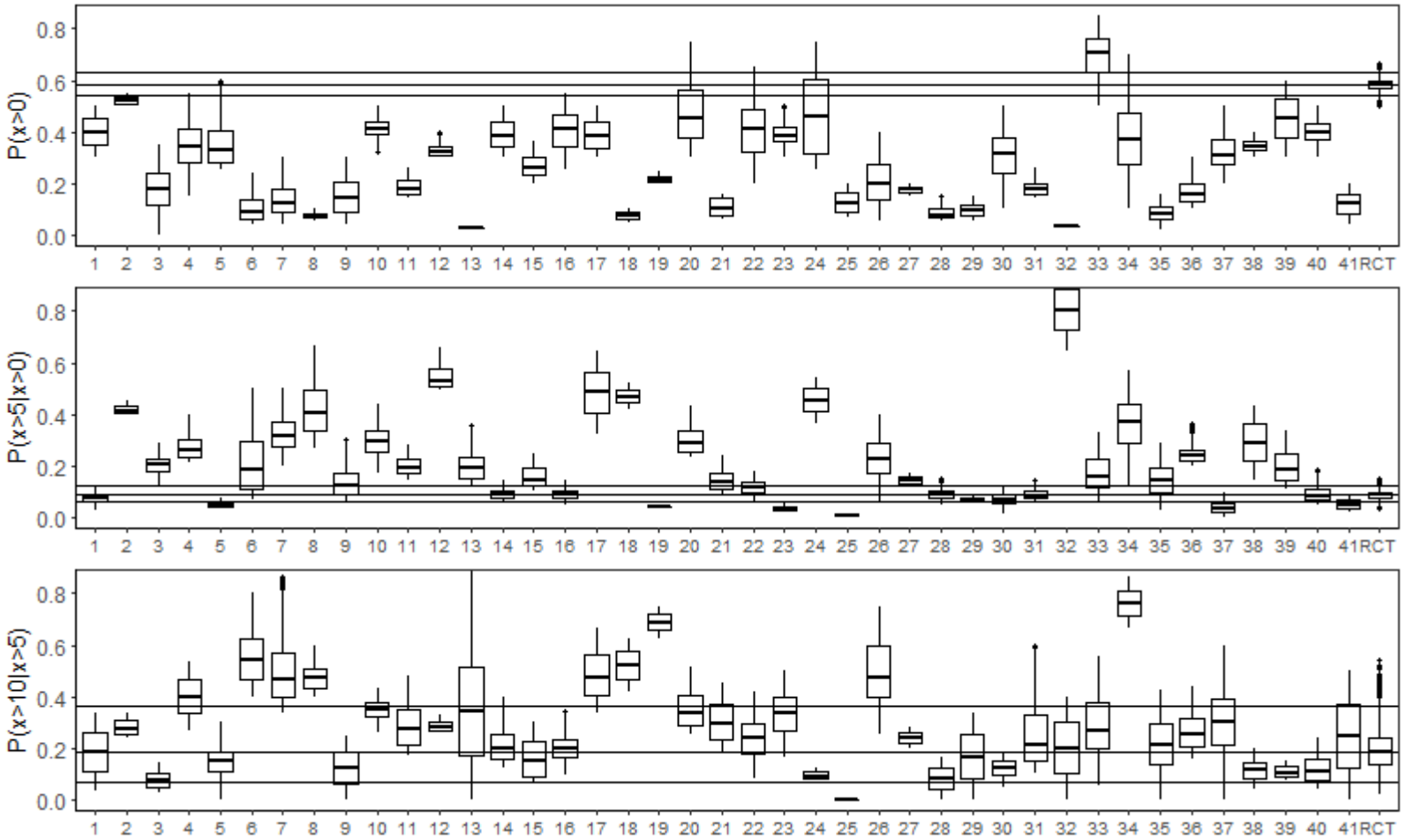


Table 4.5. Summary of experts' priors on the probability of falls and fractures without treatment.

Quantity	Mean of experts' modes (range of modes)	Plausible range	Observed in REFORM trial (95% CI)
P(x>0) (n=41)	0.241 (0.025-0.725)	0-0.850	0.585 (0.542-0.627)
P(x>5|x>0) (n=41)	0.195 (0.001-0.806)	0.004-0.968	0.088 (0.058-0.122)
P(x>10|x>5) (n=41)	0.258 (0.017-0.767)	0-1	0.192 (0.068-0.361)

Overall experts underestimated the proportion of fallers– their priors suggested that on average the most likely proportion of fallers in the control arm would be less than one quarter of all patients (0.241) compared to 0.585 reported in the REFORM trial. The conditional probabilities of falling more than five and more than ten times were higher in the elicited priors (the average most likely values were 0.195 and 0.258 respectively) than observed in the trial (0.088 and 0.192 respectively).

Experts' priors varied substantially – experts believed that the most likely probability of falling (i.e. the mode) could be as low as 0.025 and as high as 0.725 while the plausible range was between 0 and 0.85, and the conditional probabilities of falling more than five and more than ten times included almost the entire possible range 0-1.

Rate of falls

The rates of falls derived from experts' priors are shown in Table 4.6 and Figure 4.9.

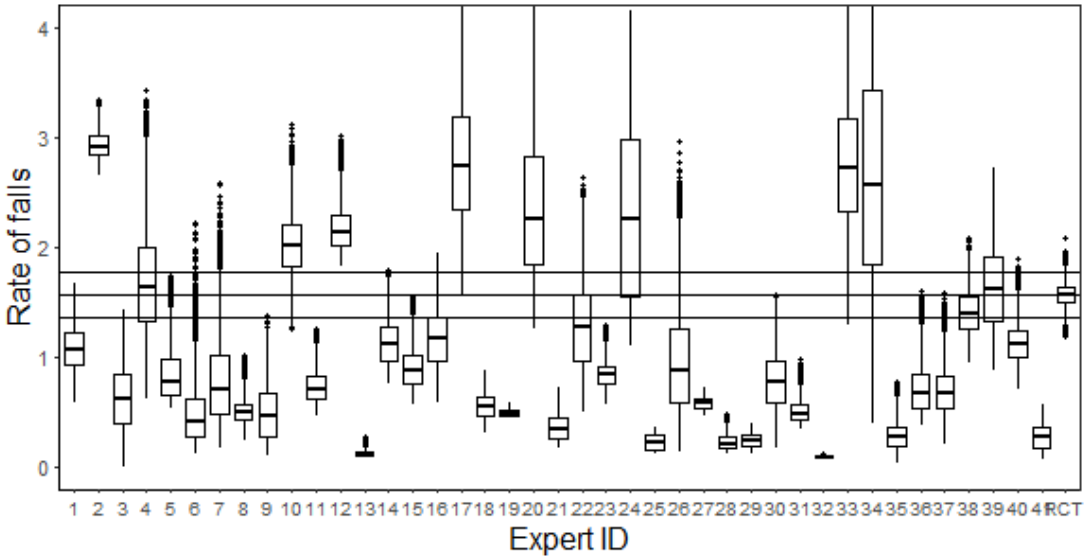
Table 4.6. The summary of the rate of falls derived from experts' priors.

Elicited quantity	Mean of experts' modes (range of modes)	Plausible range	Observed in REFORM (95% CI)
Derived rate of falls (n=41)	0.945 (0.085-2.825)	0.002-7.054	1.57 (1.366-1.777)

Overall experts underestimated the rate of falls – on average experts' most likely number of falls was 0.945, compared to the 1.5 rate observed in the trial, likely because they underestimated the proportion of fallers (as shown in Figure 4.8), which subsequently influenced the rate of falls derived from their priors. Indeed, the majority of experts who underestimated the proportion of fallers (Experts 13, 18, 19 and 33) also underestimated

the rate of falls. However, some experts (such as Expert 12) overestimated the proportion of fallers also overestimated the rate of falls, by assigning high probabilities to the probabilities of having more than five and more than ten falls.

Figure 4.9. Rate of falls derived from experts' priors without treatment and the probability observed in the REFORM trial (RCT). Box = median and interquartile range, whiskers = range. Horizontal lines = mean and 95% confidence interval observed in the trial.



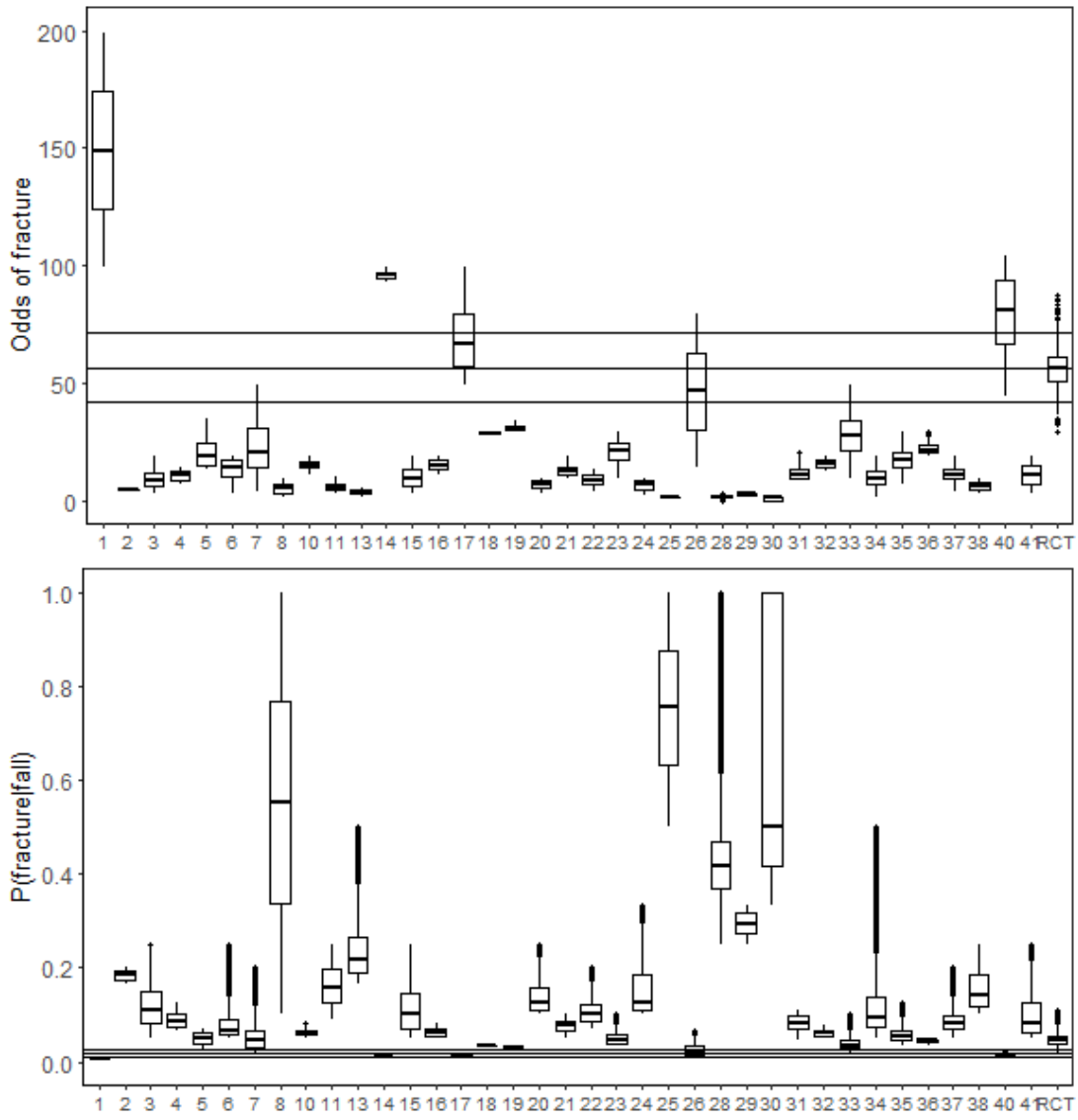
Risk of fracture

Experts' priors on the odds and probabilities of having a fracture after a fall is shown in Table 4.9 and Figure 4.10. On average, experts underestimated the odds of fracture (21.4 compared to 55.9 observed in the trial), although their modes varied from every fall to one in 149, while plausible range varied between every fall to one in 1000 falls.

Table 4.7. A summary of the risk of fractures derived from experts' priors.

Quantity	Mean of experts' modes (range of modes)	Plausible range	Observed in REFORM trial (95% CI)
Odds of having a fracture (n=37)	21.4 (1-149)	1-1000	55.9 (42-71)
P(fracture|fall) (n=37)	0.127 (0.007-0.5)	0.001-1	0.018 (0.010-0.027)

Figure 4.10. Experts' priors on the odds and probabilities of having a fracture in patients who do not receive the intervention and the probability observed in the REFORM trial (RCT). Box = median and interquartile range, whiskers = range. Horizontal lines = mean and 95% confidence interval observed in the trial.



The probability of suffering a fracture is inversely proportional to the probability of falling, and so it was higher in the elicited priors than observed in the trial (0.127 compared to 0.028 observed in the trial). The scale between the two parameters is different – for example, probabilities between 0 and 0.2 equate to odds of 4-infinity and so experts whose priors on the probability of fractures were very low and precise (such as Expert 1 and Expert 17) lead to uncertain priors, indicating high odds of fractures. The opposite is also the case, as experts whose priors on the odds place high probability on a narrow range of low values (such as Expert 30) lead to uncertain priors on the probability of falling.

4.4.3. Treatment effect during the trial

The treatment effect derived from experts' priors are summarised in Table 4.8, Figure 4.11 and Figure 4.12.

Table 4.8. Summary of experts' priors on the treatment effect of the intervention.

Elicited quantity	Mean of experts' modes (range of modes)	Plausible range	Observed in REFORM trial (95% CI)
Relative risk for P(x>0) (n=41)	0.872 (0.109-9.935)	0.003-10.645	0.915 (0.819-1.022)
Relative risk for P(x>5|x>0) (n=41)	1.543 (0.081-14.516)	0.064-241.936	1.014 (0.594-1.733)
Relative risk for P(x>10|x>5) (n=41)	1.339 (0.088-8.571)	0.178-12.857	1.809 (0.688-4.754)
Odds ratio (n=41)*	1.527 (0.318-12.498)	0.1-50	0.751 (0.364-1.550)

On average, experts predicted the direction of change in the probabilities of falling and the odds of fractures correctly - their priors indicate that the proportion of fallers would decrease but the conditional probability of falling more than five and more than ten times would increase after treatment.

Priors on fractures (in Figure 4.12) are generally more uncertain than those on the probabilities of falling (in Figure 4.11), possibly because the odds of fractures in patients who receive the intervention were not elicited conditional on outcomes in patients who do receive the intervention.

The predicted change in falling and fractures could imply that those who fall the least will benefit from the intervention the most, and that the conditional probabilities of falling more than five and more than ten times will increase because of the lower denominator, but it could also imply that those who fall the most will fall even more.

Figure 4.11. Experts' priors on the relative risk of falling and the treatment effect observed in the trial (RCT). Box = median and interquartile range. Whiskers = range excluding outliers. Points = outliers. Solid horizontal lines = mean and 95% confidence interval observed in the trial.

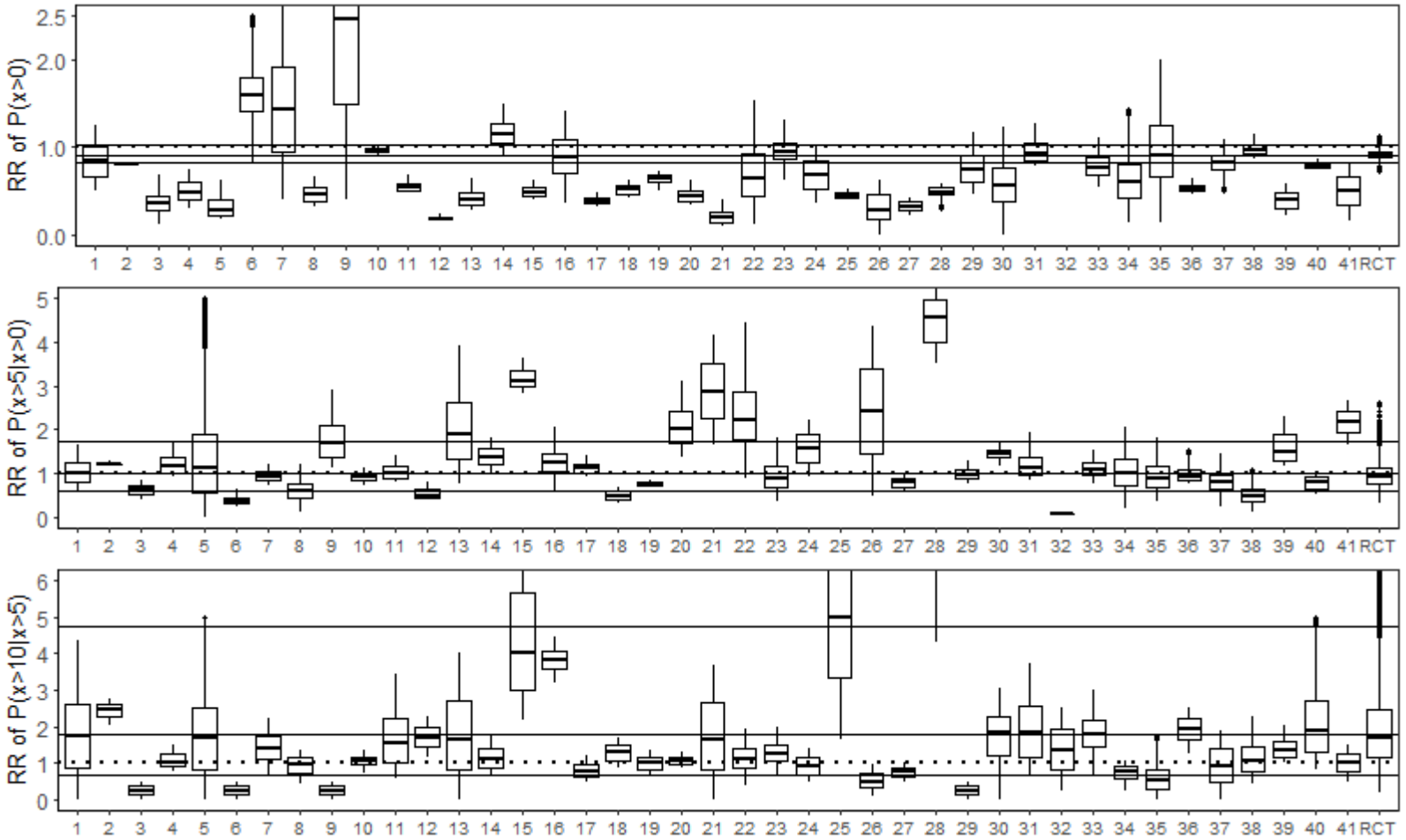
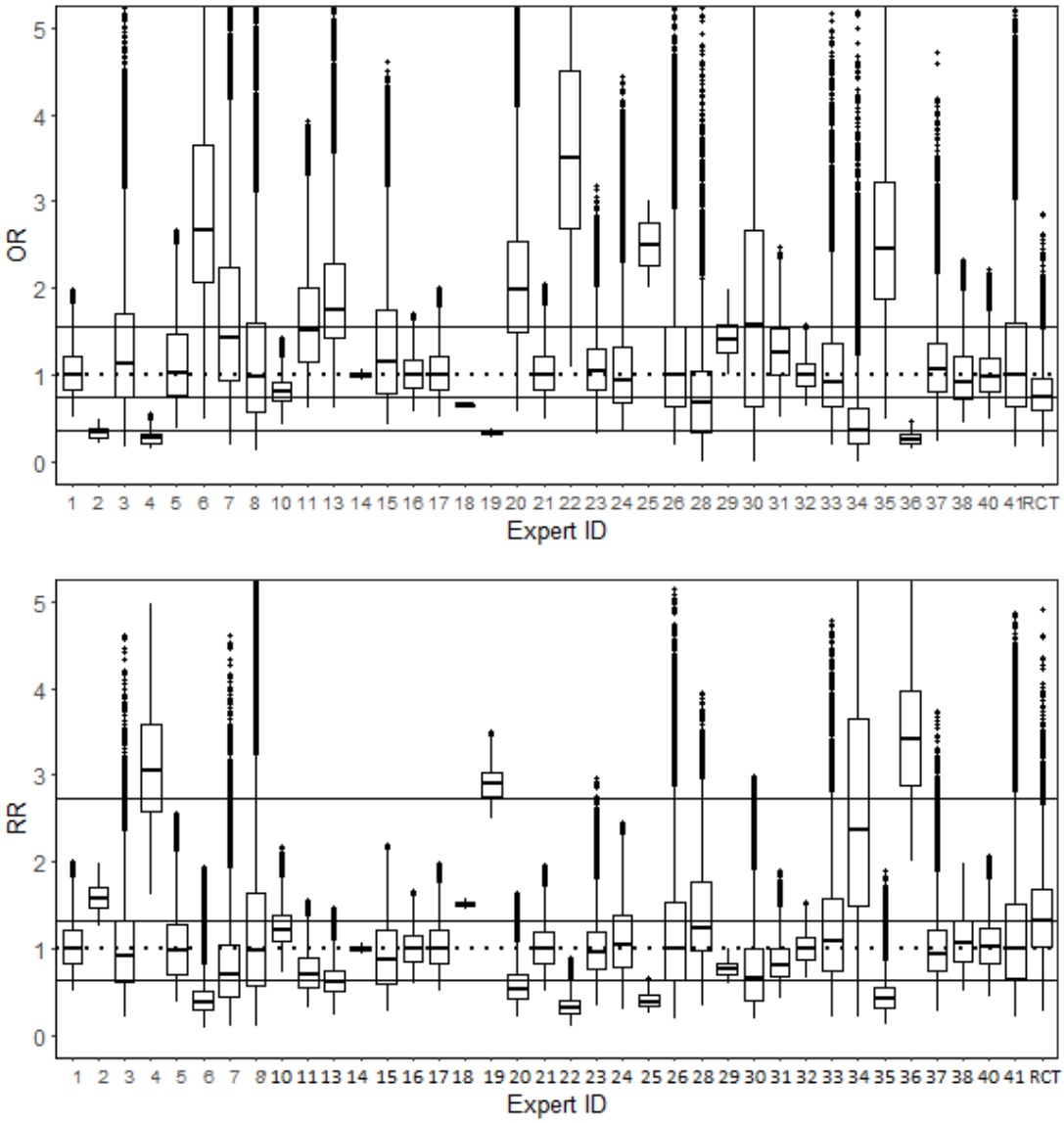


Figure 4.12. Experts' priors on the OD and RR of fractures, and the treatment effect observed in the REFORM trial (RCT). Box = median and interquartile range. Whiskers = range excluding outliers. Points = outliers. Solid horizontal lines = mean and 95% confidence interval observed in the trial.



Experts' priors are analysed in further detail to understand their beliefs about the nature of the treatment effect.

First, experts' priors were used to derive the rate of falls and the rate of fractures, to understand their beliefs about the effect of the intervention on the overall number of falls and fractures. The rate of fractures was derived by multiplying the rate of falls by the probability of having a fracture after a fall. The mode for the rate of fracture was generated for each expert by multiplying their mode probability of fracture by their mode rate of falls.

The range for each expert was derived by sampling from experts' priors on each quantity and multiplying out the random samples.

Experts' priors on the proportion of fallers, the rate of falls and the rate of fractures in the control and treatment arm are shown in Table 4.9.

Table 4.9. Summary of experts' priors on the outcomes of REFORM trial.

		Mean of experts' modes (range of modes)	Plausible range
REFORM trial outcomes, control arm	Elicited P(x>0) (n=41)	0.247 (0.025-0.725)	0-0.850
	Derived rate of falls (n=41)	0.945 (0.085-2.825)	1.368-1.779
	Derived rate of fractures (n=37)	0.100 (0.007-0.679)	0.001-1.329
REFORM trial outcomes, treatment arm	Derived relative risk for P(x>0) (n=41)	0.872 (0.109-9.935)	0.033-10.645
	Derived rate of falls ratio (n=41)	0.696 (0.110-1.990)	0.002-8.667
	Derived rate of fractures ratio (n=37)	0.766 (0.024-2.475)	0.022-3.765

Overall experts' priors suggested that the proportion of fallers, the rate of falls and the rate of fractures would all decrease, although the highest mode elicited from experts' was higher in treatment arm than in control arm (3.364 compared to 2.825), suggesting that not all experts agreed that the rate of falls would decrease.

Table 4.10 shows the elicited beliefs on the treatment effect for individual experts. Out of 41 experts who expressed their beliefs about the risk and rate of falls, 36 believed that the proportion of fallers would be lower in the treatment arm and 33 believed that the rate of falls would be lower in treatment arm. Furthermore, 31 out of the 37 experts who expressed their beliefs about the rate of fractures, believed it would be lower in the treatment arm. Out of 37 experts who expressed their beliefs about all elicited outcomes, 26 believed that the proportion of falls, the rate of falls and the rate of fractures would all decrease. Furthermore three believed that the proportion and rate of falls would decrease but the rate of fractures would increase, while four believed that the proportion of fallers and rate of fractures would decrease but the rate of falls overall would increase after

receiving the intervention. Priors from two experts indicated that the intervention would increase the proportion of fallers, the rate of falls and the rate of fractures, suggesting that overall the intervention was harmful.

Table 4.10. Experts' beliefs about the effect of the podiatry intervention on the risk and rate of falls and the rate of fractures.

		Fracture rate lower in treatment arm	Fracture rate higher in treatment arm	Did not express beliefs about the rate of fractures	Total
Probability of falls lower in treatment arm	Rate of falls lower in treatment arm	26	3	3	32
	Rate of falls higher in treatment arm	4	0	0	4
Probability of falls higher in treatment arm	Rate of falls lower in treatment arm	0	1	0	1
	Rate of falls higher in treatment arm	1	2	1	4
Total		31	6	4	41

4.4.4. Temporal change in the treatment effect

As discussed in section 4.3.1, 39 experts expressed their beliefs about the temporal change in the treatment effect after the trial verbally in MCQs, and 37 of those completed the exercise. Their responses are shown in Table 4.11. Experts predominantly believed that the treatment effect would diminish after the trial (33/39 experts), while one expert was not certain whether it would change, two believed the treatment would become more effective after the trial and two believed the treatment effect would change, but were not certain whether it would diminish or potentiate.

Table 4.11. Experts' responses to MCQs regarding treatment effect after the trial end point.

	Certain treatment effect would not change	Uncertain whether the treatment effect would change	Certain the treatment effect would change		
			Certain it would diminish	Certain it would increase	Uncertain whether it would increase or diminish
N	1	1	33	2	2

Experts who thought the treatment effect would potentiate thought it would plateau after 3 years on average, while experts who thought it would decrease thought it would diminish after 3.2 years.

One expert believed the treatment effect would not change- their temporal change was assumed to be 0. One further expert believed that the treatment effect would wear off immediately -their temporal change was assumed to equal the treatment effect so that the treatment effect after the trial is 1.

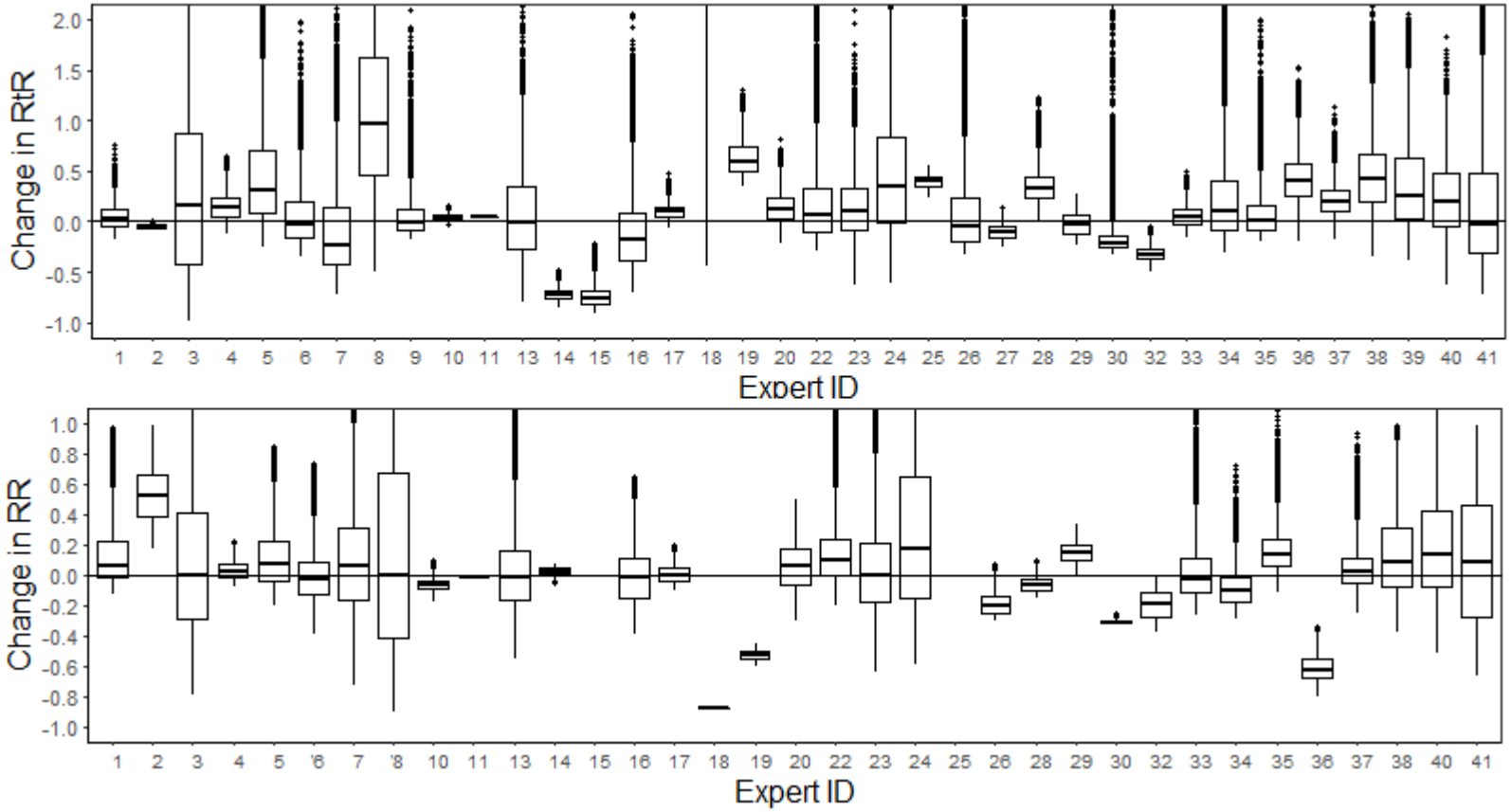
The temporal change derived from experts' priors is shown in Table 4.12 and Figure 4.13. Section 4.2.5 discussed that ΔATE can take any value between -1 and infinity, where negative values indicate the treatment effect would decrease, 0 indicates no change in the treatment effect, and positive values indicate that the treatment effect would increase. An increase in the treatment effect means that any harmful effect of the treatment is potentiating, or that the beneficial effect is diminishing, while a decrease in the treatment effect indicates that harmful effect is diminishing or that any beneficial effect is potentiating.

Table 4.12. Summary of experts' priors on the temporal change in the treatment effect.

	Mean of experts' modes (range of modes)	Plausible range
Annual change in rate ratio (n=38)	0.370 (-0.843-6.86)	-1.562-193
Annual change in odds ratio (n=34)	0.039 (-0.98-2.33)	-0.995-8.255

Overall the priors are consistent with experts' verbal responses - the mean modes in Table 4.12 are positive (0.370 and 0.039) suggesting that on average experts believed the treatment effect would diminish over time.

Figure 4.13. Probability distributions of the annual change in treatment effect derived from experts' priors. Box = median and interquartile range. Whiskers = range excluding outliers. Points = outliers. Solid horizontal line=no temporal change.



RtR = rate ratio; RR = risk ratio.

In Figure 4.13, priors elicited from 23 experts expressed priors that included 0 suggesting the rate ratio (of falls) could both increase and decrease, whereas 6 thought the same about the relative risk. Experts were more certain about the change in the relative risk than the rate ratio, possibly because it was a simpler parameter – the change in the rate ratio was derived from 9 other priors.

Furthermore, 10 experts in Figure 4.13 expressed priors that were strictly negative, suggesting they were certain the treatment effect would potentiate – this is contrary to the verbal responses in the MCQs where only 2 experts were certain the treatment effect would potentiate.

The change presented in Table 4.12 and Figure 4.13 assumed that the change in the treatment effect was linear. The implications of assuming a log-linear change in the treatment effect were also explored in sensitivity analysis; the results are presented in section 4.5.1.4.

Internal consistency of individual experts is considered in further detail in section 4.5.2.

4.5. Results: Evaluation of the elicitation exercise

This section evaluates the plausibility of assumptions in the elicitation exercise (4.5.1) and the coherence and internal consistency of the elicited priors (4.5.2), the methods of which are described in sections 4.2.6.1 and 4.2.6.2, respectively.

4.5.1. Exploring the plausibility of assumptions in the elicitation methods

Section 4.2.6 highlighted four assumptions imposed in the elicitation exercise and the methods used to evaluate them. Sections 4.5.1.1 - 4.5.1.4 presents the results for each assumption in turn.

4.5.1.1. Correlation between conditional probabilities of having >0, >5 and >10 falls

The correlation coefficients between the proportion of fallers and the conditional probabilities P(x>5|x>0) and P(x>10|x>5) are shown in Table 4.13. The correlation coefficient can take any value between -1 and 1, where positive values indicate positive correlation, and negative values indicate negative correlation. The higher the absolute value the stronger the correlation. The coefficients in Table 4.13 range from -0.11 to 0.27 –

these are relatively low values and with varying signs suggesting that the probabilities are not correlated.

Table 4.13. Correlation between conditional probabilities of different outcomes (1-5, 6-10 and >10 falls)

Conditional probabilities		Correlation coefficient
No treatment	P(x>0), P(x>5|x>0)	-0.05
	P(x>0), P(x>10|x>5)	-0.05
	P(x>5|x>0), P(x>10|x>5)	0.17
Treatment within trial	P(x>0), P(x>5|x>0)	-0.11
	P(x>0), P(x>10|x>5)	0.1
	P(x>5|x>0), P(x>10|x>5)	0.27
Treatment after the trial	P(x>0), P(x>5|x>0)	-0.07
	P(x>0), P(x>10|x>5)	0.08
	P(x>5|x>0), P(x>10|x>5)	0.12

4.5.1.2. The rate of falls derived from experts' priors accurately represent their' beliefs

Section 4.2.2 explained that the method used to derive the rate of falls required predicting the probabilities of having 1, 2, 3, etc. falls, and that the derived probabilities of having at least one fall, more than five falls and more than ten falls may deviate from experts' expressed beliefs. In order to assess whether such discrepancies existed, five experts from the sample were chosen at random, using the random number generator in R. For these five experts the elicited probabilities were compared with those predicted by the regression model in Equation 3.2. Details of the methods are described in section 4.2.6.1, while the results are shown in Table 4.14.

Of the fifteen analysed priors (three priors from five experts), only three probabilities differed by more than two percentage points -these are highlighted in the table. The highest difference between predicted and elicited probabilities was 0.07 (or 7 percentage points); Expert 19 believed that the probability of falling between one and five times was 0.217 (mean) whereas the predicted probability was 0.143.

Table 4.14. Elicited and predicted mean probabilities of 1-5, 6-10 and more than 11 falls (range) for five experts chosen at random.

		P(1-5 falls)	P(6-10 falls)	P(>11 falls)
Expert 11	**Elicited**	0.186 (0.141-0.260)	0.038 (0.022-0.073)	0.012 (0.005-0.034)
	Predicted	0.173 (0.118-0.263)	0.045 (0.025-0.088)	0.011 (0.004-0.030)
Expert 15	**Elicited**	0.267 (0.201-0.360	0.043 (0.022-0.090)	0.008 (0.002-0.027)
	Predicted	0.258 (0.171-0.385)	0.047 (0.022-0.102)	0.007 (0.002-0.024)
Expert 19	**Elicited**	0.217 (0.201-0.250)	0.010 (0.009-0.011)	0.007 (0.007-0.008)
	Predicted	0.143 (0.132-0.166)	0.026 (0.024-0.031)	0.004 (0.004-0.005)
Expert 34	**Elicited**	0.396 (0.101-0.800)	0.145 (0.017-0.451)	0.111 (0.013-0.367)
	Predicted	0.356 (0.080-0.764)	0.195 (0.030-0.546)	0.095 (0.009-0.320)
Expert 36	**Elicited**	0.166 (0.101-0.300)	0.042 (0.021-0.108)	0.012 (0.004-0.043)
	Predicted	0.161 (0.090-0.310)	0.045 (0.022-0.11)	0.011 (0.004-0.040)

Section 4.2.2 explained that two alternative methods for deriving rates were explored. The rates derived using the three different methods are presented in Table 4.15.

When the probabilities for each number of falls are assumed to be decreasing, the direct method and predicted probabilities yields very similar results to those obtained using regression – the maximum difference in rates was 0.09 for Expert 34. Assuming an equal probability of all outcomes within a category resulted in higher rates than with the other two methods, as it assigned greater probability to higher number of falls within each category than the other two methods.

Table 4.15. Rates of falls derived using three different methods.

	Mean rate of falls derived using different methods		
	Uniform distribution assumes across outcomes of the same category	**Decreasing probabilities assumed across outcomes of the same category**	**Predicted probabilities of each outcome**
Expert 11	0.89 (0.59-1.54)	0.74 (0.47-1.28)	0.73 (0.47-1.28)
Expert 15	1.11 (0.74-1.85)	0.90 (0.57-1.56)	0.90 (0.57-1.55)
Expert 19	0.79 (0.63-0.91)	0.56 (0.52-0.65)	0.50 (0.46-0.58)
Expert 34	3.30 (0.59 -9.12)	2.73 (0.46-7.7)	2.82 (0.45-7.76)
Expert 36	0.86 (0.46-1.86)	0.72 (0.39 -1.61)	0.71 (0.38 -1.57)

4.5.1.3. The treatment effect of the podiatry intervention is independent of the rate of falls and risk of fractures in patients who do not receive the intervention

The correlation coefficients between conditional probabilities are shown in Table 4.16.

The correlation coefficient between outcomes and the treatment effect ranged from -0.35 to -0.06. The coefficient is always negative, indicating the higher the risk of falls and fractures, the lower the treatment effect. However, the coefficient is relatively low, indicating a weak correlation.

Table 4.16. Correlation between outcomes and the treatment effect on those outcomes.

Treatment effect		Correlation coefficient
Within trial (after 1 year)	RR for P(x>0)	-0.21
	RR for P(x>5|x>0)	-0.19
	RR for P(x>10|x>5)	-0.32
	RtR	-0.10
	OR for fractures	-0.16
Treatment effect after the trial	RR for P(x>0)	-0.28
	RR for P(x>5|x>0)	-0.32
	RR for P(x>10|x>5)	-0.35
	RtR	-0.06
	OR for fractures	-0.14

4.5.1.4. Temporal change in the treatment effect is linear

Section 4.2.6 described that the plausibility of the assumption that the rate of change in the treatment effect is linear was explored by observing the predicted treatment effect in those experts who believed the treatment effect would diminish, at time point $t3$ when they expected the treatment effect to diminish completely. The results are shown in Table 4.17.

Twenty three experts believed the treatment effect on the rate of falls would diminish over time. Constant rate of depreciation led to a treatment effect between 0.8 and 1.2 in 6 of those 23 experts. Thirteen experts believed the treatment effect would diminish (TE = 1) before $t3$.

Priors elicited from 8 experts suggested that the treatment effect on the risk of fractures would diminish. Constant rate of depreciation led to a treatment effect between 0.8 and 1.2 in all 8 experts. However, none of the predicted treatment effects had reached 1 (diminished completely) by that point, suggesting it would take longer for the treatment effect to diminish completely.

Table 4.17. Predicted treatment effect at *t3* derived from priors that indicated the treatment effect would diminish over time.

	N	$0.8 < TE_{t3} < 1.2$	*TE diminishes by t3*
Rate ratio (falls)	23	6	13
Relative risk (fractures)	8	8	0

In addition, the implications of assuming a log-linear change in the treatment effect (derived using Equation 4.7 in section 4.2.5) were also explored. The annual change in the treatment effect ΔATE was then applied to the treatment effect, using Equation 4.11.The predicted treatment effect at *t3* were judged to be implausibly high – for one expert who believed the treatment would increase the risk of fractures, and the treatment would potentiate over time, the predicted relative risk of fractures after 5 years was 3125. In another expert, who believed the treatment would decrease the risk of fractures, but the treatment effect would depreciate over time, the predicted relative risk of fractures after five years was 92. Since assuming a log change in the treatment effect led to predictions that were judged to be implausible, the change was assumed to be linear in the rest of the thesis.

4.5.2. Visual evaluation of experts' priors

All priors were visually inspected to identify any features that could indicate that experts had difficulty completing the task.

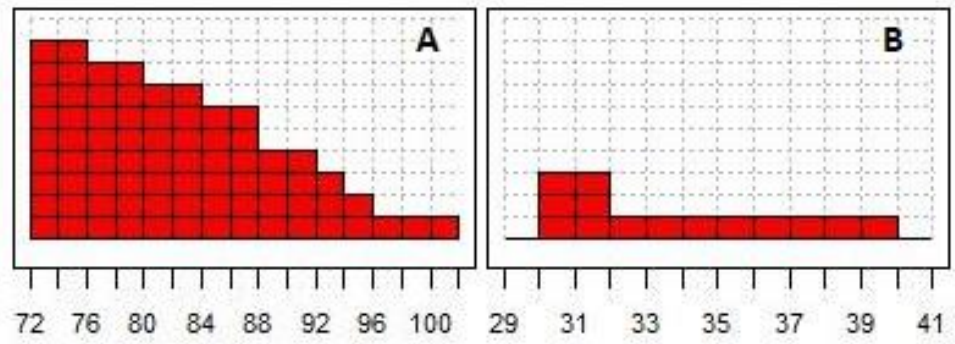
Experts' used a range of shapes to express their beliefs. The number of bricks used varied from 1 to 80 and the shapes included uniform, bimodal, bell-shapes and skewed distributions.

One expert provided a very narrow range for the non-domain seed, suggesting that 28 to 30 days every September were rainy, and placed one chip in the 30-31 bin suggesting they were certain the number of rainy days every September (out of 30) was between 30 and 31. The same expert provided a range of values for every question but only used one chip in every histogram suggesting they were disengaged. One additional expert completed all questions on the probabilities of falling, and provided ranges for the odds of fracture, but did not utilise the entire grid to express their uncertainty around the odds of fracture – they only placed one brick in each grid.

Four additional features were identified that could indicate that experts had difficulty completing the exercise:

- Bimodal priors. While all distribution shapes are theoretically possible, bimodal distribution of the mean is unlikely, yet three experts used bimodal priors to express their beliefs.
- Probability placed in bins outside the stated range. Experts were trained that the range should include values so that they believe that the quantity of interest is highly unlikely to be outside this range; placing bricks that represent relatively high probabilities outside their stated range could indicate that either the range or the histogram did not represent experts' beliefs accurately. In the REFORM elicitation study 20 experts placed more than 0.05 probability outside their range, and 9 did it consistently, in more than three of their priors.
- Mode close to the end of the range. An example of such distribution is shown in Figure 4.14, where Expert A placed the greatest number of chips in the first bin, suggesting that there was a 0.24 probability (18/76 chips) that the risk of falling at least once was between 0.072-0.076 and zero probability that it was any lower than 0.072. It is difficult to judge when this is implausible – the chips represent blocks of probability, thus the sensitivity of the chips and bins methods is limited.
- Very narrow range. For example, in Figure 4.14, Expert B expressed a very narrow range of patients who may suffer a fall – the risk of falling derived from their priors was between 3 - 4% (30/1000-40/1000). However, it is not clear how narrow a range is too narrow.

Figure 4.14. Examples of priors on the risk of falling. The x-axis represents the number of people out of 1000 who may suffer a fall.



Furthermore, experts' internal consistency was measured by comparing the temporal change in the rate ratio and the relative risk of fractures derived from their priors with their

responses to MCQs about the same. The definition of internal consistency is provided in section 4.2.6.

Table 4.18 shows the results. Experts' priors on the risk of fractures were more likely to be consistent with their verbal responses, likely because it was a less complex parameter. Out of the 36 experts in the sample, 19 were consistent on both measures of treatment effect.

Table 4.18. Experts' internal consistency

Experts' beliefs about the treatment effect after the trial, expressed verbally in MCQs	**N**	**Number of experts whose priors were inconsistent with their verbal responses**		**Number of experts whose priors were <u>not</u> inconsistent with their verbal responses**
		Rate ratio (N=36)	**Relative risk of fractures (N=32)**	
The effect of the intervention is most likely to stay the same although the expert is not certain.	1	1	0	0
The intervention is most likely to become less effective over time.	31	8	8	18
The intervention is most likely to become more effective over time.	2	1	0	0
The intervention effect is likely to change over time, but the expert is not certain whether it will get more or less effective.	2	1	0	1
Total	36	11	8	19

4.6. Summary

This chapter provided an overview of the results of the elicitation exercise, conducted as part of the REFORM elicitation study. The chapter had four objectives:

- To describe how experts' elicited quantities were used in the analysis;
- To describe the sample of experts who took part in the elicitation;
- To give an overview of what experts' priors suggest about their beliefs;
- To evaluate the elicited priors, and the methods used to analyse them.

Section 4.2 describes how the parameters required in the cost-effectiveness analysis (Chapter 6) were derived from experts' elicited quantities, given the indirect elicitation methods employed in the study (objective 1). Section 3.6 in Chapter 3 has highlighted that

it is not clear whether the employed methods were optimal for eliciting the required parameters – they were chosen because they were considered more intuitive for experts than direct elicitation of the treatment effect and the change in the treatment effect.

Section 4.3 described the sample of experts who took part (objective 2). Overall the sample size was relatively large (n=41, compared to the average sample size of 8.83 in CEDM (Soares *et al.*, 2018)) and within the target set out in Chapter 3. Experts from all targeted professions were recruited, although the sample of academics was small (n=1), likely because of a smaller pool of experts available to recruit from. One additional expert who was not in the listed occupations was recruited on recommendation – they had experience in fall prevention in a specific patient subpopulation.

Section 4.4 gave an overview of what experts' priors suggest about their beliefs (objective 3). Experts predominantly thought that the podiatry intervention in the REFORM trial would be effective, and that the treatment effect would diminish over time, although there was some variation between experts.

The accuracy of the reported priors is analysed and discussed in Chapter 5, but their coherence and internal consistency were assessed in section 4.5.2. The results were varied - internal consistency in priors on P(fracture|fall) was better than in the priors on the rates of falls, possibly because it was a simpler parameter to elicit - the change in the rate ratio was derived from nine elicited priors compared to the change in the risk of fractures that was derived from three priors. Incoherence and inconsistency in priors are likely to be a common challenge when elicitation is delivered remotely, Chapter 6 explores the effect of including inconsistent experts in the aggregate prior.

Section 4.5.1 evaluated the assumptions imposed by the elicitation methods:

1. Independence between conditional probabilities;
2. The rate of falls derived from experts' priors accurately represent their' beliefs;
3. Independence between the treatment effect of the podiatry intervention and the baseline rate of falls and risk of fractures;
4. Linear change in the treatment effect.

The outlined assumptions can affect the validity of the results – if the assumptions do not hold, then the derived probability distributions of the rate of falls, the treatment effect, and the change in the treatment effect don't represent experts' uncertainty, their score's don't represent their accuracy, and any effect of experts' characteristics on their scores is invalid.

The analysis found no evidence that assumptions 1-3 were implausible – there was no correlation between experts' assessments of conditional probabilities (section 4.5.1.1), no correlation between baseline falls and fractures and the treatment effect (section 4.5.1.3), and the methods for deriving the rate of falls predicted similar probabilities of falling 1-5, 6-10 and more than 11 times to those suggested by experts (section 4.5.1.2).

The plausibility of assumption 4 is less clear. In section 4.5.1.4 a linear diminishing effect of the treatment effect was applied to the risk of fractures elicited from 8 experts, whose priors suggested that the treatment effect would diminish. All eight predicted treatment effects at time t3 were between 0.8-1.2 suggesting that the linear change was a plausible assumption. For the rate ratio the 26 experts believed that the treatment effect would diminish. Predicted treatment effect at time t3 was between 0.2 and 1.2 in only six of those 26 experts. The majority of priors (13) suggested that the treatment effect would diminish before t3.

Chapter 5. Exploring factors that motivate experts priors: results of the REFORM elicitation study

5.1. Introduction

Chapter 2 identified four factors that have been proposed to affect the accuracy of experts' priors in the literature: field-specific knowledge and experience (substantive expertise), their perspective, their ability to express their beliefs in the required format (normative expertise), and their ability to make accurate probabilistic assessments of their uncertainty. When experts' priors are aggregated mathematically, the investigator can weight priors elicited from individual experts if they believe that some experts should contribute more towards the overall prior.

In Chapter 2, two general approaches for deriving weights were identified: 1) based on experts' observed characteristics and 2) based on their measured performance in elicitation. Both aim to improve the accuracy of the aggregate prior but they differ in the way that they capture experts' 'contribution'. Chapter 2 also highlighted that both methods were associated with limitations, and their effectiveness in improving the accuracy of the aggregate prior is uncertain.

Chapter 3 described the protocol of the REFORM elicitation study, designed to compare different weighting methods, while Chapter 4 provided an overview of the results of the elicitation exercise. This chapter uses the results of the study to explore factors that affect experts' priors.

In particular, the following three objectives were set in this chapter:

1) To score experts' priors;
2) To define the characteristics to be used as proxies for the factors believed to affect experts' priors in the REFORM elicitation study: substantive expertise, perspective, normative expertise, and the ability to make accurate probabilistic assessments.
3) To explore whether the identified characteristics explain variation in priors elicited from different experts.

The findings can be used to determine what is captured by different weighting methods and how these methods should be applied. For example, if priors are predominantly affected by experience (substantive expertise and perspective) then weights should be either based on characteristics (in particular substantive expertise) or experts' performance on domain seeds so that scores reflect experts' experience with the target variable. If, however, they are predominantly affected by expert' ability to make accurate probabilistic assessments, then performance-weighting may be preferred, and seeds may not have to be domain specific. Furthermore, understanding factors that affect experts' priors can resolve methodological uncertainties in other steps of the elicitation process, such as how to define experts for elicitation, i.e. what type of experience improves their accuracy.

Section 5.2 describes the methods employed to achieve the objectives of the Chapter and sections 5.3 - 5.5 analyse the results. Section 5.3 shows experts' scores, section 5.4 provides an overview of experts' characteristics and section 5.5 then explores whether variation between priors elicited from different experts is explained by the captured characteristics. Section 5.6 then summarises the findings.

5.2. Methods

Section 5.1 set out three objectives for this Chapter. Sections 5.2.1 – 5.2.3 describe the methods used to achieve each objective, in turn.

5.2.1. Scoring experts' elicitation performance

In total, 13 seeds were used to assess experts' performance: one non-domain seed regarding rainfall in York and the following 12 domain seeds:

1) Elicited P(x>0) in control arm,
2) Elicited P(x>5|x>0) in control arm,
3) Elicited P(x>10|x>5) in control arm,
4) Rate of falls in control arm, derived from experts' priors on 1-3,
5) Elicited odds of having a fracture,
6) Derived P(fracture|fall) derived from experts' priors on 5,
7) Relative risk of Elicited P(x>0), P(x>5|x>0), P(x>10|x>5),
8) Odds ratio for fractures,
9) Rate of falls ratio,
10) Relative risk of fractures,

The rationale for the use of each seed and the methods used to derive them were discussed in Chapter 4. This section describes the scoring methods, by discussing how the reference values of each seed were obtained (section 5.2.1.1), how the scoring methods were selected (section 5.2.1.2), and how the scores were derived (sections 5.2.1.3 and 5.2.1.4).

5.2.1.1. Reference values of the seeds

The reference value of the non-domain seed, to which experts' priors were compared, was the average number of rainy days in September recorded by the Met Office between 1980 and 2010, at the Lynton-on-Ouse weather station. The reference values of the domain seeds were observed in the REFORM trial.

In order to take into account the uncertainty surrounding the value of each seed, parametric probability distributions were fitted to the observed data. The observed values of each seed and their fitted distribution are shown in Table 5.1.

Table 5.1. Probability distributions fitted to each trial outcome. Methods for deriving RR, RtR and OR are described in Chapter 4.

Parameter	Observed value (95% CI)	Probability distribution of trial outcome
P(x>0)	0.585 (0.542-0.627)	Beta ($n_{x>0}$, $n_{x=0}$)
P(x>5|x>0)	0.088 (0.058-0.122)	Beta ($n_{x>5}$, $n_{0<x<6}$)
P(x>10|x>5)	0.192 (0.068-0.361)	Beta ($n_{x>10}$, $n_{5<x<11}$)
Rate of falls	1.57 (1.366-1.777)	Gamma (μ^2/σ^2, σ^2/μ)
Odds of fracture	55.9 (42-71)	Poisson (odds)
P(fracture|fall)	0.018 (0.010-0.027)	Beta ($n_{f>0}$, $n_{f=0,x>0}$)
Risk ratio for P(x>0)	0.915 (0.819-1.022)	N (log(RR), $\sigma_{RR}/\sqrt{\boldsymbol{n}}$)
Risk ratio for P(x>5|x>0)	1.014 (0.594-1.733)	N (log(RR), $\sigma_{RR}/\sqrt{\boldsymbol{n}}$)
Risk ratio for P(x>10|x>5)	1.809 (0.688-4.754)	N (log(RR), $\sigma_{RR}/\sqrt{\boldsymbol{n}}$)
Rate of falls ratio	0.906 (0.818-1.003)	N (log(RtR), $\sigma_{RtR}/\sqrt{\boldsymbol{n}}$)
Odds ratio for fractures	0.751 (0.364-1.550)	N (log(OR), $\sigma_{OR}/\sqrt{\boldsymbol{n}}$)
Risk ratio for fractures	1.679 (0.569-2.790)	N (log(RR), $\sigma_{RR}/\sqrt{\boldsymbol{n}}$)

$n_{x>0}$ = number of patients who fell at least once;

$n_{x=0}$ = number of patients who did not fall;

$n_{x>5}$ = number of patients who fell more than five times

$n_{0<x<6}$ = number of patients who fell 1-5 times;

$n_{x>10}$ = = number of patients who fell more than ten times;

$n_{5<x<11}$= number of patients who fell 6-10 times;

$n_{f>0,x>0}$ = number of patients who had a fracture;

$n_{f=0,x>0}$ = number of patients who fell but did not have a fracture.

5.2.1.2. Choosing the scoring methods

Chapter 2 suggested that the optimal scoring method depends on the type of parameter used (discrete or continuous), whether precision in priors is desirable, and the certainty with which the seed is known. All seeds elicited in the REFORM elicitation study are continuous or ordinal (number of rainy days) variables. Experts' were asked to assess the outcomes of a clinical trial; it was assumed that they could not know the value of the seed with more certainty than that derived from the trial results, and so the scoring method needed to penalise overconfidence. Finally precision was judged to be an important indicator of experts' performance, and so the scoring method needed to penalise imprecision. In Chapter 2 (Table 2.9 in section 2.4.3.2) KL scores were proposed to penalise uncertainty, overconfidence, and can be used on continuous variables, and so they were utilised in this chapter.

Chapter 2 proposed that experts' performance depends on their bias and precision. While the KL score is affected by both it does not reflect specific characteristics of priors. Secondary measures of bias and precision were thus used to understand what determined experts' scores.

The next two sections describes how KL scores and the secondary measures of bias and precision were derived.

5.2.1.3. Deriving KL scores

Chapter 2 (section 2.2.2) described the methods for deriving KL scores. To summarise, the scores are derived using Equation 5.1 (copy of Equation 2.1 in Chapter 2) (Cooke, 1991).

$$I(S,P) = \sum_{i=1}^{M} S(i) ln \frac{S(i)}{P(i)}$$ **Equation 5.1**

Where i is one of M outcomes,

$S(i)$ is the observed probability of i,

$P(i)$ is the expert's probability of i.

In the Classical Model each i represents a seed question for which a point estimate is observed. In this chapter, in order to account for seed uncertainty, i represents an interval of the possible values of the seed.

The width of the intervals compared could affect experts' scores. Wide intervals could fail to capture subtle differences in experts' performance, and so the lowest possible bin width was chosen as the i for each parameter.

For the non-domain seed, i was the number of rainy days so that M=31, and $i \in$ {0,1,2,…,30}. The probabilities that an expert placed on each i were then compared to the integral of the beta distribution between two values – for example, the integral between 0 and 1 days (0 and 1/30), 1 and 2 days (1/30 and 2/30) and so on.

When the proportion of fallers were scored, i was the number of patients out of 1000 (or a proportion of 0.001) so that M=1000 and $i \in$ {0, 0.001, 0.002,…,1}

Intervals for the probabilities elicited in the control arm were probabilities of 0.001 and for odds they were 1, as these was the narrowest possible bin width for those quantities. The treatment effect was derived from elicited priors on the control and treatment arms, and so there was no minimum bin width that could be used as an interval. Several different interval widths were tested and width of 0.01 was chosen for the analysis as it gave the most reasonable number of intervals (20-406) for which probabilities could be compared.

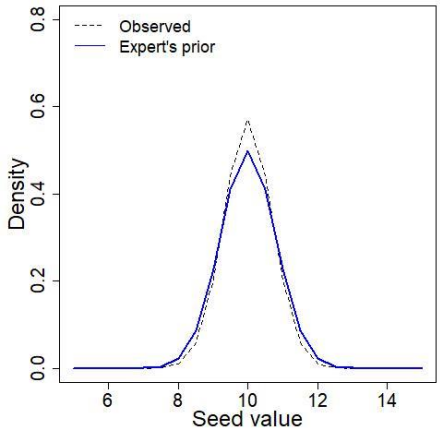
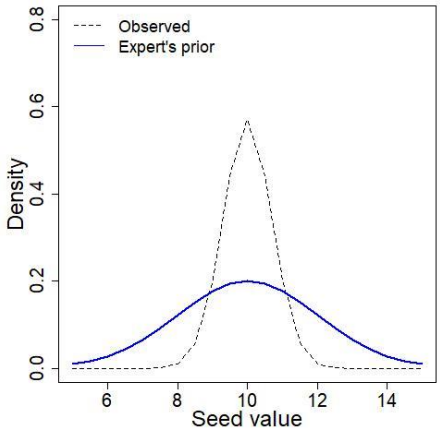
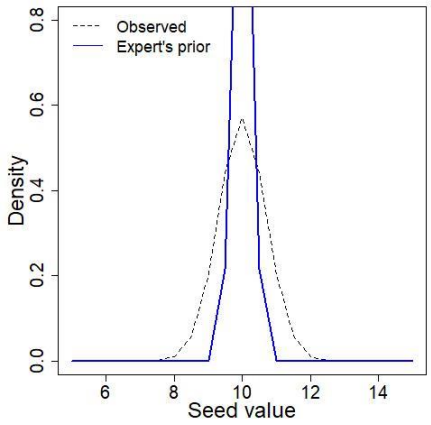
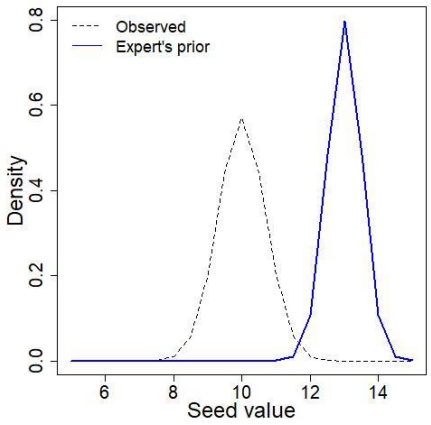
When the interval of parameter values was narrower than the expert' bin width, a uniform distribution across the bin width was assumed. Uniform distributions were chosen instead of fitting parametric probability distributions because fitting was likely to lead to varied goodness of fit across priors. If fitting affected the scores of some priors more than others, then the resulting scores would have been less representative of that experts' performance, potentially biasing the results of the analysis.

The KL scores can be difficult to interpret. In order to gain insight into how much experts' accuracy varied between different seeds, for each domain seed a prior that scored close to the mean KL score for that seed was identified. The selected priors were then compared qualitatively by plotting them (and the observed value of each seed) in a histogram and noting any differences between them.

5.2.1.4. Methods for scoring experts' bias and precision

As described in Chapter 2, experts' priors vary in bias and precision, where bias refers to a tendency for experts' priors to be consistently higher or lower than the observed relative frequency, whereas precision refers to experts' certainty about their priors. In order to understand what drives experts' KL scores, their bias and precision were measured using CICP scores. Chapter 2 highlighted that CICP scores represent the proportion of the observed probability distribution that experts' place positive probability on (see section 2.2.2 in Chapter 2 for details), while CICP precision scores represent the probability that experts place on the observed probability distribution (see section 2.4.3 in Chapter 2). The implications of different combinations of scores are shown in Table 5.2.

Table 5.2. Implications of different combinations of CICP accuracy and CICP precision scores.

		CICP precision score	
		High	Low
CICP accuracy score	High	Unbiased, precise 	Unbiased, imprecise 
	Low	Unbiased, overconfident 	Biased, overconfident 

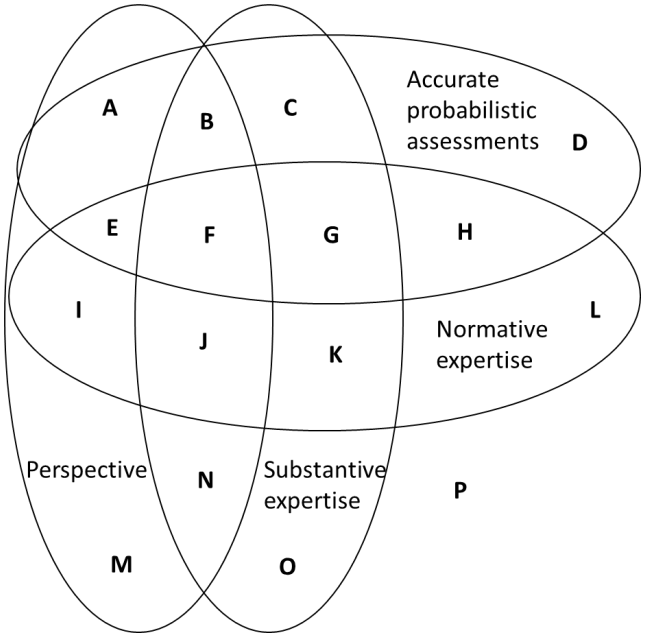
5.2.2. Defining the proxies for the factors proposed to affect experts' priors

Chapter 2 suggested that experts' priors are affected by:

- Substantive expertise,
- Perspective,
- Normative expertise,
- Their ability to make accurate probabilistic assessments.

The interaction between these four factors is shown in Figure 5.1; the definition of each factor and the interactions between them were discussed in Chapter 2 (section 2.3).

Figure 5.1. Venn diagram showing the interaction between different factors believed to affect their priors. (Copy of Figure 2.3 in Chapter 2).



The reform elicitation study explored characteristics that can be used as proxies for each of these factors, in order to understand to what extent they influence experts' elicitation performance. Sections 5.2.2.1 - 5.2.2.4 described the methods for measuring each of the four factors.

5.2.2.1. Substantive expertise

As outlined in Chapter 3, substantive expertise was measured using information on experts' professional experience, collected as part of the REFORM elicitation study. As described in Chapter 3 (section 3.4.1) the following characteristics were recorded:

1) Role, recorded in free text.
2) Research experience determined by two categorical variables: the number of publications (0-3, 4-20, 21-50, or more then 50) and the number of successful research grant proposals co-written (0, 1-5, more than 5).
3) Time spent with patients who are at an increased risk of falling, either helping them prevent falls or treating fall related injuries. Experts indicated whether they spent less than 10%, 10-30%, 30-50% or more than 50% of their time with the relevant patient population.

4) Awareness of research into podiatry interventions designed to reduce the risk of falls in the elderly (yes/no).

Points 1) and 2) were used to reflect experts' experience or seniority, while 3) and 4) reflected how specialised they are in the field of falls prevention.

5.2.2.2. Perspective

Experts' perspective was determined by their profession indicated in the MCQ prior to the elicitation exercise. The perspectives thus included: physiotherapists, geriatricians, nurses and occupational therapists, and academics.

5.2.2.3. Normative expertise

Chapter 4 evaluated experts' priors and identified five possible indicators that experts' lacked normative expertise: bimodal priors, chips in bins outside the stated range, mode close to the end of the range, very narrow ranges and internal inconsistency, defined as inconsistency of experts' priors on the treatment effect after the trial with their verbal responses (in MCQs) about the same.

Three of these characteristics were used as indicators of normative expertise:

- Bimodal priors. While all distribution shapes are theoretically possible, a bimodal distribution of the mean is unlikely and so they were used as an indicator of a lack of normative expertise.
- Chips in bins outside the stated range. While probabilities outside the stated range are not statistically incoherent, experts were trained that the range should include values so that they believe that the quantity of interest is highly unlikely to be outside this range. Placing bricks that represent more than 5% of their probability outside their stated range suggested that either the range or the histogram did not represent experts' beliefs accurately, and so it was used as an indicator of a lack of normative expertise.
- Internal consistency. A lack of internal consistency indicates that experts' priors do not represent their beliefs. The definition is described in detail in section 4.2.6 in Chapter 4.

Experts were classified as normative or not normative depending on whether they exhibited the above characteristics or not.

Mode close to the end of the range and very narrow ranges were not used to define normative expertise as it was difficult to determine a precise definition of implausible priors using the described features – for example, it is not clear how narrow a range is implausibly narrow.

5.2.2.4. Ability to make accurate probabilistic assessments

As outlined in Chapter 3, experts' ability to make accurate probabilistic assessments was based on their prior on the non-domain seed. Details on the non-domain seed were provided in section 3.4.3. Experts' priors were scored using CICP scores, penalising only experts who were biased and overconfident. For details on how to derive CICP scores, see section 2.2.2 in Chapter 2.

5.2.3. Exploring the relationship between experts' characteristics and their priors

The aim of this chapter was to measure the extent to which each of the four factors affected experts' accuracy in elicitation.

The effect of experts' characteristics was explored using two types of seeds: non–domain, about the number of rainy days in York, and 12 domain seeds about REFORM trial outcomes. For each type of seed different characteristics were used to reflect the factors believed to affect their priors. Sections 5.2.3.1 and 5.2.3.2 describe the methods for measuring the effect of experts' characteristics on their priors on non-domain and domain seeds in turn.

5.2.3.1. Exploring factors that affect experts' priors on the non-domain seed

The non-domain seed question was 'Out of 30 days in September, how many days does it rain in York, on average?'

The effect of three characteristics on experts' priors on non-domain seeds were explored:

- Normative expertise, measured using statistical coherence and internal consistency, as described in the previous section.
- The ability to draw inference, measured using the inference section from the Glacier Watson (GW) test for adaptive reasoning (see section 3.4.3 in Chapter 3 for details).
- Location of recruitment

The ability to draw inference is one of the requirements for accurate probabilistic assessment and so it was important to ascertain if this characteristic can be used as a proxy for it.

Location of recruitment was used to reflect experts' substantive expertise and perspective. While the aim was to use a seed independent of their field of expertise (fall prevention), some experts may be more familiar with York weather than others and so can be considered either to be more substantive, or to have a perspective resembling the elicited parameter more closely. In order to take into account this potential variation in substantive expertise and perspective, experts' venue of recruitment was noted, and priors elicited from experts who were recruited in the Y&H region were compared to those elicited from experts recruited elsewhere.

Including expert' normative expertise, GW inference scores and their venue of recruitment meant that the effect of three characteristics was explored. Two of those were binary (region of recruitment and normative expertise) and one was ordinal (GW inference scores could take values between 0 and 4). Given the relatively small sample size (target sample size was 30-50), the GW test scores were converted into a binary variable of scores ranging 0-2 and 3-4. The effect of using different cut-offs was explored in a sensitivity analysis.

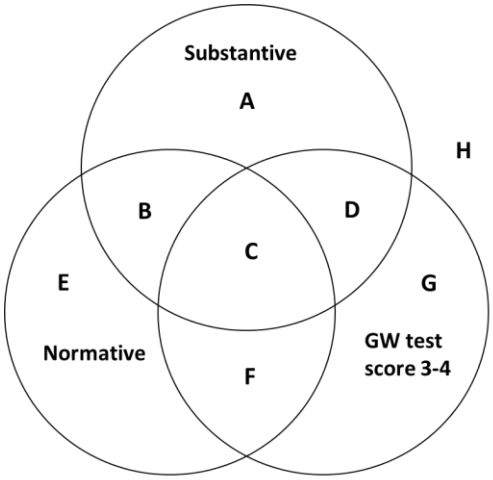
With an unlimited sample size the comparison would be performed using regression analysis where the calibration score would be the independent variable and each characteristic would be an independent variable. However, the sample size in the REFORM elicitation study was unlikely to afford sufficient power to detect the effect of each characteristic using regression analysis. If each of the three factors thought to affect experts' priors were represented using one binary variable, there would be 8 (2^3) possible combinations of characteristics to compare. In a sample of up to 50 experts (as was the target population size in the REFORM elicitation study) there would be, on average, up to 6.25 (50/8) experts with each combination of characteristics. This is a relatively small sample size and regression analysis would be unlikely to detect any statistically significant results.

Instead, a stepwise approach was used to observe the effect of specific characteristics on experts' scores and patterns of belief. Experts with a specific set of characteristics were grouped together and their mean score was derived. In each step, a different set of characteristics was compared. Comparison of individual characteristics does not lead to conclusive results as any variation in scores could be confounded by other characteristics.

For example, if substantive experts also tend to be more normative, then comparison of scores between substantive and not substantive and normative and not normative experts could lead to the same conclusion but it cannot be said which of the two caused the difference. Interaction between different characteristics can further complicate analysis – for example, the effect of normative expertise can depend on experts' substantive expertise. Comparison of individual characteristics and combinations of characteristics in a stepwise manner allowed exploration of characteristics that consistently led to higher or lower scores.

The comparison of three binary characteristics resulted in 8 combinations of characteristics in total and 7 possible comparisons as shown in Figure 5.2.

Figure 5.2. Venn diagram showing possible combinations of experts' characteristics and resulting comparisons in the analysis.



Characteristics compared	Areas in the diagram
Substantive vs not substantive	ABCD vs EFGH
Normative vs not normative	BCEF vs ADGH
GW test score 0-2 vs 3-4	CDFG vs ABEH
Substantive and normative	AD vs BC vs EF vs GH
Substantive and GW test score	AB vs CD vs FG vs EH
Normative and GW test score	EF vs CF vs DG vs AH
All three characteristics	A vs B vs C vs D vs E vs F vs G vs H

A=substantive, not normative, GW test score below 3; B=substantive, normative, GW test score below 3; C= Substantive, normative, GW test score 3-4; D=substantive, not normative, GW test score 3-4; E= not substantive, normative, GW test score below 3; F= not substantive, normative, GW test score 3-4; G= not substantive, not normative, GW test score 3-4; H= not substantive, not normative, GW test score below 3.

In the analysis, all seven comparisons in the table in Figure 5.2 were explored in two steps:

- The Kruskal-Wallis test was used to detect whether differences in scores between groups were statistically significant. The Kruskal-Wallis test of statistical significant was used as it is the recommended test for non-parametric comparison of 3 or more groups in statistical literature (Corder and Foreman, 2009).

- Mean scores for all experts with similar characteristics were compared by visual inspection. Visual comparison was used because the sample size was unlikely to be sufficiently large to spot a statistically significant difference, as discussed earlier in this section, yet the results could still provide a useful insight into factors that affected experts' priors.

Visual inspection involved identifying any characteristics that consistently led to higher or lower scores. For example, with reference to Figure 5.2, if substantive expertise (A,B,C,D) scored higher than non-substantive experts (E,F,G,H), the effect of substantive expertise in other comparisons was observed – e.g. whether AD (substantive and normative) experts also score higher than EF (not substantive but normative) and BC (substantive but not normative) higher than GH (not substantive and not normative), and so on.

5.2.3.2. Exploring factors that affect experts' priors on the domain seeds

The effect of the four characteristics on experts' priors on domain seeds were explored: substantive expertise, perspective, normative expertise, and their ability to make accurate probabilistic assessments. The methods for capturing each factor were described in section 5.2.2.

The effect of experts' characteristics on their priors was explored in two stages. Stage 1 involved measuring the effect of substantive and normative expertise, and accuracy of probabilistic assessments. The role of perspective was not explored in stage 1 as it was not clear which professionals were the most familiar with the elicited parameter; in other words, it was not possible to determine whether experts represent areas B, F, J and N, or areas C, G K and O in Figure 5.1. Stage 2 then compared priors elicited from different professionals to explore the extent to which experts' perspective affected their priors. Each stage is described here in further detail.

<u>Stage 1: The effect of substantive and normative expertise, and accuracy of probabilistic assessment on experts' priors</u>

The REFORM elicitation study collected information about four categorical variables that defined experts' substantive expertise (see section 5.2.2.1 for details), leading to a minimum of 144[12] possible combinations of characteristics. This number increased to 576 if

[12] A minimum of two types of role, three categories of publications, three categories for experience in research grant proposal writing, four categories for the amount of working hours spent with target patient population and two categories for awareness of research into podiatry interventions for fall prevention. 2x3x3x4x2=144

normative expertise and the ability to make accurate probabilistic assessments were taken into account. The final sample size in the REFORM elicitation study was 41 (as reported in Chapter4), and so it was likely to be too small to detect the effect of all six variables, and interactions between them. Instead, the information collected was used to classify experts as substantive or not substantive, normative or not normative and accurate or inaccurate when making probabilistic assessments.

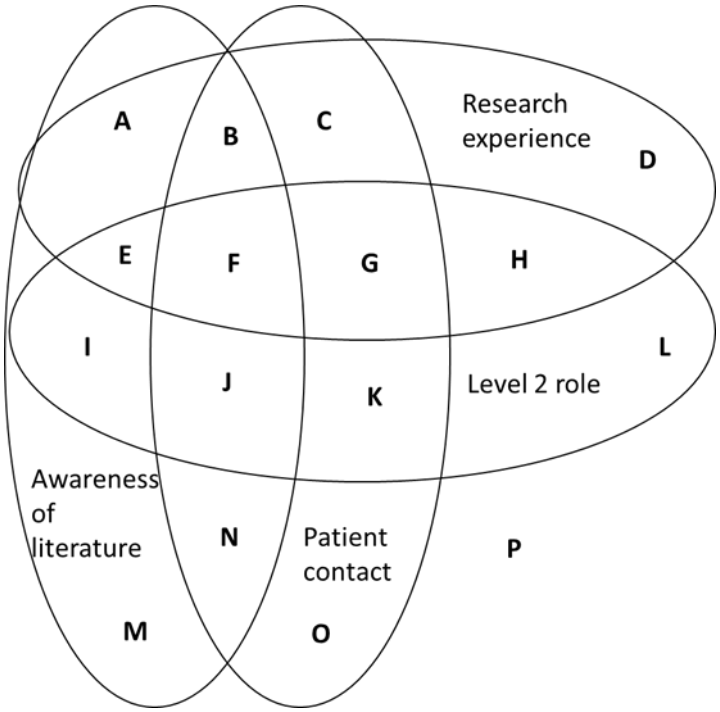
Normative expertise was a binary variable, as described in section 5.2.2.3.

In order to derive a binary variable for substantive expertise, first experts' role, research experience and patient contact were simplified into binary variables, as follows:

- Experts' role was categorised as Level 1 and Level 2, where Level 2 included specialist clinical roles such as consultant geriatricians, and other healthcare professionals specialising in fall prevention, and Level 1 included clinicians who do not have a specialist role, such as physiotherapists without a speciality and trainee geriatricians.
- Research experience was simplified into 'research experience' and 'no research experience'.
- Patient contact was defined as less than or more than x% of time spent with the stated patient population.

This resulted in 16 possible combinations of characteristics, represented by letters A-P in Figure 5.3**.** Experts were then classified as substantive if they were in a Level 2 role, had research experience, spent over x% of their time with the relevant patient population and were aware of research into podiatry interventions designed to reduce the risk of falls, represented by the area F in Figure 5.3. The exact definition of research experience and the time spent with patients were determined in the analysis based on the sample, to ensure at least six experts in the sample satisfied all four criteria, possessed normative expertise and made accurate probabilistic assessments. Six was chosen to represent a sample of substantive experts that may be recruited for an elicitation exercise, considering that the average sample size in elicitation in HTA was 8.83. (Soares *et al.*, 2018)

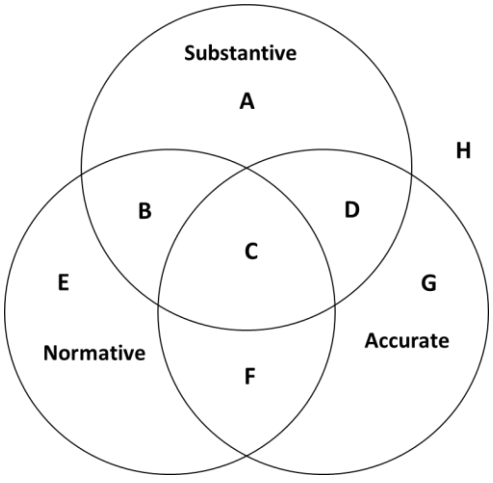
Figure 5.3. Venn diagram showing interaction between different measures of experience and speciality.



The accuracy of experts' probabilistic assessments was converted into a binary variable by classifying experts whose CICP score on the domain seed was less than 0.5 as inaccurate, and those whose score was higher than 0.5-1 as accurate.

In the analysis the mean score derived from experts who were substantive, normative and made accurate probabilistic assessments were compared to those who were not – i.e. by comparing the average score for experts in area C in Figure 5.4 to the average score of experts in all other areas.

Figure 5.4. Skills explored in the first stage of the analysis.



The effect of experts' skills was first compared using:

- Experts' mean KL scores for all domain seeds,
- Experts' mean KL scores for seeds on trial outcomes in control arm only,
- Experts' mean scores for seeds on the treatment effect only,
- Experts' KL scores non-domain seeds.

It is important to note that definitions of categories (e.g. characteristics that define substantive expertise) may affect the findings. Sensitivity analysis was conducted to explore whether changing the definition of substantive expertise and accuracy of probabilistic assessments affected the KL scores. The definitions explored in the sensitivity analysis are shown in Table 5.3.

Table 5.3. Definitions of expertise explored in the sensitivity analysis.

Factors believed to affect experts' priors		Characteristic used to reflect each factor	Sensitivity analysis
Substantive expertise	Experience	Seniority level	Exclude the variable from the definition of substantive expertise
		Research experience (more than three publications or at least one successful research grant application)	Vary the number of publications that define research experience
	Speciality	Time spent with patients	Vary the proportion of time spent with patients for an expert to be considered specialised
		Awareness of research	Exclude the variable from the definition of substantive expertise
Normative expertise		Statistical coherence	None (binary variable)
Ability to make accurate probabilistic assessments		Non-domain seed score	Change the score required for an expert to be considered 'accurate'

Stage 2: The effect of experts' perspective on their scores

Stage 2 of the analysis compared mean scores derived from experts in each individual profession. The results were used to assess whether the effect of experience, normative expertise and accuracy in elicitation were affected by experts' perspective.

As described in Stage 1, the analysis was performed on all domain seeds combined, as well as on outcomes in patients who do not receive the intervention, and the treatment effect separately.

5.3. Results: experts' scores

This section compares experts' scores for each seed. The scores for the probabilities of falling and the rate of falls include 41 experts, while those on the odds and probability of fractures include 37 experts, as the chips and bins grid for the odds of fractures failed to show for 4 experts (as outlined in section 4.3.1 in Chapter 4).

First, section 5.3.1 shows the KL scores, then 5.3.2 shows and analyses the CICP scores.

5.3.1. KL scores for different seeds

KL scores represent the discrepancy between two probability distributions, where the lower the score the lower the discrepancy (i.e. the more accurate the expert) (Kullback and Leibler, 1951).

Experts' KL scores for different seeds are summarised in Table 5.4. All parameters scores in the control arm were higher than the treatment arm, indicating that the experts were less accurate when assessing baseline probabilities (retrospective domain seeds). Out of the three probabilities of falling in the control arm, experts were the most accurate on the probability of falling more than ten times – the mean score for all experts was 0.1995 in comparison to scores 4.681 and 3.673 on the risk of falling and the probability of falling more than five times, respectively. This contrasts with the general literature suggesting experts are less accurate when predicting rare events.

Experts were less accurate (higher KL scores) on the directly elicited parameter (P(x>0, P(x>0|x>5), P(x>5|x>10) and odds of fracture) than they were on the indirectly elicited ones (rate of falls, P(fracture|fall) and the treatment effect) – their mean KL score on the directly and indirectly elicited parameters were 4.078 and 1.963, respectively.

Table 5.4. KL scores for each seed.

Seed		Mean KL score (range)
Non-domain seed	Number of rainy days	1.814 (0.078-3.986)
Seeds regarding outcomes in control arm	P(x>0)	4.681 (1.82-5.391)
	P(x>5|x>0)	3.673 (0.158-5.684)
	P(x>10|x>5)	1.995 (0.233-4.19)
	Rate of falls	2.065 (0.154-2.917)
	Odds of having a fracture after a fall	5.962 (0.807-6.474)
	P(fracture|fall)	6.184 (0.9-6.963)
Seeds regarding outcomes in treatment effect	Risk ratio for P(x>0)	1.995 (0.071-2.913)
	Risk ratio for P(x>5|x>0)	0.895 (0.074-2.916)
	Risk ratio for P(x>10|x>5)	1.059 (0.055-2.524)
	Rate of falls ratio	1.815 (0.089-4.136)
	Odds ratio for having a fracture after a fall	0.702 (0.065-2.878)
	Relative risk for the probability of fractures	0.987 (0.052-2.914)

Priors on the odds of fractures, and the probability of fractures conditional on falling, and odds ratio and relative risk for fractures were derived from the same priors. The scores on these seeds are similar, although not identical (5.962, 6.184 for odds and probabilities and 0.702, and 0.987 for OR and RR), suggesting that the format in which priors are elicited can affect experts' perceived performance. Scores on odds and the odds ratio were better (lower) than on the P(fracture|fall) and the relative risk of fractures (5.962 and 0.702 for odds compared to 6.184 and 0.987 for probabilities).

The odds and probabilities of fractures were the least accurate of all seeds.

Experts' priors on the treatment effect were consistently lower (the priors were more accurate) than on outcomes without treatment, and are comparable to their scores on the number of rainy days in York. There is no apparent difference in the scores when the treatment effect was elicited conditional on the mode outcome without treatment (risk ratios for P(x>5|x>0), P(1x>5|x>0) and P(1x>10|x>5)), and those elicited conditional on the

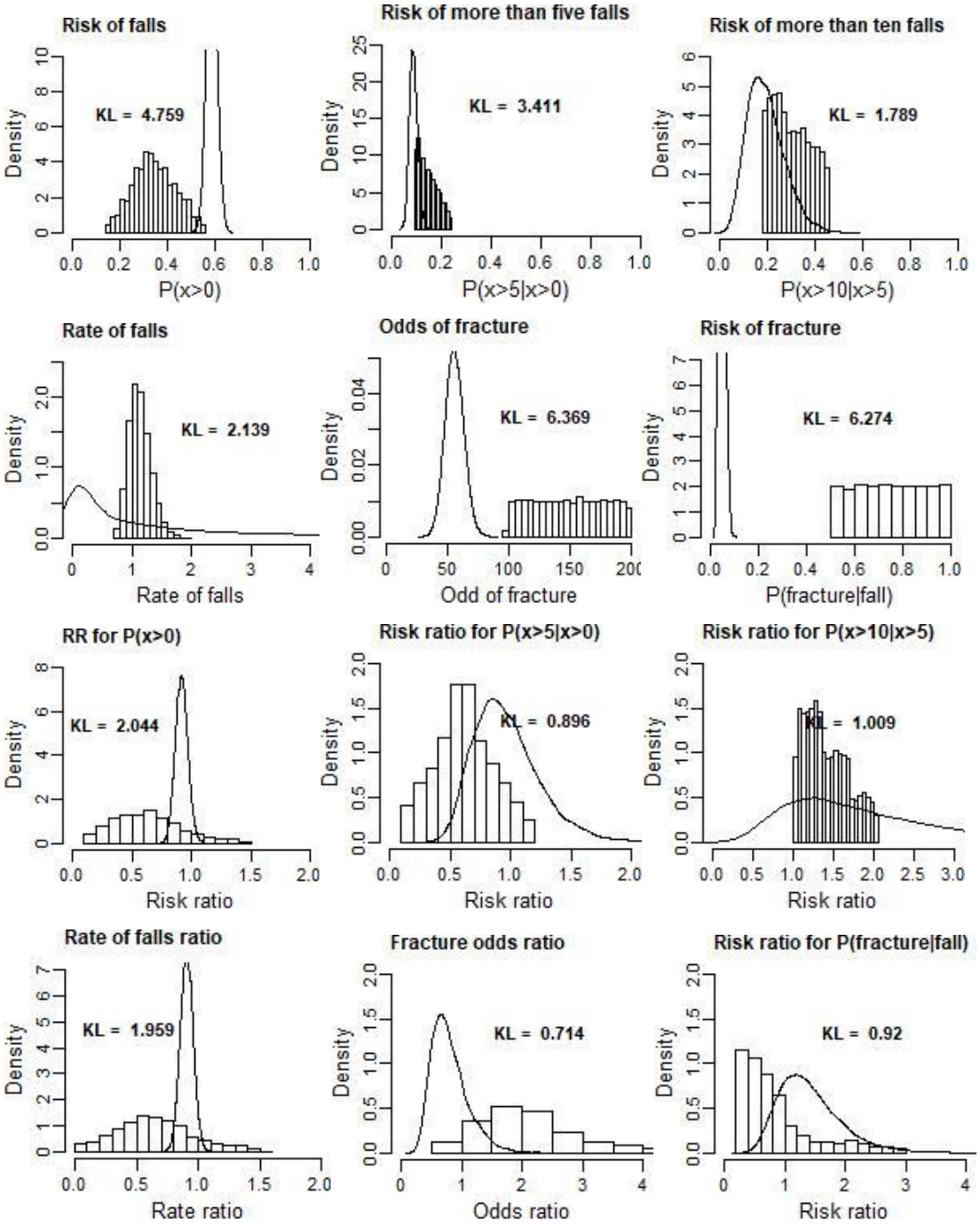
probability distribution of outcomes without treatment (odds ratio and relative risk of fractures).

Section 5.2.1.3 highlighted that the differences in KL scores across different seeds are difficult to interpret if the seeds are not measured on the same scale. In order to gain insight into how much experts' accuracy varied between different seeds, for each domain seed, a prior that scores close to the mean KL score for that seed was identified.

Figure 5.5 shows the selected priors in comparison to the probability distribution of the seed against which they were scored.

The priors on the odds and the risk of fractures - on which the experts attained the highest (worst) scores (5.962 and 6.184, respectively) - did not overlap with the observed probability distribution. The remaining scores appear to be correlated with the uncertainty in the value of the seed. For example, the priors on the risk of more than five and more than ten falls appear to have comparable distance from the observed value of those seeds, however there is less uncertainty around the observed value of the risk of falling more than five times, and so there is a greater disparity between the probability density of experts' prior and the observed value, than there is for the risk of falling more than ten times. A similar pattern follows for all seeds. Experts were more accurate on the treatment effect than on the control arm outcomes, and there is more uncertainty in the former.

Figure 5.5. Examples of priors that achieved mean scores for each domain seed. The histogram represents experts' priors whereas the line represents the probability distribution of the seed value.



5.3.2. CICP scores

Chapter 2 proposed that experts' performance depends on their bias and precision. While KL score captures all three, it does not determine what is driving the score. CICP accuracy and precision scores were thus used to understand what motivates experts' scores, as described in section 5.2.1. The results are shown in Table 5.5.

As discussed in section 5.2.1, KL scores can take values between 0 and infinity, and the lower the KL score the lower the discrepancy between experts' priors and the observed probability distribution. CICP scores can take any value between 0 and 1, and in contrast to KL scores, it is assumed that the higher the CICP precision score, the more precise the expert is, as the score represents the probability that experts' placed on the 95% confidence interval of the observed probability distribution. Similarly, the higher the CICP accuracy score, the less overconfident they are, as the score represents the proportion of the observed probability distribution included in their prior.

CICP precision scores were lower for the baseline outcomes (falls and fractures without treatment) than they were for the treatment effect – the average CICP score for all baseline outcomes was 0.183, compared to 0.500 for the treatment effect. Similar is the case for CICP accuracy scores – experts scored 0.338 on baseline parameters compared to 0.565 on the treatment effect.

The lower CICP accuracy and precision scores suggest that experts were both more biased and overconfident on the parameters regarding outcomes without treatment, compared to those regarding the treatment effect.

It is important to note that CICP and KL scores are not perfectly correlated; for example, CICP scores on P(x>5|x>0) (0.254 and 0.37) were higher than those on P(x>0) (0.026 and 0.214) indicating the experts were less biased and overconfident, yet the mean KL score for on P(x>5|x>0) was worse than the KL score for P(x>0) (KL scores were 5.062 and 4.681, respectively). This can occur if experts were unbiased, and include all observed values in their prior, but their probability is concentrated on a small range of values, leading to a high discrepancy between the priors and the observed probability distribution and consequently a worse KL score.

Table 5.5. Secondary measures of performance for each seed. Rates and proportion of fractures conditional on falling were not assessed for coherence as they were not elicited, but derived from elicited priors. RR=risk ratio; OR=odds ratio.

Seed		KL scores (range)	Mean CICP precision scores (range)	Mean CICP accuracy scores (range)
Non-domain seed	Number of rainy days	1.814 (0.078-3.986)	0.562 (0-1)	0.527(0-0.997)
Domain seeds: falling and fractures without treatment	P(x>0)	4.681 (1.82-5.391)	0.026 (0-0.197)	0.214 (0-1)
	P(x>5|x>0)	3.673 (0.158-5.684)	0.254 (0-1)	0.37 (0-1)
	P(x>10|x>5)	1.995 (0.233-4.19)	0.597 (0-1)	0.531 (0-1)
	Rate of falls	2.065 (0.154-2.917)	0.077 (0-0.526)	0.699 (0-1)
	Odds of having a fracture after a fall	5.962 (0.807-6.474)	0.041 (0-0.59)	0.107 (0-1)
	P(fracture|fall)	6.184 (0.9-6.963)	0.101 (0-1)	0.104 (0-0.97)
Domain seeds: treatment effect on falling and fractures	RR for P(x>0)	1.995 (0.071-2.913)	0.152 (0-0.955)	0.424 (0-1)
	RR for P(x>5|x>0)	0.895 (0.074-2.916)	0.623 (0-1)	0.534 (0-1)
	RR for P(x>10|x>5)	1.059 (0.055-2.524)	0.801 (0.049-1)	0.51 (0-0.98)
	Rate of falls ratio	1.815 (0.089-4.136)	0.146 (0-0.819)	0.746 (0-1)
	OR for having a fracture after a fall	0.702 (0.065-2.878)	0.609 (0-1)	0.523 (0.004-1)
	RR for P(fracture|fall)	0.987 (0.052-2.914)	0.656 (0-1)	0.652 (0-1)

5.4. Results: Experts' characteristics

Experts who took part in the REFORM elicitation study completed a questionnaire on their professional experience, as described in Chapter 3. Their responses were used to define

their substantive expertise and perspective, while their statistical coherence and internal consistency were used to measure their normative expertise.

Sections 5.4.1-5.4.3 describe the results of each.

5.4.1. Experts' normative expertise

Section 5.2.2 outlines that experts' normative expertise was based on the shape of their priors, the probabilities placed outside their stated range, and their internal consistency.

Section 4.5.2 showed that three experts used bimodal priors to express their uncertainty. Of those three, only one expert expressed a single bimodal prior, and they highlighted they were having difficulty adding chips on the tablet they were using. The remaining two experts completed the exercise on a laptop and elicited at least two bimodal priors. Only the latter two were classified as 'not normative'.

Furthermore, section 4.5.2 showed that 20 experts added bricks, that represented probabilities greater than 0.05, to bins outside their stated range. Probabilities outside the stated range is not strictly incoherent, instead it suggests that the expert could have misunderstood the task. Thus not all 20 experts who added more than 0.05 probability to bins outside their range were judged to be not normative, only those who did it repeatedly. The cut off was set at more than 3 priors.

The expert who only used one brick in each histogram, and expressed that the number of rainy days in York was between 30-31 was also classified as non-normative.

Table 4.18 in Chapter 4 reported experts' internal consistency, while Table 5.6 summarises experts' normative expertise.

Table 5.6. Summary of experts' statistical coherence and internal consistency.

Experts with less normative expertise				Normative	Total
Incomplete priors	Bimodal	Probability outside stated range	Inconsistent		
2	2	9	17	17	41

In total, 17 out of 41 experts elicited priors that were inconsistent with their MCQs. Of those, six experts also elicited priors that were statistically incoherent. Furthermore seven experts elicited priors that were statistically incoherent but internally consistent. Seventeen

experts provided consistent and coherent priors and so were considered to be normative experts.

5.4.2. Experts' substantive expertise

The questions about experts' professional experience aimed to capture how specialised experts are in fall prevention (or treatment), their level of seniority based on their role, research experience, patient contact (proportion of their working time spent with patients who are at an increased risk of falling, either helping them prevent falls or treating fall related injuries), and awareness of research into podiatry interventions designed to reduce the risk of falls in the elderly. The details of how information was collected on each of the four characteristics is provided in section 5.2.2.

As highlighted in section 5.2.3, the sample size was too small to explore individually the effect of every stated characteristic on experts' scores. Instead, the information on experts' experience was used to classify them as substantive or not substantive. The aim was to have a minimum of 6 experts who were substantive, normative and accurate when making probabilistic assessments to represent a typical sample of experts in elicitation. The definition of 'substantive' experts was thus based on the characteristics of the sample.

First, research experience and time spent with patients were simplified into binary variables: research experience or no research experience, and patient contact or no patient contact. The decision on how to define research and patient experience was based on the sample size, as follows. A summary of experts' characteristics is provided in Table 5.7.

Table 5.7 shows that only one expert had published over 50 research papers, and so experts with 20-50 and more than 50 publications were combined into a single category and four different definitions of research experience were considered: 1) More than 3 publications or at least one successful research grant proposal; 2) More than 20 publications or at least one successful research grant proposal; 3) More than 3 publications and at least one successful research grant proposal; and 4) More than 20 publications and at least one successful research grant proposal. Table 5.8 shows the number of experts in the REFORM elicitation study with different levels of experience, per role. Only two experts in the sample had more than 20 publications and experience in writing at least one research grant proposal (research experience 4 in Table 5.8) and so this definition of research experience was not chosen as the baseline scenario in the analysis.

Table 5.7. Summary of experts' characteristics.

		Level 1 role	Level 2 role	Total
Number of publications	0-3	6	24	30
	4-20	2	7	9
	21-50	0	1	1
	>50	0	1	1
Number of protocols	0	7	24	31
	1-5	1	9	10
	>5	0	0	0
Patient contact	0-10%	0	4	4
	11-30%	2	5	7
	31-50%	0	4	4
	>50%	6	20	26
Aware of research		1	19	20
Total		8	33	41

Table 5.8. The number of experts with different levels of patient contact and research experience.

	Research Experience				Percentage of working time spent with target patient population			Aware of domain specific research	Total
	1*	2*	3*	4*	>50%	>30%	>10%		
Level 1	2	1	0	0	6	6	8	1	8
Level 2	10	9	8	2	20	24	29	19	33

***Research experience 1 = More than 3 publications or at least one successful research grant proposal; research experience 2 = More than 20 publications or at least one successful research grant proposal; research experience 3 = More than 3 publications and at least one successful research grant proposal; and research experience 4 = More than 20 publications and at least one successful research grant proposal.**

Further decisions on how to define substantive experts were based on the combinations of characteristics of experts in the sample. Table 5.9 shows the sample size for experts who were in Level 2 roles and were aware of relevant research with different levels of research experience and patient contact. When research awareness and a Level 2 role were used to define substantive expertise, the sample size of experts who were substantive, normative and made accurate probabilistic assessments was never greater than 5. Removal of research awareness from the definition of substantive expertise was considered in order to obtain a greater sample size.

Table 5.9. Number of experts who were in Level 2 roles and were aware of research into podiatry interventions designed to reduce the risk of falls. *Definitions of different research experience are shown in Table 5.8.

Research experience	Patient contact	Total	Number of experts who were normative	Number of experts who were normative and accurate
Any	Any	19	8	5
	>10%	17	7	5
	>30%	14	6	4
	>50%	12	4	3
1*	Any	10	4	3
	>10%	9	4	3
	>30%	6	3	2
	>50%	4	1	1
2*	Any	9	4	3
	>10%	8	4	3
	>30%	6	3	2
	>50%	4	1	1
3*	Any	8	4	3
	>10%	7	4	3
	>30%	5	3	2
	>50%	3	1	1
4*	Any	2	1	0
	>10%	2	1	0
	>30%	2	1	0
	>50%	1	0	0

The sample size of experts who were in Level 2 roles with different levels of research experience and patient contact are shown in Table 5.10. The sample size of experts with research experience was the same with and without research awareness indicating that

everyone with any research experience was also aware of literature on podiatry interventions designed to reduce the risk of falls. As a result, only excluding research experience from the definition of substantive expertise led to a sample size greater than 5.

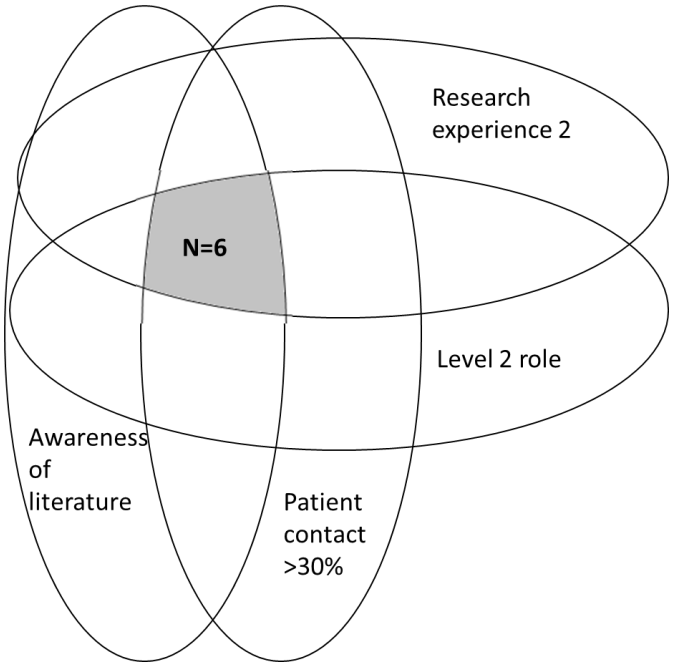
Table 5.10. Number of experts who were in Level 2 roles. *Definitions of different research experience are shown in Table 5.8.

Research experience	Patient contact	Total	Number of experts who were normative	Number of experts who were normative and accurate
All	Any	33	13	7
	>10%	29	11	6
	>30%	24	10	5
	>50%	20	8	4
1*	Any	10	4	3
	>10%	9	4	3
	>30%	6	3	2
	>50%	4	1	1
2*	Any	9	4	3
	>10%	8	4	3
	>30%	6	3	2
	>50%	4	1	1
3*	Any	8	4	3
	>10%	7	4	3
	>30%	5	3	2
	>50%	3	1	1
4*	Any	2	1	0
	>10%	2	1	0
	>30%	2	1	0
	>50%	1	0	0

Level 2 role and more than 10% patient contact (the only experience that led to a sample size greater than 6) were assumed to be insufficient to capture substantive expertise, and so a smaller sample of experts was used in the analysis. Substantive expertise was defined in a way that maximised their expertise, while ensuring a sample size of six or more substantive experts.

Research experience 2 and patient contact more than 30% were chosen as the definition of substantive expertise, in addition to Level 2 role and research awareness (see Figure 5.6 for diagrammatic representation). Increasing patient contact or research experience both led to a sample size below 6.

Figure 5.6. Venn diagram showing the definition of substantive expertise in the baseline scenario.

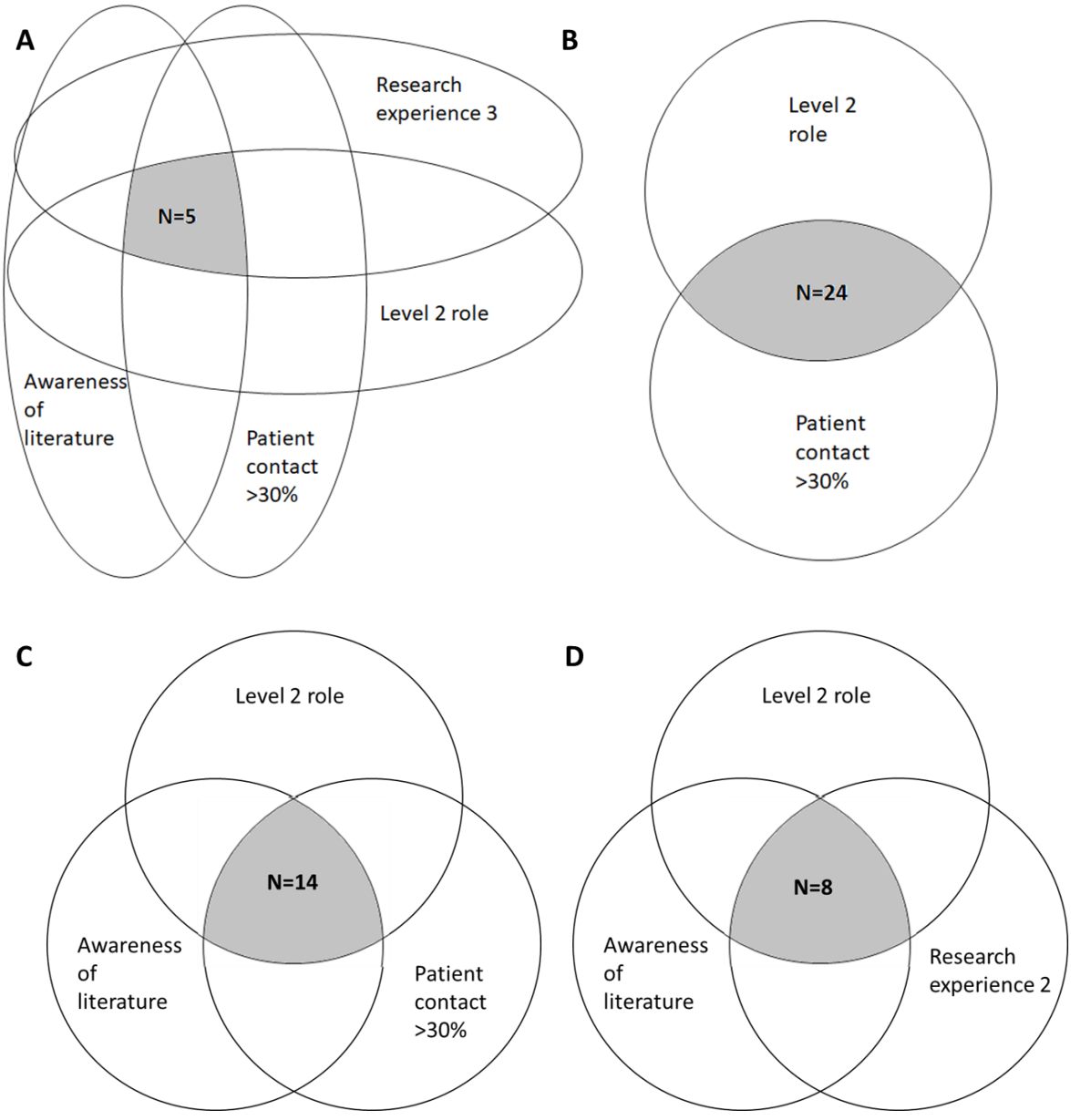


Other definitions of substantive expertise were explored in sensitivity analysis, where one characteristic was removed or altered at a time and the effect on experts' scores was observed. The definitions explored in the sensitivity analysis are shown in Figure 5.7 and included:

- A different definition of research experience (more than 3 publications and at least one successful research grant proposal, compared to only one of the two conditions used in the baseline scenario) to assess whether more research experience affected experts' priors (Figure 5.7.A),
- Excluded awareness of podiatry interventions and research experience from the definition to achieve a greater sample size for experts who were substantive, normative and made accurate probabilistic assessments (Figure 5.7.B),
- Excluded research experience from the definition (Figure 5.7.C),
- Excluded patient contact from the definition (Figure 5.7.D).

The effect of the seniority level was not explored in the definition of substantive expertise as it was considered to be fundamental to the definition of substantive expertise.

Figure 5.7. Venn diagrams showing definitions of substantive expertise explored in the sensitivity analysis. The area highlighted in grey represents characteristics of substantive experts. Research experience is defined in Table 4.22. Definitions A, C and D represent rows 15, 3 and 13 in Table 5.9 respectively, while definition B represents row 3 in Table 5.10.



5.4.3. Experts' perspective

Section 5.2.2 explained that experts' perspective was defined by their profession. Table 4.2 in section 4.3.2 showed the number of experts recruited in each role. In this section, Table 5.11 shows the characteristics of experts in each profession.

Table 5.11. Professional experience by role.

		Proportion of sample in each type of role		Years in role	Percentage of working time spent with patients**				Proportion of experts with research experience				Research awareness	Proportion of experts who were classified as experts		N
		Level 1	Level 2		0-10	11-30	31-50	>50	1*	2*	3*	4*		Subst.	Subst. + Norm. + Accurate	
Physios		0.15	0.85	8.2	0	0.08	0.15	0.77	0	0	0	0	0.38	0	0	13
Geriatricians		0.33	0.67	6.64	0	0.2	0.13	0.67	0.53	0.47	0.47	0.13	0.47	0.27	0.13	15
Nurses		0.14	0.86	10	0	0.29	0	0.71	0.14	0	0	0	0.57	0	0	7
GPs		0	1	2.5	0.5	0.25	0	0.25	0.5	0.5	0.25	0.25	0.5	0.25	0	4
Academics		0	1	5	1	0	0	0	1	1	1	0	1	0	0	1
Other		0	1	NA	1	0	0	0	0	0	0	0	1	0	0	1

*Definitions of different research experience are shown in Table 5.8.

** Either helping them prevent falls or treating fall related injuries.

The majority of experts from all professions were in Level 2 roles. Geriatricians had the highest proportion of experts in Level 1 roles, this is likely to be because regional meetings, where these experts were recruited, were organised for trainee geriatricians. Senior staff attending (consultants) were teaching or presenting.

Overall physiotherapists and nurses had high patient contact and little research experience. 77% of physios and 71% of nurses spend more than half of their time with such patients, and they had the longest experience in the role (8.2 and 10 years respectively). None of the physiotherapists had any research experience and 14% of nurses had 'Research experience 1'.

Geriatricians had similar contact time with the target patient population (67%), and they were the group most involved in research - more than half (53%) had published at least one paper or been involved in writing successful research grant proposals, and 47% had been involved in both. This may be due to the fact that all senior geriatricians involved were recruited either through their involvement in research on fall prevention, or through training days. It is possible that consultants who attend training days are more likely to be involved in teaching and research.

Geriatricians and GPs were the only ones who were considered substantive in the baseline scenario, and of those only geriatricians were also normative and substantive.

5.5. Effect of experts' characteristics on their priors

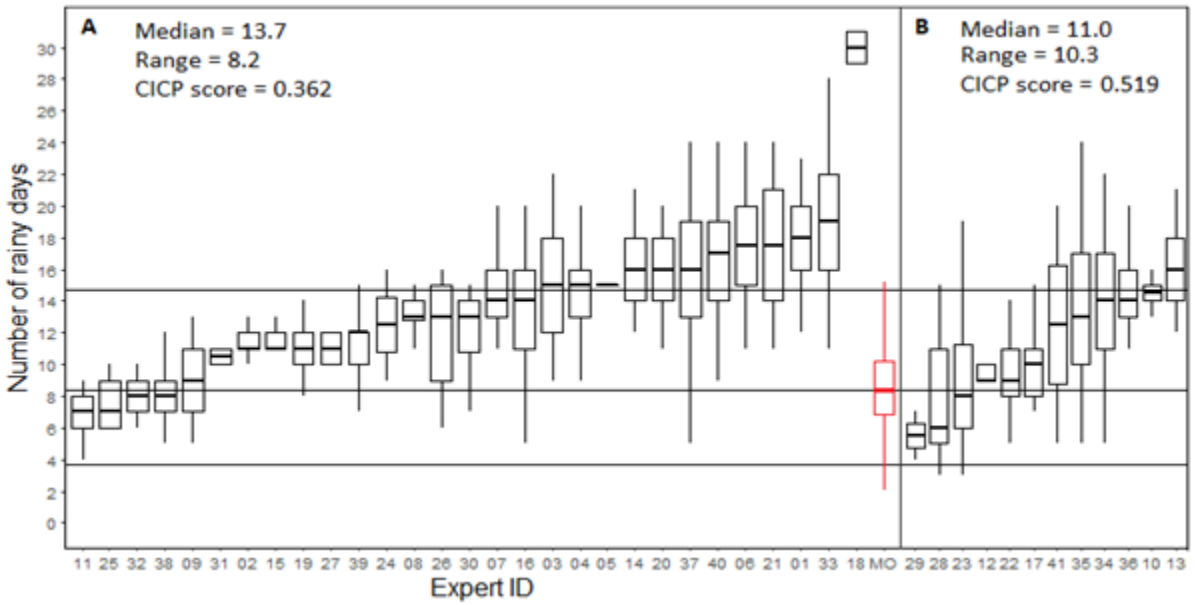
This section reports the effect of experts' characteristics on their priors on non-domain and domain seeds, in turn.

5.5.1. Exploring the effect of experts' characteristics on the non-domain seed

This section explores whether experts' Grace-Watson test scores, statistical coherence and internal consistency explain their priors on the non-domain seed.

Experts' priors and scores on the non-domain seed are shown in Figure 5.8. Experts who live in the Y&H region scored higher overall – Y&H experts were more likely to include the observed value and cover the range, although with the current sample size the difference is not statistically significant.

Figure 5.8. Priors on the number of rainy days elicited from A) experts who were recruited outside Y&H; B) experts who were recruited in Y&H. The median, range and CICP score represent the average (mean). Box = the median and the interquartile range. Whiskers = 95% confidence interval. The red box (MO) = reference value from the Met Office.



Experts' scores according to different characteristics are shown in Table 5.12. Priors elicited from experts recruited outside Yorkshire and Humber were less accurate (scored lower) than those were recruited within Yorkshire and Humber, suggesting that a lack of substantive expertise can lead to overconfidence. However, substantive expertise alone does not ensure accurate priors (low scores) - experts who were substantive but scored 0-2 on the GW critical thinking score and were not normative were most likely to be overconfident of all experts.

When GW scores were not taken into account, normative expertise improved scores both for substantive and non-substantive experts.

When normative expertise was not taken into account, higher GW scores improved the CICP score in Y&H experts, whereas they reduced it in non-substantive experts. This finding was counterintuitive, as inference skills were expected to be beneficial to the non-substantive experts who were required to extrapolate their knowledge about their local weather to assess the number of rainy days in York, compared to the substantive experts who were only required to express their beliefs.

When both normative expertise and GW scores were taken into account, no pattern in beliefs was identified.

Table 5.12. Experts' scores according to different characteristics.

Experts' characteristics		Y&H		Not Y&H		All	
		Score (N)	P-value	Score (N)	P-value	Score (N)	P-value
Normative	+	0.329 (7)	0.416	0.498 (10)	0.713	0.428 (17)	0.473
	-	0.408 (5)		0.530 (19)		0.505 (24)	
GW test score 3-4	+	0.265 (3)	0.578	0.634 (9)	0.194	0.542 (12)	0.3892
	-	0.395 (9)		0.467 (20)		0.445 (29)	
Normative and GW score	Normative, GW score 3-4	0.393 (2)	0.578	0.506 (5)	0.194	0.473 (7)	0.369
	Not normative, GW score 3-4	0.008 (1)		0.795 (4)		0.637 (5)	
	Normative, GW score 0-2	0.303 (5)		0.490 (5)		0.397 (10)	
	Not normative, GW score 0-2	0.508 (7)		0.459 (15)		0.440 (19)	
All		0.362		0.519		0.362	

None of the effects were statistically significant. This is likely to be due to the small sample size.

5.5.2. Exploring whether experts' characteristics explain their priors on substantive seed parameters

The analysis of the effect of experts' characteristics on their priors was conducted in two stages, as described in section 5.2.3: stage 1 explored the effect of substantive and normative expertise, and their ability to make accurate probabilistic assessments on experts' elicitation scores, while stage 2 explored the effect of experts' perspective. The results for each stage are presented here, in turn.

Stage 1: The effect of substantive and normative expertise, and accuracy of probabilistic assessments

Experts' priors on seed parameters were scored using KL scores; the scores were then compared for those experts who were substantive and those who were not, and between experts who were

substantive, normative and adaptive and those who were not. The results of these comparisons are shown in Table 5.13.

Table 5.13. Mean KL scores on different seeds for experts and non-experts.

Seeds included		Substantive only		Substantive, normative and accurate	
		Experts (N=6)	Non-experts (N=35)	Experts (N=2)	Non-experts (N=39)
All domain seeds	Score	2.432 (0.078-6.916)	2.717 (0.052-6.963)	2.469 (0.166-6.879)	2.686 (0.052-6.963)
	P-value	0.319		0.512	
Seeds outcomes without treatment	Score	3.795 (0.315-6.916)	4.144 (0.154-6.963)	4.443 (0.791-6.879)	4.075 (0.154-6.963)
	P-value	0.459343		0.549	
Seeds about treatment effect	Score	1.068 (0.078-2.923)	1.276 (0.052-4.136)	0.496 (0.166-1.012)	1.284 (0.052-4.136)
	P-value	0.376		0.029	
Non-domain seed	Score	2.062 (0.336-3.223)	1.772 (0.078-3.986)	0.81 (0.336-1.285)	1.866 (0.078-3.986)
	P-value	0.580		0.215	

Priors elicited from substantive experts were more accurate (achieved lower KL scores) than those elicited from non-substantive experts on all domain seeds, while the opposite was the case for the non-domain seed where substantive experts achieved a mean score of 2.062, whereas the non-experts achieved 1.772.

The effect of normative expertise and accuracy of probabilistic assessments is less clear. When non-normative and inaccurate experts were excluded from the sample, scores for priors on the treatment effect and the non-domain seed decreased (from 1.068 to 0.496, and from 2.062 to 0.81, respectively), suggesting that normative expertise and the ability to make accurate probabilistic assessments improved experts' scores both on domain and non-domain seeds. The difference between the experts' and non-experts' scores for the treatment effect was the only statistically significant difference (P-value 0.029).

However, exclusion of non-normative and inaccurate experts had the opposite effect on the scores for outcomes in the control arm – these increased from 3.795 to 4.443, and experts scored worse (higher) than non-experts (4.443 compared to 4.075).

Sensitivity analysis was used to explore the effect of changing the definition of substantive expertise on experts' scores. It also explored the effect of normative expertise and the ability to make accurate probabilistic assessments individually to understand which of the two had a greater effect on experts' scores.

The results of the sensitivity analysis are shown in Table 5.14. Substantive experts were more accurate than non-substantive experts in all scenarios. In scenario A, when the definition of expertise was changed to include more than three publications and at least one successful research grant proposal (as opposed to more than 20 publications or at least one successful research grant proposal) experts' priors were on average more accurate than in the baseline scenario (score decreased from 2.432 to 2.338). In scenario C, when research experience was not included in the definition of substantive expertise, the scores of substantive expert had increased from 2.432 to 2.64 suggesting the priors were less accurate than when experts with no research experience were added to the sample.

In scenario D, when patient contact was excluded from the definition of substantive expertise, the scores were higher than in the baseline scenario (from 2.432 to 2.595), suggesting that experts who had more patient contact were more accurate. Excluding patient contact from the definition of substantive expertise led to better (lower) scores than when research experience was excluded (2.595 compared to 2.64), suggesting that experts with research experience were more accurate than experts with patient contact.

Excluding research awareness from the definition of substantive expertise (Scenario B) slightly improved scores in comparison to scenario D (2.595 to 2.522) where research awareness was included. However, the scores in scenario B were worse than in the baseline scenario, suggesting that research experience improved experts' scores more than research awareness.

Scenario A has the highest difference in scores between substantive experts and non-substantive experts (mean KL scores 2.338 and 2.722, respectively), while scenario B was the only scenario where the comparison of the accuracy of substantive experts to non-experts was statistically significant (P-value 0.019). The statistical significance in scenario B (and a lack thereof in other scenarios) is likely to be due to the higher sample size – in scenario B there were 20 experts, compared to 2-12 experts in all other scenarios.

Table 5.14. Mean KL scores on all domain seeds for experts and non-experts, when different definitions of expertise were used.

Seeds included		Substantive only		Substantive and normative		Substantive and accurate		Substantive, normative and accurate	
		Experts	Non-experts	Experts	Non-experts	Experts	Non-experts	Experts	Non-experts
X	Score	2.432 (0.078-6.916) N=6	2.717 (0.052-6.963) N=35	2.593 (0.166-6.916) N=3	2.682 (0.052-6.963) N=38	2.457 (0.166-6.879) N=3	2.692 (0.052-6.963) N=38	2.469 (0.166-6.879) N=2	2.686 (0.052-6.963) N=39
	P-value	0.319		0.759		0.320		0.512	
A	Score	2.338 (0.078-6.916) N=5	2.722 (0.052-6.963) N=36	2.593 (0.166-6.916) N=3	2.682 (0.052-6.963) N=38	2.457 (0.166-6.879) N=3	2.692 (0.052-6.963) N=38	2.469 (0.166-6.879) N=2	2.686 (0.052-6.963) N=39
	P-value	0.175		0.759		0.320		0.512	
B	Score	2.522 (0.052-6.963) N=20	2.894 (0.065-6.956) N=21	2.702 (0.065-6.933) N=8	2.666 (0.052-6.963) N=33	2.548 (0.052-6.963) N=9	2.728 (0.055-6.956) N=32	2.791 (0.114-6.933) N=4	2.659 (0.052-6.963) N=37
	P-value	0.019		0.800		0.088		0.862	
C	Score	2.64 (0.055-6.963) N=12	2.693 (0.052-6.956) N=28	2.876 (0.114-6.916) N=4	2.64 (0.052-6.963) N=37	2.752 (0.071-6.963) N=6	2.659 (0.052-6.956) N=35	2.946 (0.114-6.879) N=3	2.646 (0.052-6.963) N=38
	P-value	0.734		0.438		0.959		0.457	
D	Score	2.595 (0.078-6.956) N=8	2.695 (0.052-6.963) N=33	2.632 (0.166-6.916) N=4	2.680 (0.052-6.963) N=37	2.530 (0.166-6.879) N=4	2.691 (0.052-6.963) N=37	2.562 (0.166-6.879) N=3	2.684 (0.052-6.963) N=38
	P-value	0.784		0.882		0.457		0.687	

X = baseline scenario (Level 2 role, research awareness, patient contact >30%, research experience 2); A = Level 2 role, research awareness, patient contact >30%, research experience 3; B= Level 2 role, patient contact >30%; C= Level 2 role, research awareness, patient contact >30%; D= Level 2 role, research awareness, research experience 2. The definitions are also presented diagrammatically in Figure 5.7 (section 5.4.2).

When experts were defined by their substantive and normative expertise and the accuracy of their probabilistic assessments, the results varied. The sample of substantive experts in the baseline scenario (scenario X in Table 5.14) and scenario A differed in one expert who was substantive in scenario X but not in scenario A. This expert was neither normative nor accurate, so the sample of substantive, normative and adaptive experts in the two scenarios is identical. In these scenarios both normative expertise and the ability to make accurate probabilistic assessments led to better (lower) scores in experts than non-experts (KL scores 2.593, 2.457 and 2.469 in experts, and 2.682, 2.692 and 2.686 in non-experts). Substantive and accurate experts attained better (lower) scores than experts who were substantive, normative and accurate (KL scores of 2.457 and 2.469). Furthermore, priors elicited from all substantive experts regardless of their normative expertise and ability to make accurate probabilistic assessments attained better (lower) scores (KL=2.432) than those elicited from experts who satisfied all three criteria (KL=2.469). The same pattern follows in scenario D – including normative and accurate experts led to better (lower) scores in comparison to non-experts, but the sample that included all substantive experts led to the best (lowest) scores.

In scenarios B and C– when research experience was not included in the definition of substantive expertise - all three (normative expertise only, accurate only and normative and accurate) led to higher (worse) scores in experts than non-experts. Furthermore, excluding non-normative and inaccurate experts from the sample worsened (increased) the scores.

Stage 2: The effect of experts' perspective on their scores

As shown in Table 5.11 geriatricians and GPs were the only experts who satisfied the criteria of substantive expertise, and of those, only geriatricians were also normative. Priors elicited from experts with different professions were compared to assess whether different professions led to different levels of accuracy and patterns of belief. The results are shown in Table 5.15.

Physiotherapists were the most accurate on the treatment effect and on the control (mean KL scores 3.864 and 1.07, respectively), followed by geriatricians (KL 1.115 and 4.043, respectively). The results should be interpreted with caution as different professions had different levels of experience. As shown Table 5.11 in section 5.4.3, geriatricians had the lowest proportion of Level 2 expert, so it is not clear whether their scores were lower because of their substantive expertise or their role.

Table 5.15. Scores and patterns of belief elicited from experts in different professions.

	Physios N=13	Geriatrics N=15	Nurses and OTs N=7	GPs N=4	Academic N=1	Other N=1
KL scores on all domain seeds	2.467 (0.055-6.963)	2.579 (0.052-6.956)	3.1 (0.114-6.949)	2.712 (0.151-6.855)	2.883 (0.197-6.561)	3.658 (0.154-6.834)
KL scores on baseline falls and fractures	3.864 (0.198-6.963)	4.043 (0.155-6.956)	4.502 (0.154-6.949)	4.056 (0.32-6.855)	4.585 (0.901-6.561)	4.618 (0.154-6.834)
KL scores on treatment effect	1.07 (0.055-3.72)	1.115 (0.052-4.136)	1.697 (0.114-3.921)	1.368 (0.151-2.913)	1.18 (0.197-2.913)	2.217 (0.84-2.913)

5.6. Summary of findings

This chapter explored whether variation in experts' priors can be explained by their characteristics. The chapter had three distinct objectives:

1) To score experts' priors on different seeds;
2) To define characteristics that can be used to reflect the factors believed to affect experts' priors in the REFORM elicitation study: substantive expertise, perspective, normative expertise, and the ability to make accurate probabilistic assessments.
3) To explore whether the identified measures explain variation in priors elicited from different experts.

The findings of each are summarised here in turn.

5.6.1. Scoring experts' performance

In this Chapter, the methods used to score experts priors were chosen based on the guiding principles developed in Chapter 2 – KL scores were chosen to capture experts' bias, overconfidence and imprecision.

The priors were scored with reference to probability distributions of the seeds rather than point estimates, to take into account uncertainty around their true value – a method commonly used in CEDM (Bojke *et al.*, 2010) but not in other sectors. (Colson and Cooke, 2017)

This is the first identified study where KL scores were used to score experts' priors on seeds associated with uncertain values, in order to compare their accuracy, and so it was not clear how best to implement the method. KL scores are derived by comparing intervals of the continuous scale of parameters values. Since probability distributions were not fitted to experts' priors (as discussed in Chapter 3), the width of the intervals compared could affect experts' scores – narrow intervals could penalise experts' in a uniform distribution across the interval is assumed, whereas wide intervals could fail to distinguish between experts with marginal differences in performance. The width of bin can affect the score but it is not clear whether the effect is comparable between experts or some are affected more than others.

Furthermore, not fitting could fail to penalise bias if both experts place the same probability on the observed values of the seed, but one is more biased than the other.

Overall experts' scored lower (better) scores on the indirectly elicited priors (such as the treatment effect and the rate of falls), compared to those that were elicited directly (the probabilities of falling without treatment) and on the treatment effect compared to the baseline probabilities of falling and fractures.

However, when priors on different seeds are compared the difference is difficult to interpret, as the scores were dependent on the certainty with which the seed is known. Priors on seeds that were known with more certainty (for example, the risk of falling) attained relatively worse (higher) scores than those that were uncertain, and so it is not clear whether differences in performance across different parameters are driven by experts' performance or the sample size used to measure the value of the seed (where larger sample size, and so less uncertainty, can lead to worse (higher) scores).

5.6.2. Can experts' characteristics be used to reflect the factors believed to affect experts' priors?

The literature review in Chapter 2 found that experts' characteristics have only been used to capture experts' substantive expertise. (Goossens, 2008b; Shabaruddin *et al.*, 2010; Haakma *et al.*, 2014) This chapter aimed to define characteristics that can be used as proxies for all four factors believed to affect experts' priors (namely substantive and normative expertise, and perspective), in the REFORM elicitation study.

Many characteristics that could potentially reflect substantive expertise were identified, (Bolger, 2017), the characteristics used in this study were role, research experience, patient contact, and research awareness, and various combinations of these.

Statistical coherence and internal consistency were used to reflect experts' normative expertise. This chapter identified two features of experts' priors that could indicate statistical incoherence: bimodal priors and probabilities placed outside the experts stated range. Additional features were identified that could indicate difficulty in expressing one's beliefs using the chips and bins method, these included implausibly narrow ranges and a mode at the end of the distribution. These were not chosen as indicators of normative expertise because of the difficulty in determining what is plausible, i.e. it requires subjective judgement.

Perspective was based on experts' role. Experts' perspective is likely to have varied within each profession as well (for example, physiotherapists could be based in primary and secondary care), but in the REFORM elicitation study the experts did not provide sufficient information to further clarify their roles.

The literature is not clear on how to measure the ability to assess probabilistic assessments. The GW test score is used here to capture experts' ability to draw inference, but this did not affect their non-domain seed so was not used in further analysis.

5.6.3. Do the identified characteristics explain variation in priors elicited from different experts?

The effect of experts' characteristics on non-domain and domain seeds were explored separately.

The non-domain seed was the number of rainy days every September in York. Priors elicited from experts who were recruited in the Yorkshire and Humber region attained better scores than those were recruited in other regions of the UK suggesting that the scores were affected by substantive expertise/perspective. Normative expertise was correlated with better performance.

The domain seeds were the REFORM trial outcomes. Experts' scores varied across seeds more than they did between experts but the effect of expertise was relatively constant across parameters.

Substantive expertise was found to improve scores on domain seeds, but not the non-domain seed, suggesting that substantive expertise does improve scores. The effect was consistent when different definitions of substantive expertise were used, although it was not statistically significant, likely because of the sample size.

The effect of normative expertise and accuracy of probabilistic assessments on the domain seeds is less certain.

The implications of the findings for the role of different weighting methods is uncertain. This chapter only explored accuracy of individual experts. It is not clear whether including only more accurate experts improves the accuracy of the aggregate prior, or whether the 'Wisdom of Crowds' outweighs any benefit incurred by only including accurate experts. Furthermore, the impact of the improvement in prior accuracy on the results of cost-effectiveness analysis is not clear. These themes are explores in Chapter 6.

Chapter 6. Comparison of different weighting methods: results and impact of the REFORM elicitation study

6.1. Introduction

Chapter 2 identified two approaches for deriving weights- based on experts' observed characteristics and their measured performance in elicitation – and concluded that it is not clear which approach leads to more accurate aggregate priors. Chapter 3 designed the REFORM elicitation study that compared different weighting methods in an elicitation exercise applied in HTA. Chapter 5 used the results of the elicitation study to explore whether experts' characteristics explained their priors and predicted their accuracy. The findings suggested that the accuracy of priors elicited from individual experts was affected by their substantive and normative expertise, as well as their ability to make accurate probabilistic assessments outside their domain of expertise. However, the discussion highlighted that eliciting from the most accurate experts only may not lead to the most accurate aggregate prior.

This chapter compares the effect of using different weighting methods applied to the REFORM elicitation study. The results from the study were used to explore the following three objectives.

1) To apply different weighting methods identified in Chapter 2 to the REFORM elicitation study.
2) To compare the effect of different weighting methods on the accuracy of the aggregate prior.
3) To observe the impact of different weighting methods on estimates of uncertainty in cost-effectiveness analysis.

Section 6.2 describes the analysis methods, while sections 6.3 - 6.5 present the results. Section 6.6 then provides a summary of the findings.

6.2. Methods

This section describes the methods used to observe the effect of different weighting methods in elicitation. Section 6.2.1 describes how weights were derived in the REFORM elicitation study (objective 1). Section 6.2.2 describes the methods used to compare the effect of the different weighting methods on the accuracy of the aggregate prior (objective 2), while section 6.2.3 describes the methods used to observe their effect on the results of cost-effectiveness analysis (objective 3).

6.2.1. Deriving weights from the results of the REFORM elicitation study

This chapter compared eight different weighting methods in total, where six were based on experts' characteristics and two were based on their elicitation performance, to unweighted priors.

Characteristics used to derive weights were based on the information collected in the REFORM elicitation study. Chapter 5 described the characteristics used to reflect three of the four factors proposed to affect elicitation performance in Chapter 2: substantive expertise, perspective and normative expertise. In Chapter 6, six different sets of weights were derived using different combinations of characteristics.

- Two methods for deriving weights from experts' substantive expertise,
- Perspective,
- Substantive expertise and perspective,
- Normative expertise,
- Substantive expertise, normative expertise and perspective.

Chapter 2 concluded that external validity of performance-weighted priors can depend on the seed, the scoring method, and the method for deriving weights. Seeds tend to be domain-specific, and so experts' performance on domain-seeds was used as one of the methods for deriving weights in this chapter. Furthermore, Chapter 2 proposed that, in theory, non-domain seeds can be used to capture experts' normative expertise and ability to make accurate probabilistic assessments, independent of their substantive expertise, and so the performance on non-domain seeds was another method for deriving weights.

The remainder of this section describes each of the ten weighting methods in detail.

6.2.1.1. Methods 1 and 2: Deriving weights from experts' substantive expertise

The accuracy of priors, weighted by experts' substantive experience only, was explored in this chapter because Chapter 2 found substantive expertise to be the most common characteristic chosen as the basis for weighting.

Chapter 5 (section 5.5.2) explored four different characteristics that could be proxies for substantive expertise: research experience, seniority, patient contact, and awareness of research in podiatry interventions designed to reduce the risk of falls. The former three of the four characteristics were found to improve experts' elicitation performance; these were thus used as basis for differential weighting.

Chapter 2 (section 2.4.2) highlighted that the relationship between experience and elicitation performance is not clear, and that weights derived from experience tend to be assigned arbitrarily. Given the lack of guidance on how to derive weights from characteristics, in this study one point was added for each of the following three characteristics:

- Level 2 role (specialist clinical roles such as consultant geriatricians, and other healthcare professionals specialising in fall prevention);
- More than 50% of working time with the relevant patient population;
- More than five publications or at least one successful research grant proposal.

In addition, experts who had both more than three publications and had written successful research grant proposals were assigned an extra point, as having both was found to improve their scores in Chapter 5 (see section 5.5.2 for details).

Experts could thus score 0-4 points. The number of experts with each number of points, and the weights assigned to those experts are shown in Table 6.1. The table shows weights derived using two different methods for converting scores into weights.

Table 6.1. Experts' scores based on substantive expertise and resulting weights.

Score	0	1	2	3	4
N	2	10	15	4	4
Weight per expert (method 1)	0	0.0147	0.0294	0.0441	0.0588
Weight per expert (method 2)	0	0.01	0.0133	0.075	0.1

Method 1 was employed by Haakma et al. (2014) and uses Equation 6.2.

$$w_i = \frac{score_i}{\sum_{i=1}^{N} score_i} \qquad \textbf{Equation 6.2}$$

Where i is one of N experts (N=35),

w_i is the weight of expert i,

$score_i$ is the number of points assigned to expert i.

It is important to note that Method 1 assigns greater weights to experts with 2 points, than those with 3 or 4 points because their sample size is greater (15 *x* 0.0294 > 4 *x* 0.0441 and 15 *x* 0.0244 > 4 *x* 0.0588). Method 2 was derived to ensure overall weight assigned to experts with the same number of points was proportional to the number of points, as shown in Equation 6.3.

$$w_j = \frac{j}{\sum_{j=0}^{4} j} \qquad \textbf{Equation 6.3}$$

Where j represents the number of points and ranges 0-4,

w_j is the weight assigned to all experts with j points.

All experts with j points were then assigned equal weights that summed to w_j.

6.2.1.2. Method 3: Deriving weights from experts' perspective

As highlighted in Chapter 1, the elicitation literature recommends using a sample of experts that represent different perspectives in the domain in which elicitation is conducted, and so experts' were assigned weights that ensured that the aggregate prior reflected the beliefs of a heterogeneous sample of experts.

Perspective was defined by experts' role: physiotherapists, geriatricians, nurses and occupational therapists (OTs), GPs and academics, and all professions were assigned equal weights (1/5) so they contribute equally towards the aggregate priors. When there was more than one expert with the same perspective, all experts with that perspective were assigned equal weights that summed to 0.2 (1/5). The resulting weights assigned to

individual experts within each profession are shown in Table 6.2. Academics were assigned the greatest weight per expert as they constituted the smallest proportion of the sample of experts, whereas geriatricians were assigned the lowest weights per expert because they were the most represented in the sample.

Table 6.2. Weights assigned to individual experts in each profession.

	N	Weight per expert
Physiotherapists	10	0.02
Geriatricians	13	0.0154
Nurses and OTs	7	0.0286
GPs	4	0.05
Academics	1	0.2
Total	40	1

6.2.1.3. Method 4: Deriving weights from substantive expertise and perspective

Deriving weights from experts' perspective alone assumed that all experts within each profession contributed equally towards the aggregate prior. Another method for deriving weights was explored, where each perspective (profession) was assigned equal weights (0.2, as described in section 6.2.1.2), but experts within each profession were not weighted equally. Instead only the most substantive experts within each profession were assigned non-zero weights.

The most substantive experts were identified based on their seniority level, patient contact and research experience. Where more than one expert existed with the same perspective and substantive expertise, they were both included and assigned equal weights that summed to 0.2 (the weight assigned to each perspective).

The overview of experts' substantive expertise in each profession are shown in Table 6.3. Only geriatricians had at least one expert with each characteristic, so characteristics of individual experts were analysed to identify the most substantive ones.

Out of the 11 physiotherapists in a Level 2 role, 6 spent over 50% of their time with the target patient population, and none had more than 3 publications or any successful research grant applications. These 6 physiotherapists were thus assigned weight of 0.0333 each (0.2/6)

Three Level 2 geriatricians had 50% patient contact or more. Of the three geriatricians, one had over 50 publications and 1-3 successful research grant applications – more research experience than all other geriatricians. This expert was chosen as the substantive expert within their 'perspective'.

Table 6.3. Number of experts with different characteristics defining substantive expertise, per profession.

	N	Level 2 role	Level 2 role and research experience	Level 2 role and >50% patient contact	Level 2 role, research experience and >50% patient contact
Physiotherapists	10	8	0	6	0
Geriatricians	13	8	6	6	3
Nurses and OTs	7	6	0	4	0
GPs	4	4	2	1	0
Academics	1	1	1	0	0
Total	35	27	9	17	3

Out of 6 nurses and OTs in level 2 roles, expert had less patient contact and research experience than the remaining five experts, and so that experts was excluded and the remaining five were assigned equal weights that summed to 0.2 (i.e. 0.04 each).

Out of 4 GPs, one had over 50% patient contact while another had less than 50% patient contact but more research experience than all other GPs. Both were included in the sample and assigned equal weights than sum to 0.2.

There was only one academic and so they were included in the sample and assigned 0.2 weight.

The resulting number of experts in each profession, and the weights assigned to them are shown in Table 6.4.

Table 6.4. Weights assigned to individual experts in each profession.

	N	Weight per expert
Physiotherapists	9	0.0222
Geriatricians	1	0.2
Nurses and OTs	5	0.04
GPs	2	0.1
Academics	1	0.2
Total	18	1

6.2.1.4. Method 5: Deriving weights from experts' normative expertise

Normative expertise was a binary variable, determined by experts' statistical coherence and internal consistency. Deriving weights from experts' normative expertise aimed to explore the effect of excluding experts whose priors were statistically incoherent, or inconsistent with their verbal responses.

Chapter 4 highlighted that 17 experts were normative, and so all were assigned weights 0.059 (1/17).

6.2.1.5. Method 6: Deriving weights from substantive expertise, perspective and normative expertise

This method ensures that a range of views is represented, and that for each perspective only the most substantive experts are included who are also normative.

Weighting method 4 included the most substantive experts from each perspective. The sample of experts who were assigned non-zero weights in Method 4 was then analysed to identify experts who were also normative. The results are shown in Table 6.5.

Of the 9 substantive physiotherapists 4 were also normative, so these were assigned equal weights that summed to 1.

The most substantive geriatrician was not normative, so geriatricians in level 2 roles who were also normative were identified, and their patient contact and research experience were analysed to identify the most substantive ones. Of 4 such geriatricians, two had less than 50% patient contact and no research experience, one had more than 50% patient contact and more than five publications, and so the latter expert was assigned a weight of 0.2.

Table 6.5. Number of experts with substantive and normative expertise, per profession.

	N	Substantive	Substantive and normative
Physiotherapists	10	6	3
Geriatricians	13	1	0
Nurses and OTs	7	5	2
GPs	4	2	1
Academics	1	1	0
Total	35	15	7

Out of 5 substantive nurses and OTs, two were normative; both had more than 50% patient contact and no research experience, so both were assigned equal weights.

Out of two normative Level 2 GPS, one had no patient contact and no research experience and the other had more than 10% patient contact and more than three publications, and so the latter was included in the sample and assigned a weight of 0.2.

The academic in the sample was not normative but they were included in the sample to represent their perspective, because they were the only academic.

The resulting weights are shown in Table 6.6.

Table 6.6. Weights assigned to individual experts in each profession.

	N	Weight per expert
Physiotherapists	4	0.05
Geriatricians	2	0.1
Nurses and OTs	2	0.1
GPs	1	0.2
Academics	1	0.2
Total	10	1

6.2.1.6. Methods 7: Deriving weights from experts' measured performance - domain seeds

Method 7 used experts' scores on the seeds about falling and fractures without treatment. Experts' priors were scored using KL divergence, as described in Chapter 4. KL scores

decrease with accuracy, and so the weights were derived by inverting the score (1/KL) to ensure that weights increase with accuracy. The inverted scores were scaled so they summed to 1.

6.2.1.7. Method 8: Deriving weights from experts' measured performance - non-domain seeds

The non-domain seed was the number of rainy days in York. Priors were scored using CICP scores, to ensure experts were not penalised for knowledge, only their bias and overconfidence (as discussed in section 3.4.3). The higher the CICP score the more accurate the expert and so their weights were derived by scaling the CICP score so that all weights add up to 1.

6.2.1.8. Method 9: Equal weights

Unweighted priors were derived by assigning equal weights (1/40) to each expert in the sample.

6.2.2. Comparing the effect of weighting methods on the accuracy of the aggregate prior.

The accuracy of weighted priors (using methods in section 6.2.1) was assessed using methods proposed by Cooke et al. (2008). The performance-weighted priors were assessed by splitting the seeds into the training and the testing set, and using training set to derive performance-based weights. The accuracy of the performance-weighted priors from the training set was then compared to the characteristic weighted and unweighted priors on the same seeds. Seed questions about falling and fractures in patients who do not receive treatment were used as the training set (to derive weights), and the treatment effect was

used as the testing set (to assess whether the weighted priors were more accurate than the unweighted ones).

The remainder of this section describes how experts' priors were aggregated (section 6.2.2.1), fitted (section 6.2.2.2) and scored (section 6.2.2.3).

6.2.2.1. Aggregating priors

Experts' priors were aggregated using linear pooling (using Equation 1.5 in Chapter 1). Linear pooling was used as this is the commonly applied aggregation method in HTA. (Soares *et al.*, 2018)

The aggregate prior on each parameter was derived by sampling from each experts' prior, where the number of random samples drawn from each expert's prior was proportional to their weight. A total of 10000 random samples was used for each seed, so for example, if an expert was assigned a weight of 0.2, 2000 samples were drawn from their prior.

6.2.2.2. Fitting

Linear aggregation can result in multimodal probability distributions; to avoid using multimodal priors in the model, the parametric distributions were fitted to the aggregate priors. The seeds were relative risk, and odds and rate ratios, and so log normal distributions were fitted to all parameters, as recommended in the literature (Briggs, Claxton and Schulpher, 2006).

6.2.2.3. Scoring aggregate priors

The aim was to compare the accuracy of aggregate priors derived using different scoring methods. Accuracy was assessed using KL scores, as they penalise bias, imprecision and overconfidence (as discussed in Chapter 2). The methods for deriving scores were described in section 6.2.1).

6.2.3. Observing the impact of different weighting methods on uncertainty in cost-effectiveness analysis

While different weighting methods can affect the accuracy of the aggregate prior, it is not clear whether the effect on the aggregate priors is clinically and economically impactful. To test this, the aggregate priors derived using nine different weighting methods described in section 6.2.1, were used to populate a cost-effectiveness decision model (CEDM), and observe whether the weighting methods affected the resulting model uncertainty.

Building an externally valid model, populated by systematically reviewed data was out of scope of this chapter – the aim was to develop a simplistic model that utilised data from the REFORM trial and the REFORM elicitation exercise to assess the cost-effectiveness of the podiatry intervention designed to prevent falls.

The remainder of this section describes the CEDM. Section 6.2.3.1 describes the model structure, section 6.2.3.2 describes the parameters used to populate the model while section 6.2.3.3 describes the methods for conducting sensitivity analysis. Section 6.2.3.4 then describes how the CEA results based on different aggregate priors were compared.

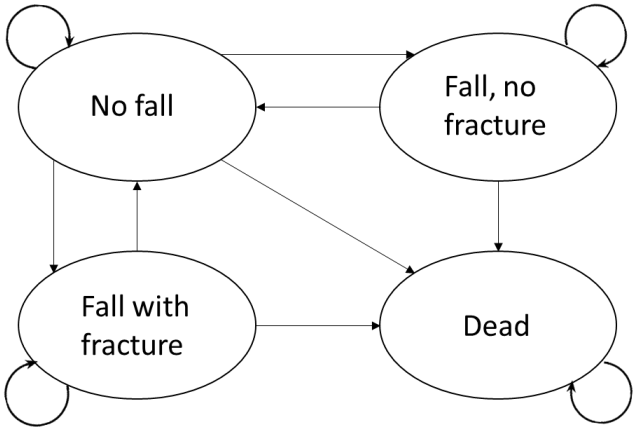
6.2.3.1. The model structure

As highlighted at the beginning of section 6.2.3, the model was designed to utilise data from the REFORM trial and the REFORM elicitation exercise, to assess the cost-effectiveness of the podiatry intervention designed to prevent falls.

Section 3.3.2 in Chapter 3 highlighted that the intervention could affect both the frequency and the severity of falls. In order to reflect any changes in the frequency and the severity of falls a probabilistic four-state Markov model was used, shown in Figure 6.1. Patients were assumed to start in the 'no fall' state; subsequently they could stay in the 'no fall' state, have a fall that did not result in a fracture, have a fall that resulted in a fracture, or die. Each state represents patients' falling behaviour in that cycle, and all health effects and costs associated with a state were assumed to be observed in the cycle in which the patient transitioned into that state. Thus, in subsequent cycles, all living patients could transition into all other states, regardless of their falling history (for example, a patient who had a fall in the first cycle, may not have a fall in the second cycle). The cycle length was 1 year based on previous models on fall prevention (Eldridge *et al.*, 2005). The analysis time horizon was a lifetime, to capture the long term effect of the intervention, and any effect of the intervention on mortality, through reduced risk in falls.

Since the model was designed solely to observe whether the effect of weighting methods was clinically and economically impactful, its clinical plausibility was not explored in further detail in this thesis.

Figure 6.1. Model schematic



6.2.3.2. Model parameters

The CEA was conducted from a health system perspective, reflecting all costs that fall on the health system, and all health effects for the patient. Health effects were measured in terms of QALYs (described in Chapter 1), as recommended by NICE (National Institute for Health and Care Excellence (NICE), 2013) to capture the effect of the intervention on both the length and quality of life.

The cost and utility values for each state were assumed to be the same with and without treatment; i.e. the intervention was assumed to affect costs and outcomes by reducing the number of falls and fractures only.

The REFORM trial reported the overall cost and utility in patients who received the intervention and those who did not. While this is sufficient in trial based analysis, the model in Figure 6.1 required the cost and utility of each individual state, and so other sources were sought.

In order to identify model parameter values from the team of health economists and statisticians involved in the trial based analysis of the REFORM trial were approached. The health economist recommended studies that evaluated cost-effectiveness of intervention designed to prevent falls. The recommended studies were selected for use in the model based on their time and country of publication, where the most recent UK based studies were prioritised. The parameters defined in the model and their sources are shown in Table 6.7.

The cost of treating falls were derived form a falls prevention economic model published by the Chartered Society of Physiotherapy (Chartered Society of Physiotherapy, 2016). The model is a decision tree that calculates the average cost of treatment of minor (£355.00),

moderate (£469.00) and severe falls (£18,352.40). The model in this chapter assumed that falls that results in a fracture were severe, falls without fracture but with injury were moderate, and falls without fracture or injury were mild. REFORM trial reported that 0.58 of falls without fracture resulted in injury. The cost in death state was assumed to be £0.00, all other states were added a cost of £2125.00 – the average NHS spending per head in England (Nuffield Trust, 2014). The cost of social care (for example, incurred by patients moving into sheltered accommodation or care homes after suffering a fall) was not included.

The utility values for no fall and fracture states were derived from a recent UK-based paper on the utility values in the most common fractures. (Svedbom *et al.*, 2018) The authors reported that the average QALY state before having a fracture was 0.84, and having a fracture led to a utility decrement of 0.22. Brazier et al. (2002) reported the utility decrement of falls without fractures to be 0.018. As discussed earlier in this section, the REFORM trial reported that 0.58 falls without fracture resulted in injury, and so a 0.018 utility decrement was applied to 0.58 of falls without fracture. The utility in the death state was assumed to be 0.

The transition probabilities in the first cycle were derived from the REFORM trial. The trial reported the rate of falls and the proportion of falls that resulted in a fracture. The risk of death was used directly in the model, whereas the latter two were converted into transition probabilities using Equation 6.4 to Equation 6.7.

Equation 6.4

$$P(fall) = 1 - e^{-\lambda t}$$

Equation 6.5

$$P(fall, fracture) = P(fall)\ x\ P(fracture|fall)$$

Equation 6.6

$$P(fall, no\ fracture) = P(fall) - P(fall, fracture)$$

Equation 6.7

$$P(no\ fall) = 1 - P(death) - P(fall)$$

The transition probabilities into 'no fall', 'fall-no fracture' and 'fall and fracture' were assumed to be the same irrespective of which state the patient was in.

Table 6.7. Model parameters. SE= standard errors, N=sample size

	Mean (standard error)	Source
<u>QALYs</u>		
Utility in non-fallers (SE)	0.84 (0.13)	(Svedbom *et al.*, 2018)
Disutility after a fall - no fracture, no injury	0.00	Assumption
Disutility after a fall - no fracture, no injury (SE)	0.018 (0.014)	(Brazier and Green, 2002)
Disutility after a fracture (SE)	0.22 (0.17)	(Svedbom *et al.*, 2018)
<u>Costs</u>		(Nuffield Trust, 2014)
Non-fallers	£2125	(Chartered Society of Physiotherapy, 2016)
Fallers - no fracture, no injury	£360	REFORM trial
Fallers – no fracture, injury	£474	
Fallers (fracture)	£18,352	
Cost of intervention - fixed (SE)	£155.79 (55.02)	
<u>Probabilities</u>		
Mortality risk – no fall (age 65-75)	0.023	(Eldridge *et al.*, 2005)
(age 75-85)	0.059	(Eldridge *et al.*, 2005)
(age 85+)	0.144	(Eldridge *et al.*, 2005)
Mortality risk - fall, no fracture (age 65-75)	0.033	(Eldridge *et al.*, 2005)
(age 75-85)	0.069	(Eldridge *et al.*, 2005)
(age 85+)	0.154	(Eldridge *et al.*, 2005)
Mortality risk - fall, fracture (age 65-75)	0.193	(Eldridge *et al.*, 2005)
(age 75-85)	0.229	(Eldridge *et al.*, 2005)
(age 85+)	0.314	(Eldridge *et al.*, 2005)
Rate of falls in control arm (SE)	1.503 (0.105)	REFORM trial
Rate ratio in year 1 (SE)	0.915 (0.057)	REFORM trial
Annual change in the rate ratio	***	Elicited
P(fracture|fall) (N)	0.0175 (762)	REFORM trial
P(no fracture|fall) (N)	0.9825 (762)	REFORM trial
P(no injury|no fracture) (N)	0.42 (748)	REFORM trial
P(injury|no fracture) (N)	0.58 (748)	REFORM trial
Relative risk of fracture in year 1 (SE)	1.324 (0.362)	REFORM trial
Annual change in the relative risk of fracture	***	Elicited

*** varied depending on the aggregation method

The risk of death was obtained from Eldridge et al. (2005). The authors reported the mortality risk in patients who fear falling, patients who do not fear falling and the mortality risk following a fracture. The study reported that patients generally feared falling as result of a fall in the past, and so the risk of death in no fall and fall-no fracture states was assumed to be equal to the risk in patients who do not fear falling and those who do, respectively (Eldridge *et al.*, 2005). The risk was assumed to increase every ten years, as reported in the study. (Eldridge *et al.*, 2005)

The transition probabilities after the first cycle were informed by the priors elicited in the REFORM elicitation study. Chapter 4 described how experts' priors on the treatment effect after the trial end point were converted into the annual change in the treatment effect. The change in the treatment effect could take any value between -1 and infinity, where negative values indicated that the treatment effect would decrease over time, while positive values indicated the treatment effect would increase. When the treatment effect was expected to diminish, it was assumed to change until it became ineffective; so the change in treatment effect was only applied to the transitional probabilities until the treatment effect was 1. When the treatment effect was expected to potentiate, the change in the treatment effect was assumed to decrease for up to 3 cycles – the mean time when experts indicated the treatment effect would plateau (as reported in section 4.4.4).

All costs and effects were discounted at 3.5% rate, as recommended by NICE (National Institute for Health and Care Excellence (NICE), 2013).

6.2.3.3. Sensitivity analysis

Chapter 1 emphasised the need to characterise uncertainty in decision models, and discussed why probabilistic sensitivity analysis was the optimal method for doing this. To this effect, each model parameter was sampled from a probability distribution. The distributions applied to each parameter are shown in Table 6.8.

The model outcomes were measured in terms of the Net Monetary Benefit (defined in Chapter 1, section 1.2.3) and the probability the intervention was cost-effective.

The value of further research (EVPI) was estimated using Equation 1.4 in Chapter 2, while population EVPI ($EVPI_P$) was estimated by multiplying EVPI by the effective population size using **Error! Reference source not found.**.

$$Effective\ population = \sum_{t=1}^{T} \frac{I_t}{(1+r)^t}$$ **Equation 6.8**

Where t is the year of treatment,

I_t is the population affected in year t,

r is the discount rate.

I_t was assumed to be 1,093,894 over 10 years (Chartered Society of Physiotherapy, 2016), while the discount rate was 0.035 (or 3.5%), as recommended by NICE (National Institute for Health and Care Excellence (NICE), 2013).

Table 6.8. Probability distributions of model parameters

Parameter	Distribution
QALYs	
Utility in non-fallers	Beta
Disutility after a fall (no fracture)	Gamma
Disutility after a fall (fracture)	Gamma
Costs	
Non-fallers	Gamma
Fallers (no fracture)	Uniform*
Fallers (fracture)	Uniform*
Intervention fixed cost	Gamma
Transition probabilities	
Risk of death	Beta
Rate of falls in control arm	Gamma
Rate ratio in year 1	Log normal
Annual change in the rate ratio	Log normal**
Probability of fracture after a fall	Beta
Relative risk of fracture in year 1	Log normal
Annual change in the relative risk of fracture	Log normal**

* The authors did not provide information about uncertainty, so the parameter was samples from a uniform distribution and the range of costs was assumed to be ± 10% of the reported value.

** The annual change in treatment effect can take any value between -1 and infinity. The values were thus added 1 before fitting a log normal distribution.

6.2.3.4. Comparison of weighting methods

Nine aggregate priors on the change in the rate ratio for falls and the relative risk of fractures (derived using methods described in section 6.2.1) were used to inform the temporal change in the treatment effect in the model shown in Figure 6.1. The model was ran for each set of aggregate priors separately. Furthermore, the model was once ran without the aggregate priors, where the treatment effect was assumed to remain constant over time. The latter scenario was used to simulate the results of the analysis if experts' priors hadn't been elicited.

For each method, the NHB, probability of cost-effectiveness, and $EVPI_p$ were derived and compared.

6.3. Results: Overview of weighted priors

The aggregate priors were derived using 8 different weighting methods, and compared to unweighted priors. The weighting methods were described in detail in section 6.2.1, and include the following.

- Method 1: Substantive expertise;
- Method 2: Substantive expertise (greater weight applied to the most substantive experts);
- Method 3: Perspective;
- Method 4: Perspective and substantive expertise;
- Method 5: Normative expertise;
- Method 6: Substantive and normative expertise, and perspective;
- Method 7: Domain seeds;
- Method 8: Non-domain seeds;
- Method 9: Unweighted.

Section 6.3.1 shows the aggregate priors on the seed parameters, while section 6.3.2 shows the aggregate priors on the target parameters.

6.3.1. Aggregate priors on seed parameters

The priors on seed parameters derived using different weighting methods, and probability distribution fitted to them are shown in figures Figure 6.2 to Figure 6.7.

Visual inspection of the priors suggests that the goodness of fit varied between priors. Generally, bimodal priors (such as Method 5 in Figure 6.4, Method 4 in Figure 6.6, and all priors in Figure 6.7) and those with pronounced peaks (such as Method 6 in Figure 6.3) were led to less-well fitted probability distributions.

The impact of different weighting methods varied across parameters.

For the relative risk of P(x>0) (in Figure 6.2), all aggregation methods resulted in priors where the majority of the probability density was between 0 and 2. Methods 3,5,7,8 and 9 included values between 9 and 11 because one expert (Expert 32 in Figure 4.11 in section 4.4.3) believed that the RR was close to 10. That expert was normative but not substantive, and so they were assigned 0 weight in methods 1,2,4 and 6, and non-zero weights in methods 3, 5, 7, 8 and 9. However, Expert 32 was an outlier and so the probability distributions fitted to the priors did not include their range (i.e. the probability of the relative risk being between 9 and 11 was very low).

For the relative risk of P(x>5|x>0) (in Figure 6.3) the majority of the probability density was between 0 and 4. Methods that included experts' perspective (3, 4 and 6) assigned greater probability to values in the tail of the distribution. The skewedness was caused by Expert 13 who was the only academic in the exercise and so were assigned a relatively high weight of 0.2 in methods 3, 4 and 6). Expert 13 believed that the relative risk was 3-4 and their inclusion affected the fitted probability distribution as the distributions fitted to priors in methods 1, 3 and 6 placed 0.05 probability on the range 3-4, compared to methods 1,2 and 5 where Expert 13 was assigned 0 weight, where the fitted probability distributions of the same range were 0.013-0.017. (The probabilities were derived by integrating the fitted probability distribution between the two values.)

Priors on the rate of falls (in Figure 6.5) all suggest that the rate ratio is most likely between 0.5 and 1. Method 7 places visibly higher probability on values between 0.25 and 0.5, suggesting that experts who accurately assessed the rate of falls in patients who do not receive the intervention believed the treatment effect would be higher (lower rate ratio). Expert 25 believed that the rate ratio was between 1.8 and 4.2. They were normative and so they were assigned higher weights in Methods 5 and 6 (weights 0.067 and 0.05 respectively, compared to weights 0.013-0.029 in the remaining methods), leading to higher probability being assigned to values between 2 and 4 in those two methods.

Priors on the odds of fracture (in Figure 6.6) appear to be more precise when experts are weighted according to their substantive expertise (Methods 1, 2, 4 and 6).

Priors on the probability of fracture (in Figure 6.7) were derived from the same priors as the odds of fracture, but on an inverted scale. The priors on the relative risk of fractures are overall more precise than those on odds. Three experts believed that the relative risk was between 2.5 and 4 – all three were normative and substantive compared to other experts in their profession, and so they were assigned relatively high weights in Methods 4, 5 and 6 (weights 0.17, 0.2 and 0.3, respectively), compared to Methods 1, 2, 3, 7, 8 and 9 (weights 0.05-0.10), leading to bimodal priors with the second mode between 2.5 and 4.

Comparing across parameters, weights derived from experts' perspective generally leads to less precision (in all priors in Figure 6.2 - Figure 6.7 except the relative risk of P(x>0)). No other consistent effects were identified by visual inspection, further analysis is provided in section 6.4 when scores for each method are presented.

Figure 6.2 Aggregate priors on the RR of P(x>0)

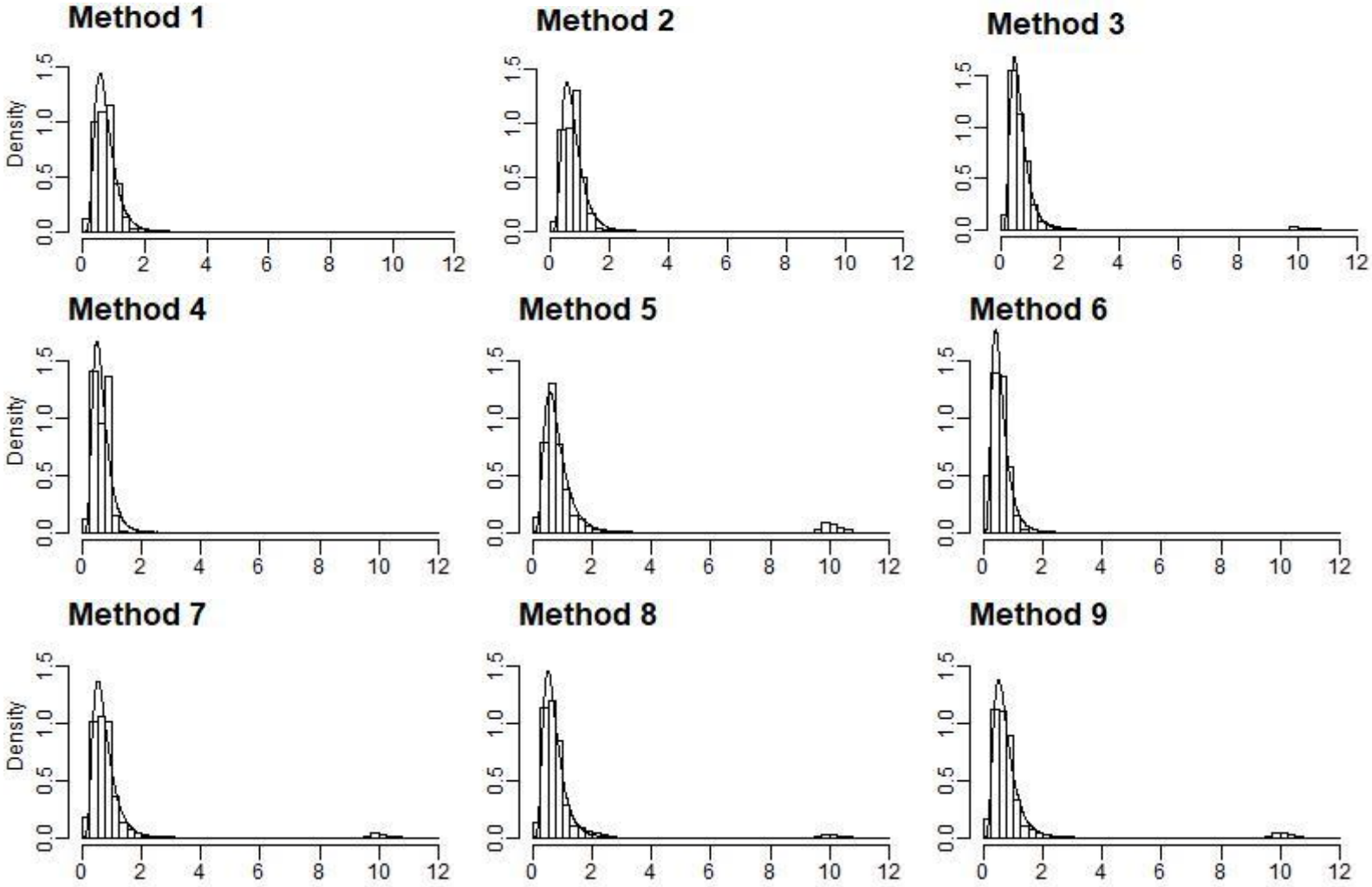


Figure 6.3. Aggregate priors on the RR of P(x>5|x>0)

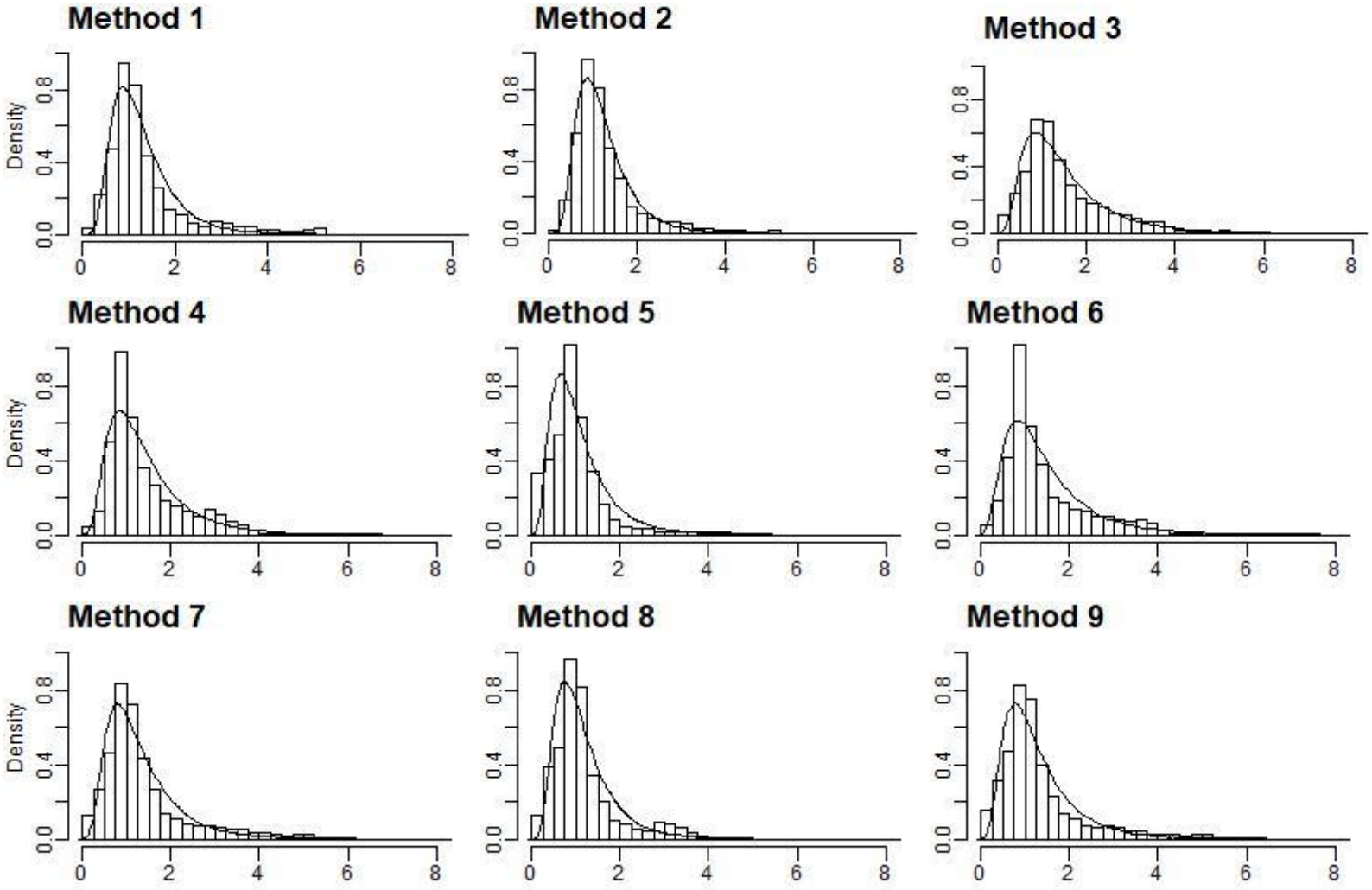


Figure 6.4 Aggregate priors on the RR of P(x>10|x>5)

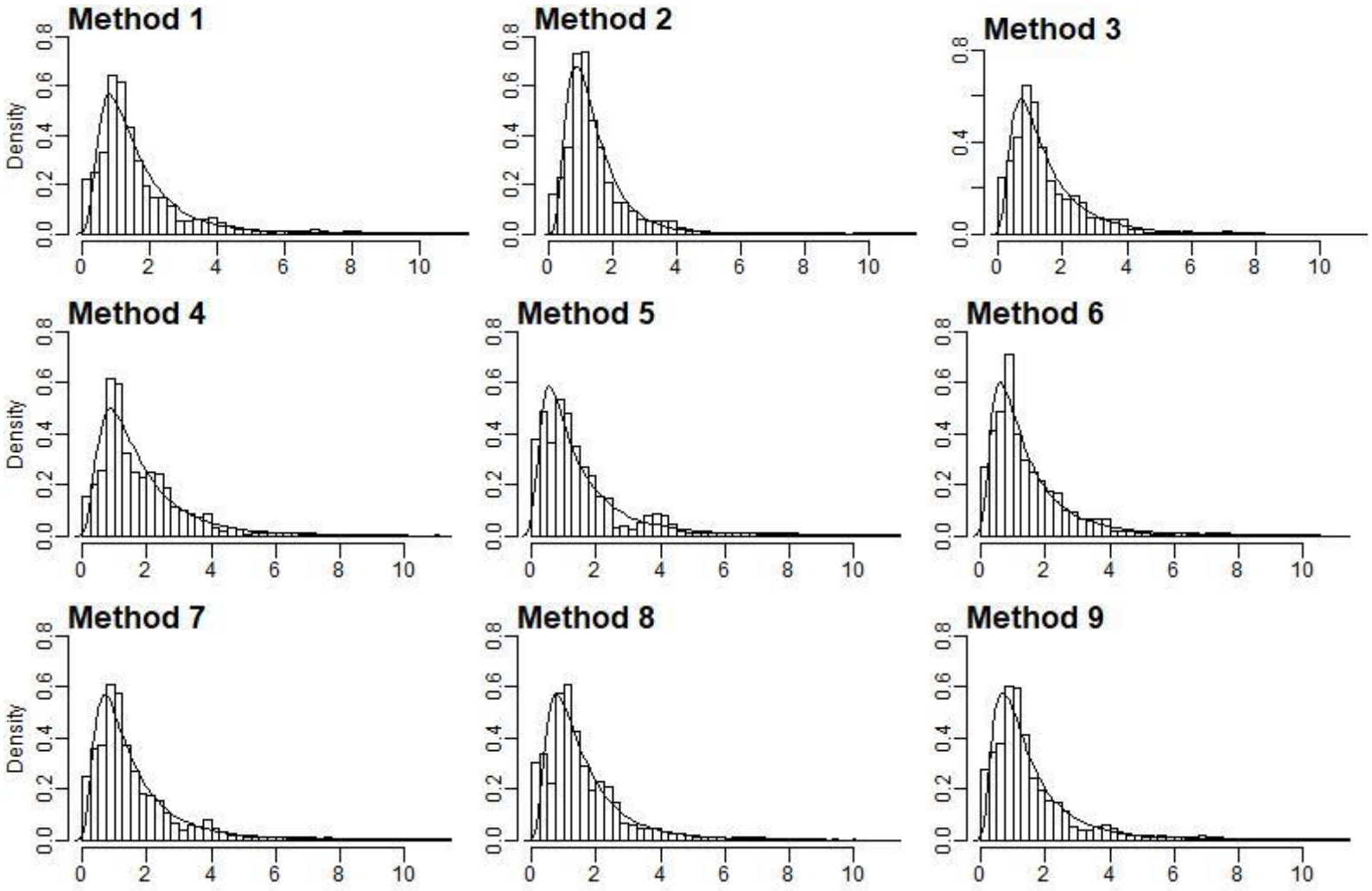


Figure 6.5. Aggregate priors on the rate of falls

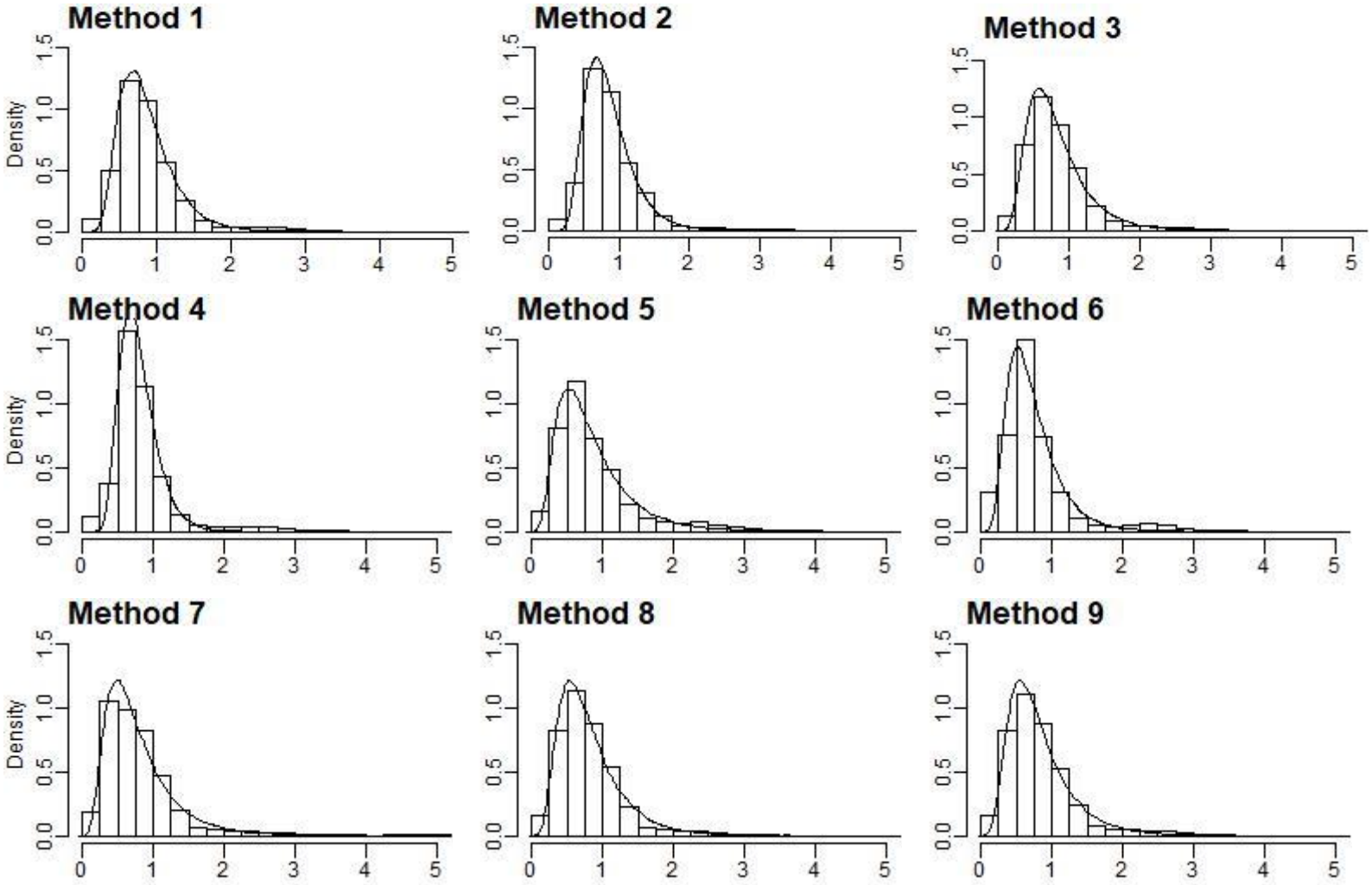


Figure 6.6 Aggregate priors on the OR for fractures

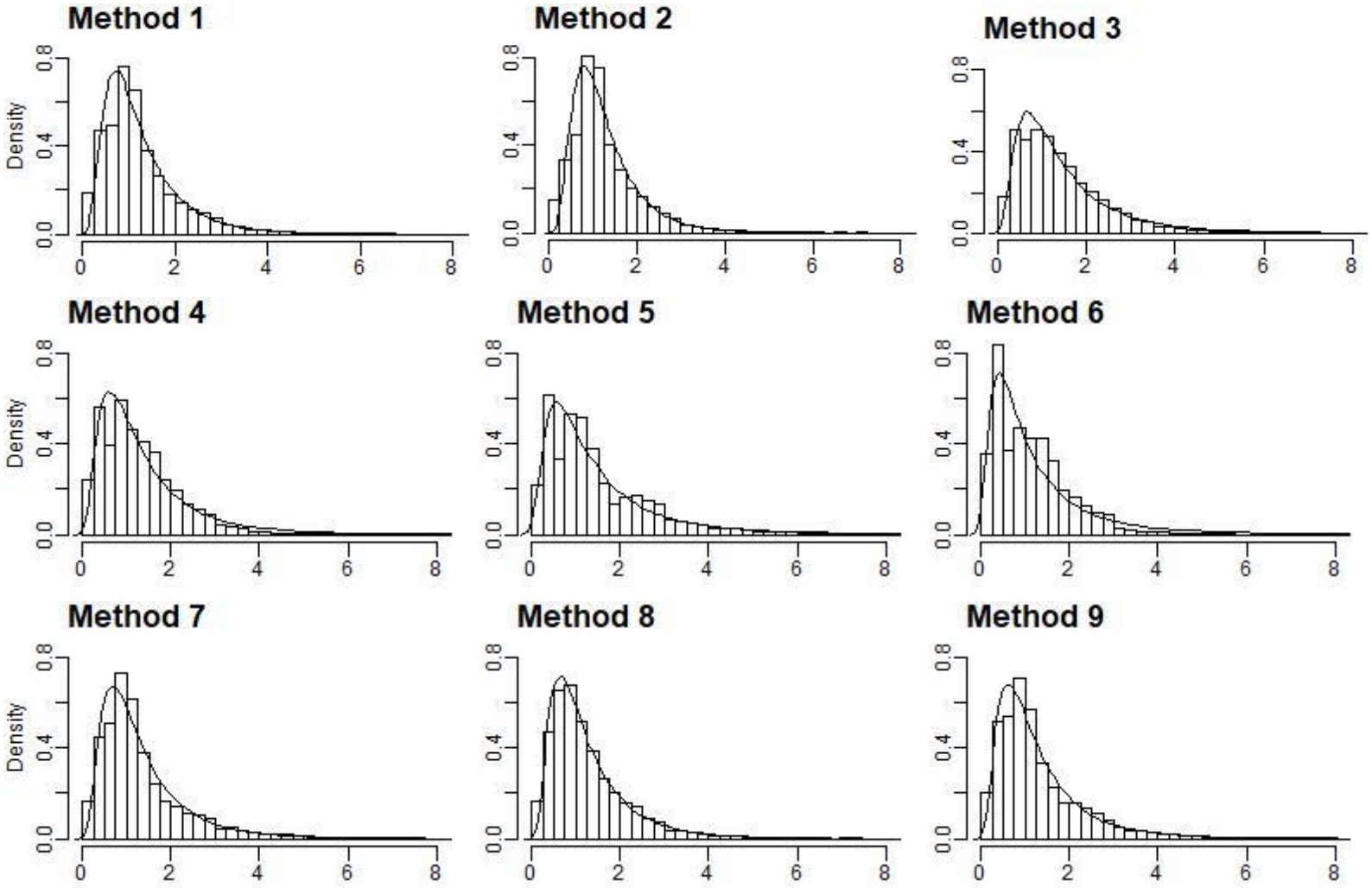
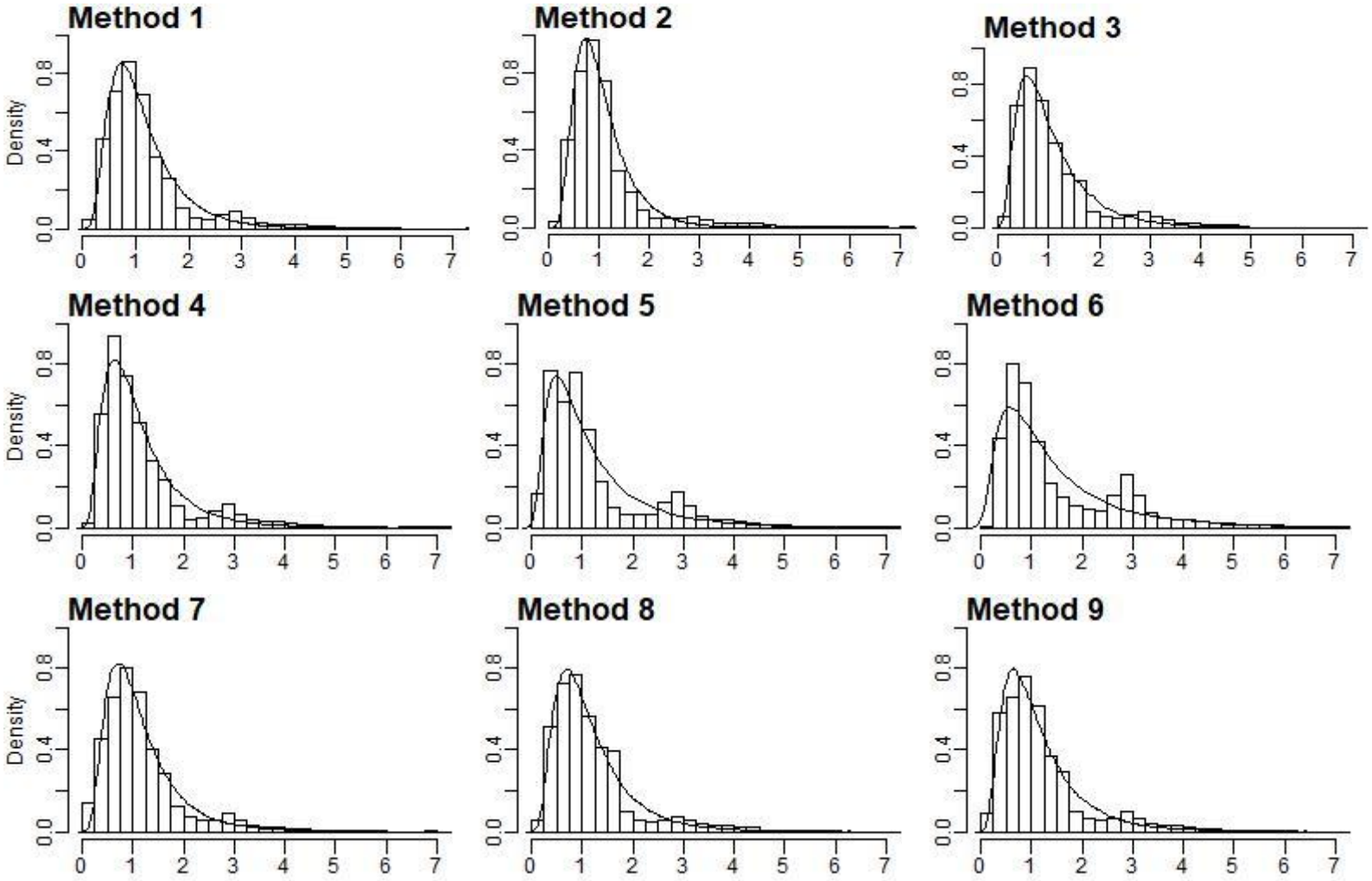


Figure 6.7. Aggregate priors on the RR of P(fracture|fall)



6.3.2. Aggregate priors on target parameters

Priors on target parameters derived using different weighting methods are shown in Figure 6.8 and Figure 6.9.

The majority of priors on the change in the rate ratio are positive and close to 0, suggesting a small increase in the treatment effect over time. Experts predominantly believed that the treatment would be effective, and so an increase in the treatment effect indicates that the treatment effect would diminish.

Priors on the annual change in the rate ratio derived using methods 1 and 2 (substantive expertise) have multiple peaks, caused by two substantive experts (Experts 14 and 40 in Figure 4.13, section 4.4.4) who assessed the change to be more extreme than others in the sample. They were not normative, nor the most substantive in their profession, and so they were not included in methods 4, 5 and 6.

The aggregate priors, like those of individual experts shown in Figure 4.13 (section 4.4.4), were more precise on the relative risk of fractures than the rate ratio of falls.

The change in the relative risk of fractures appears to be more likely to decrease over time than the rate ratio. Figure 6.7 suggested that experts on average believed that the risk of fracture would be higher in patients who receive treatment, and so a decrease in the treatment effect suggests the effect would diminish (the risk of fractures would decrease over time).

Figure 6.8. Aggregate priors on the annual change in rate ratio

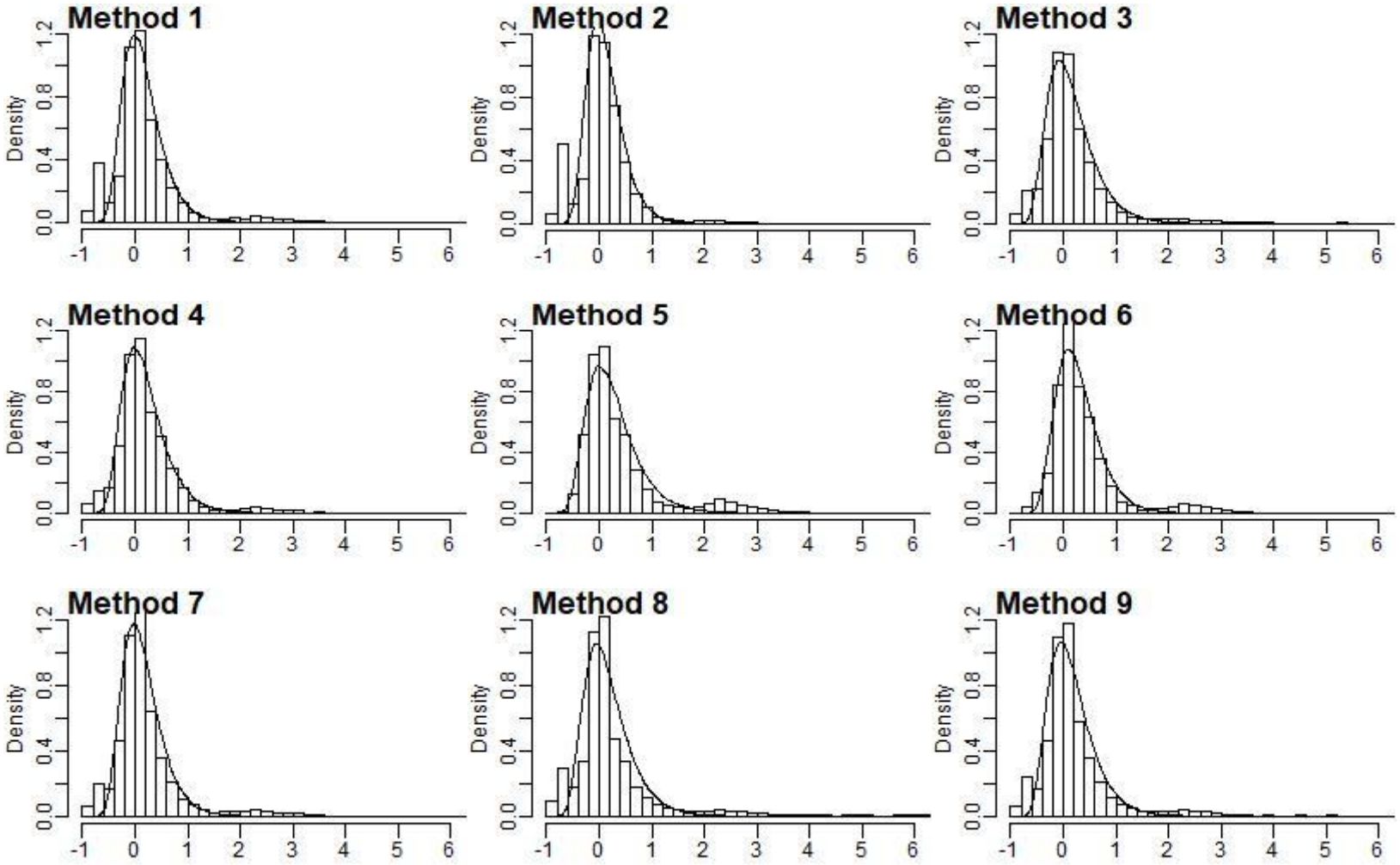
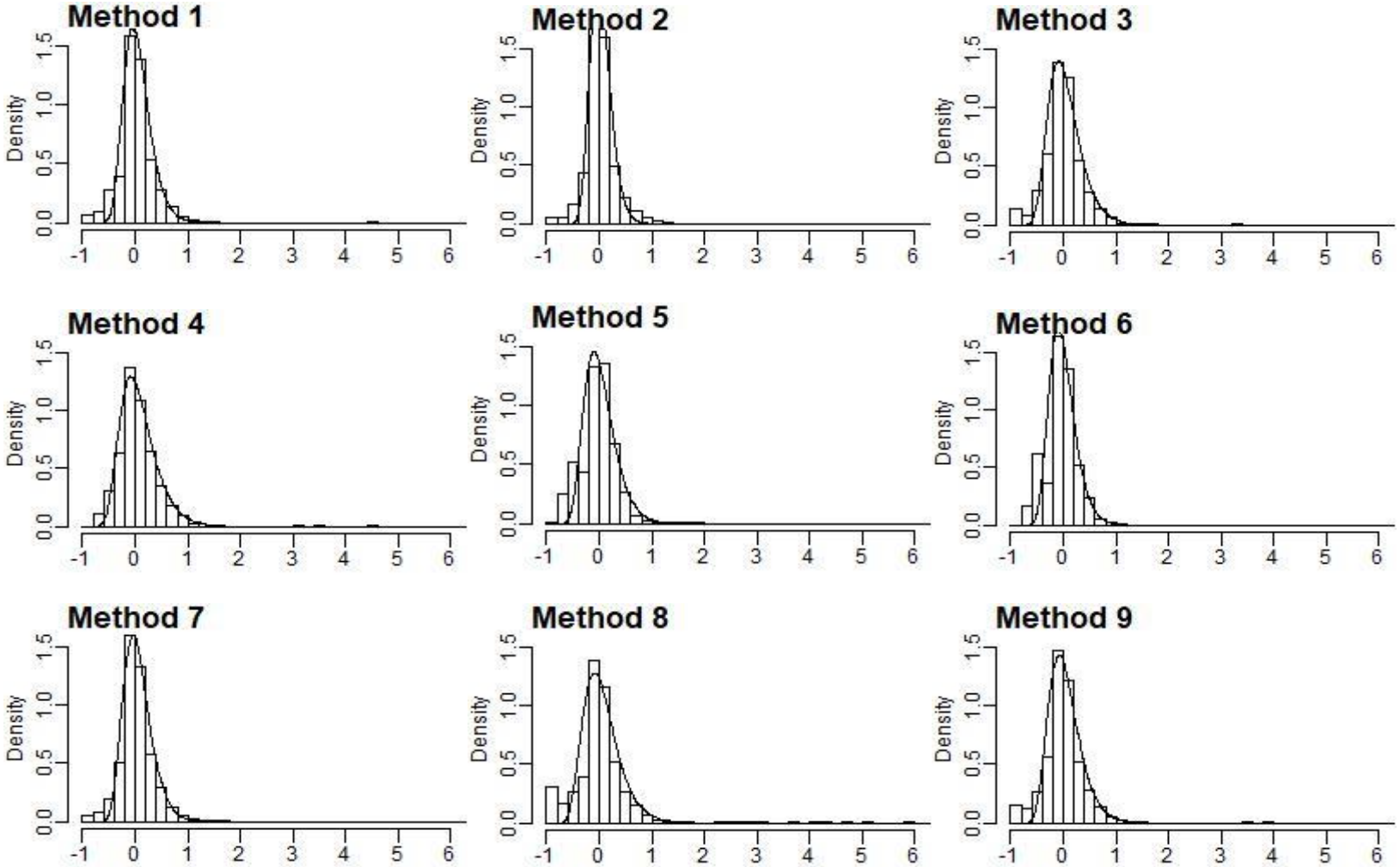


Figure 6.9. Aggregate priors on the annual change in relative risk of fractures.



Aggregate priors on the annual change in the relative risk of fractures were more precise in methods 1 and 2 (when weights were based on substantive expertise) while weights derived from non-domain seeds (method 8) led to the least precise priors. This is likely to be because the scores on the non-domain seed penalised overconfidence, and so assigned greater weight to uncertain experts.

6.4. Results: the effect of weighting methods on the accuracy of the aggregate prior

The accuracy of the aggregate priors on seed parameters was measured using KL scores, ranging from 0 to infinity, where the lower the score the more accurate the prior. The results are shown in Table 6.9. Experts' priors on the relative risk of P(x>0) and the rate ratio for falls were the most biased and overconfident, and so their scores are higher than for the remaining parameters. Differences in scores achieved using different weighting methods also varied as result; for example, the scores on the relative risk of falling more than ten times ranged from 0.286 – 1.061, whereas the scores for the relative risk of falls ranged from 40.5-93.8.

Overall, weights based on substantive expertise (methods 1 and 2) consistently improved the score in every seed compared to the unweighted prior (method 9). Furthermore, method 2 (where greater weights were assigned to more substantive experts) led to more accurate priors than method 1 in 4 out of 6 parameters in Table 6.9, suggesting that placing greater weight on the most substantive experts improved the score of the aggregate prior.

Weights based on substantive expertise and perspective (method 4) were more accurate (attained lower scores) than weights based on perspective only (method 3), although they were not always more accurate than the equally weighted priors (method 9), suggesting that including experts with a range of perspectives does not necessarily improve the accuracy of the aggregate priors, while assigning greater weight to substantive experts does.

Table 6.9. Mean scores for aggregate priors on different seeds, derived using different weighting methods. The lower the score the more accurate the aggregate prior.

		RR for P(x>0)	RR for P(x>5|x>0)	RR for P(x>5|x>0)	Rate ratio	Odds ratio for fractures	P(fracture|fall)
Characteristics	**S** (method 1)	44.9	0.590	0.434	29.9	0.79	0.605
	S (method 2)	40.5	0.481	0.380	25.4	0.75	0.604
	P (method 3)	71.4	1.517	0.619	44.4	1.53	1.207
	SP (method 4)	59.7	1.189	0.286	23.1	1.37	0.932
	N (method 5)	44.1	1.130	1.061	56.6	1.80	1.724
	SNP (method 6)	93.8	1.464	0.847	56.0	1.69	1.276
Performance	**Domain seeds** (method 7)	50.8	1.022	0.621	65.5	1.03	0.683
	Non domain (method 8)	55.3	0.750	0.441	51.7	0.94	0.731
Unweighted (method 9)		55.8	1.083	0.642	50.0	1.09	0.877

Weights derived from substantive and normative expertise, and perspective (method 6) consistently led to worse (higher) scores than equal weighting. Furthermore, weights derived from normative expertise and perspective both led to worse (higher) scores than equal weighting – normative expertise increases scores for all but one seed (relative risk of P(x>0)), while weights based on perspective attained higher scores than unweighted priors in 4 out of 6 seeds.

Performance-weighted priors (methods 7 and 8) led to more accurate aggregate priors than equal weighting on all parameters except the rate ratio. However, they were consistently less accurate (higher scores) than methods 1 and 2. Weights based on non-domain seeds (method 8) led to more accurate priors than those derived from domain seeds (method 7) in 4 out of 6 parameters.

6.5. Results: the effect of weighting methods on the results of the cost-effectiveness analysis

The results of the CEA informed by different weighting methods are shown in Table 6.10.

Overall, the intervention led to a QALY loss in all scenarios, likely because the REFORM trial reported that more patients had a fracture in the treatment arm than in the control arm. While the rate of falls was lower in the control arm, the increase in the risk of fractures meant that overall more people suffered fractures after treatment. Mortality risk was higher following a fracture, and so higher rate of fractures also led to higher mortality, and consequently further QALY loss.

The QALY loss was greater when effectiveness of the intervention was assumed to be constant over time (i.e. when experts' priors were not used in the model). This is likely to be because, as discussed in section 4.4.4, the majority of experts believed that the treatment effect would diminish.

While the intervention incurred additional costs (mean cost £155.79 as discussed in section 6.2.3.2), the incremental cost was negative – i.e. the intervention reduced the cost of treatment. Since the treatment of fractures was costly, and the intervention increased the rate of fractures, the cost reduction is likely to result from the increased mortality.

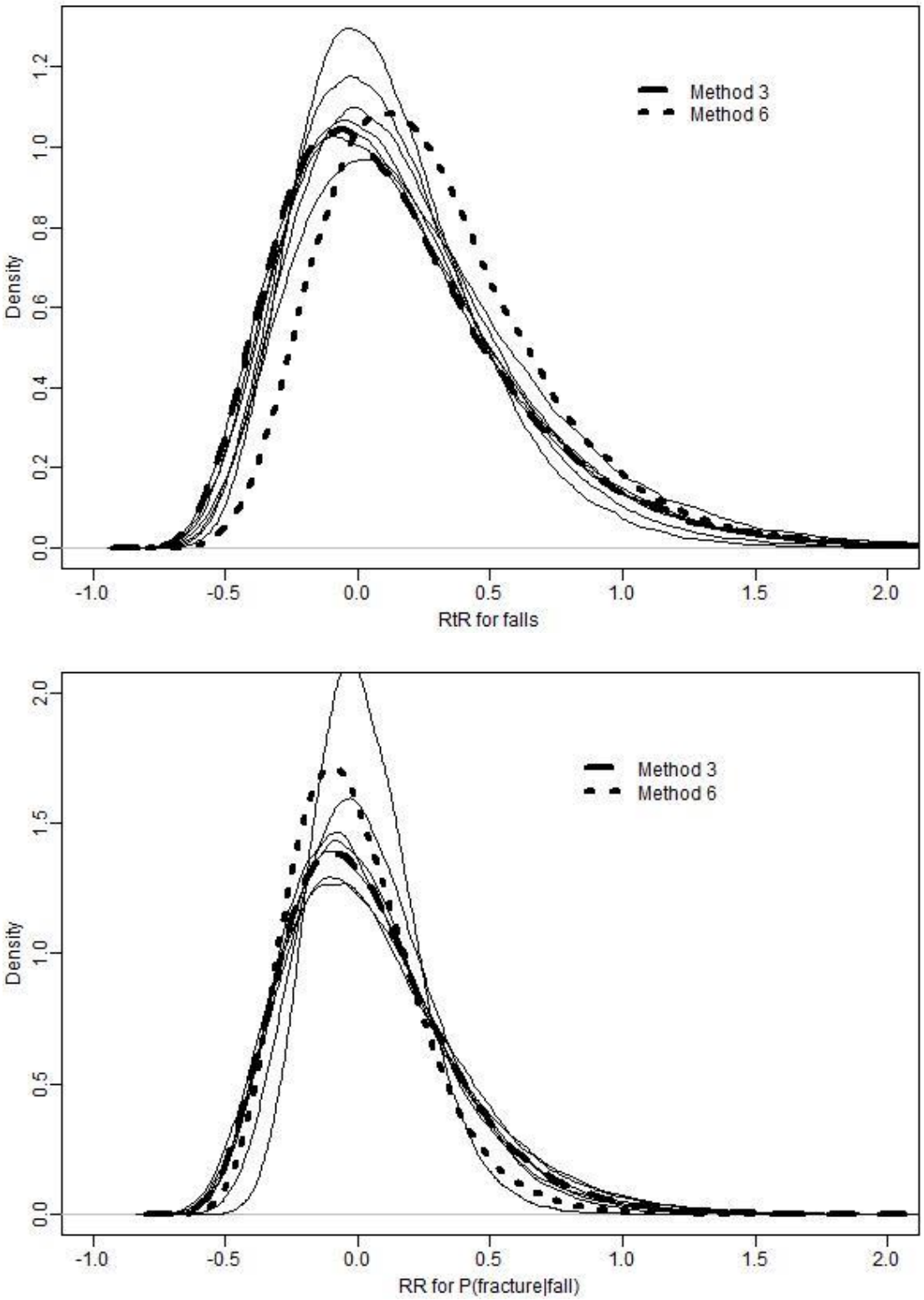
Table 6.10. Results of CEA when different weighting methods are used, in comparison to the assumption of a constant treatment effect over time.

Weighting methods		QALY gain	Incremental cost	NHB	Prob. cost effective	$EVPI_P$ (billions)
Characteristics	**S** (method 1)	-0.328	-7863.50	1291.35	0.4200	8.794
	S (method 2)	-0.307	-8048.75	1898.48	0.4328	8.311
	P (method 3)	-0.200	-8061.59	4056.94	0.4517	8.708
	SP (method 4)	-0.342	-7649.11	817.25	0.4105	9.563
	N (method 5)	-0.356	-7872.31	754.86	0.4064	9.279
	SNP (method 6)	-0.525	-7803.51	-2688.21	0.3609	6.712
Performance	**Domain seeds** (method 7)	-0.300	-7897.65	1896.44	0.4301	8.759
	Non domain (method 8)	-0.242	-7890.75	3154.13	0.4391	8.934
Unweighted (method 9)		-0.246	-7993.20	3074.75	0.4394	8.675
Assume indefinite effectiveness of the intervention (ΔTE=1)		-0.525	-7897.65	-5974.07	0.3258	2.816

When constant treatment effect was assumed over time (i.e. experts' priors were not used in the model) the NHB was negative, while using experts' priors led to positive NHB in all but one scenario (method 6 where weights were based on substantive and normative expertise and perspective).

The methods that resulted in the most extreme disparity in the results of the CEA were explored in further detail to determine what caused the change: these were method 3 (weights based on experts' perspective only) that led to a NHB of 4056.94 and method 6 (weights based on substantive and normative expertise and perspective) that led to negative NHB of -2688.21 (see Table 6.10 for details). The aggregate priors derived using different methods are shown in Figure 6.10.

Figure 6.10. Aggregate priors on the annual change in the rate ratio for falls and the relative risk of fractures derived using methods 3, and 6, in comparison to other methods.



Method 6, which was the only method that led to a negative NHB, shows a higher annual change in the rate ratio (the probability distribution is further right than the other eight priors). This means that, when method 6 was used in CEA, the beneficial effect of the intervention on the rate of falls diminished more quickly than when other aggregate priors were used. Furthermore, the relative risk of fractures derived using method 6 was relatively precise, and close to 0, indicating that the (detrimental) effect of the intervention on the risk of fractures would not change. The opposite is the case when method 3 was used to derive weights – the change in the rate ratio was lower than for other experts, suggesting the beneficial effect of the intervention on the rate of falls would diminish at a lower rate, or potentiate.

When experts' priors were used in the decision model the estimates of cost-effectiveness were more uncertain – the probability of cost-effectiveness was greater than 0.4 in all nine scenarios, compared to 0.3258 when priors were not used. Consequently, the $EVPI_P$ was also higher, although the value of information was high in all scenarios (£2.861bn or greater). In practice, the intervention is unlikely to be recommended or warrant further research, as the positive NHB was driven by the cost reduction due to an increase in mortality.

6.6. Summary of findings

Chapter 6 had three specific objectives:

- To apply different weighting methods identified in Chapter 2 to the REFORM elicitation study.
- To compare the effect of different weighting methods on the accuracy of the aggregate prior.
- To compare the effect of different weighting methods on the results of the cost-effectiveness model used to analyse the results of the REFORM trial.

Section 6.2.1 described 9 different weighting methods derived from various characteristics collected about experts, and their elicitation performance (objective 1). The overview of the priors derived using different weighting methods (shown in Figure 6.2 - Figure 6.9 in section 6.3) did not show any consistent patterns in how weights affected priors (for example in terms of the direction of bias or precision).

However, section 6.4 explored the effect of weighting methods on the accuracy of the aggregate prior (objective 2) and found that weights derived from substantive expertise consistently led to lower (better) scores on seed parameters, and substantive and perspective improved accuracy in comparison to perspective alone (see Table 6.9 for details). Performance weighted priors also improved accuracy in comparison to unweighted ones.

Normative, substantive and perspective consistently led to highest (worst) scores and were consistently less accurate than equally weighted priors.

Section 6.5 explored the effect of different weighting methods on estimates of uncertainty in cost-effectiveness analysis (objective 3).

Including experts' priors in the analysis changed the decision generated by the model in 8 out of 9 scenarios. One weighting method (based on substantive and normative expertise, and perspective) resulted in a different decision to the remaining 8 methods.

The two weighting methods that led to the most disparate NHBs were compared to explore the reason for this variation. The variation was caused by the interaction between the nature of the treatment effect and the change in the treatment effect.

All scenarios indicate high value of information although further research is unlikely to be recommended as the intervention reduced costs by increasing mortality.

It is important to note that the model was founded on several assumptions of uncertain clinical plausibility. For example, all health effects and costs of falls and fractures were assumed to be observed within the year in which they occurred (one cycle), the cost of care following falls and fractures were not taken into account, and the risk of falls and fractures was assumed to be independent of patients' history of falling. Assessment of the plausibility of the stated assumptions was beyond the scope of this chapter. Nevertheless, the results presented in section 6.5 demonstrate an important point – that apparently small differences in the assessment of parameter uncertainty (such as those arising from different weighting methods presented in Figure 6.8 and Figure 6.9), can be clinically and economically impactful, emphasising the importance of developing the methodology for deriving weights.

Chapter 7. Discussion

The aim of this thesis was to develop an understanding of the methodology for deriving weights in expert elicitation, when used to characterise uncertainty in cost-effectiveness decision modelling in health.

Chapter 1 introduced the role of expert elicitation as a tool for characterising uncertainty in cost-effectiveness decision models (CEDM) and proposed that the aim of an elicitation exercise is to capture the current state of knowledge around uncertain quantities (such as model parameters). To do this, elicitation is conducted using formal processes that encourage experts to use all available information and express their priors in an unbiased way. However, in navigating the choices available in designing and conducting an elicitation exercise, there are many methodological uncertainties.

This thesis explored a particular aspect of the elicitation process: the methods for assigning weights to experts' priors when the priors are elicited from multiple experts individually, and mathematically aggregated into a single probability distribution that captures uncertainty in the parameter of interest.

Differential weighting assumes that some experts should be 'given more say' than others and there are multiple methods for deriving weights for experts. The choice of method for deriving weights can affect the resulting estimates of uncertainty, (Cooke, ElSaadany and Huang, 2008) yet it is not clear which method is optimal.

The thesis improved understanding of the existing methods for opinion pooling to ensure that the aggregate priors are an unbiased representation of the current state of knowledge. In order to achieve this, three objectives were set.

1. To identify the existing methods for deriving weights and develop a set of guiding principles for choosing between different options.
2. To apply the principles developed in Chapter 2 to a case study.
3. To observe the consequences of using different methods for opinion pooling.

This sections starts by providing a summary of the findings in each chapter (section 7.1), then a discussion is provided on the key contributions of the thesis (section 7.2). Section 7.3 discusses the limitations, while section 7.4 provides recommendations for further research.

7.1. Summary of findings

Chapter 2 identified the existing methods for deriving weights and developed a set of guiding principles for deriving weights (objective 1). A literature review was first conducted to identify the existing methods for deriving weights – the review identified two general approaches (based on experts' observed characteristics and their measured performance in elicitation), and multiple methods within each approach.

In order to derive guiding principles for choosing between the various options for deriving weights, section 2.3 in Chapter 2 revisited the aims of an elicitation, and then discussed factors that could affect experts' contribution towards achieving those aims.

The chapter proposed that weighting can potentially compensate for methodological challenges in elicitation by giving 'more say' to experts who are believed to be less affected.

Four factors were identified that could affect experts' contribution: substantive expertise, perspective, normative expertise and ability to make accurate probabilistic assessments. Variation in the four factors could provide a basis for differential weighting.

The importance of each factor was proposed to depend on the elicitation process. Expert recruitment, provision of background information and opportunity for discussion with other experts are likely to improve substantive expertise and minimise bias due to perspective, whereas elicitation process design, training and evaluation and feedback help reduce cognitive biases in assessing quantities and expressing uncertainty.

Section 2.4 in Chapter 2 then analysed the assumptions that underpin the existing methods for deriving weights, exploring their role in elicitation.

Chapter 2 concluded that different weighting methods can be used to capture different factors, and understanding where the process lacks can inform which weighting method to use. For example, performance-based weights are affected by experts' normative expertise and their ability to make accurate probabilistic assessments, whereas weights derived from experts' characteristics can be used to capture their substantive expertise independently of their normative expertise.

The challenge in implementing the proposed principles arises from the lack of understanding of how to determine what the challenges in the elicitation process are. For example it is not clear how to demonstrate that training and planning were optimal and that the only basis for differential weighting is substantive expertise.

Furthermore, there are many methodological challenges in deriving weights that make it unclear whether they successfully achieve their objective.

For example, weights derived from experts' characteristics have only been based on their substantive expertise. It is not clear whether characteristics can be used as proxies for other factors that can affect experts' priors.

The characteristics that have been used as proxies for substantive expertise and the derived weights tend to be chosen arbitrarily. It is not clear whether they successfully minimise bias and uncertainty in the weighted aggregate priors.

When seeds are used to derive performance-based weights that are domain-specific, their score on the seed will represent their performance on the target parameter only if their substantive expertise and perspective equally affect the seed and the target parameters. When this is not the case, there is a risk that lower weights will be assigned to experts with unique but important perspective, reducing the heterogeneity of the expert sample. Several applied elicitation exercises have reported this challenge. (Fischer, Lewandowski and Janssen, 2013; Grigore *et al.*, 2016)

The literature review in Chapter 2 highlighted that studies comparing different weighting methods, based on different approaches, are sparse.

The remainder of the thesis thus applied the principles developed in Chapter 2 to a case study in CEDM in order to explore to what extent the different factors that provide basis for differential weighting affected experts' priors and how this affected the role of different methods for deriving weights.

In Chapter 3 different weighting methods were compared in an elicitation exercise applied in CEDM (thesis objective 2). The study was based on a clinical trial (REFORM trial) conducted to measure the clinical and cost-effectiveness of a multifaceted podiatry intervention designed to prevent falls in the elderly.

The REFORM elicitation study was designed to recruit a relatively large sample of experts, collect information about experts' professional experience and elicit a range of seed and target parameters. The trial outcomes were used as seeds, while the change in the treatment effect was used as the target parameter.

The information about experts and their priors on seed parameters were used later in the thesis (in Chapter 5) to explore factors that affect experts' priors – by measuring the effect

of the captured characteristics on their priors on the seed parameters. They were also used to derive weights in Chapter 6, and observe the effect of different weighting methods on the target parameter and on the results of cost-effectiveness model populated by the priors.

Chapter 4 gave an overview of the results of the elicitation exercise, while Chapters 5 and 6 analysed the results to assess the effect of different weighing methods (thesis objective 3).

Specifically, Chapter 5 applied the guiding principles developed in Chapter 2 to score experts' priors on seed parameters, and used the information about experts' professional experience and the derived scores to explore the effect of the captured characteristics on their elicitation performance. The effect of experts' characteristics on non-domain and domain seeds were explored separately.

The non-domain seed was the number of rainy days every September in York. Priors elicited from experts who were recruited in the Yorkshire and Humber region attained better scores than those recruited in other regions of the UK suggesting that the scores were affected by substantive expertise/perspective. Normative expertise was also correlated with better performance.

The domain seeds were the REFORM trial outcomes. Experts' scores varied across seeds more than they did between experts but the effect of expertise was relatively constant across parameters.

Substantive expertise was found to improve scores on domain seeds, but not the non-domain seed, suggesting that substantive expertise does improve scores. The effect was consistent when different definitions of substantive expertise were used, although it was not statistically significant, likely because of the sample size.

The effect of normative expertise and accuracy of probabilistic assessments on the domain seeds was less clear.

While Chapter 5 provided useful insight into factors that affect experts' priors, the implications of the findings for the role of different weighting methods were uncertain. Chapter 5 only explored accuracy of individual experts and it is not clear whether including only more accurate experts improves the accuracy of the aggregate prior, or whether the 'Wisdom of Crowds' outweighs any benefit incurred by only including accurate experts. Furthermore, the impact of the improvement in prior accuracy on the results of cost-effectiveness analysis is not clear. These themes were explored in Chapter 6.

In Chapter 6 the results from the REFORM elicitation study were used to apply different weighting methods identified in Chapter 2, and to compare the effect of the derived methods on the accuracy of the aggregate priors on seed parameters, as well as on the cost-effectiveness analysis of the REFORM trial.

Eight different weighting methods were derived from various characteristics collected about experts, and their elicitation performance. The choice of methods was found to affect the accuracy of experts' priors on the seed parameters.

The findings from Chapter 6 suggest that weights based on experts' substantive expertise and performance-based weights both improved the accuracy of the aggregate priors in comparison to equal weighting, and the former were more accurate than the performance-weighted priors. Weights derived from experts' perspective and normative expertise was detrimental to experts' scores in comparison to unweighted priors.

Chapter 6 also applied the derived weighting methods to the temporal change in the treatment effect. The treatment effect observed in the trial, and the temporal change in the effect derived from experts' priors were used to populate a CEDM, and used to observe the effect of different weighting methods on the cost-effectiveness decision generated by the model, and the resulting value of further research.

Despite apparently small differences in the estimated temporal change in the treatment effect, the different weighting methods were found to affect the decision generated by the model.

7.2. Key contributions of this thesis to the literature

The focus of the thesis has been methods for deriving weights. This thesis has developed an understanding of the differences between existing methods used to derive weights, the effect they may have on aggregate priors and how this affects their role.

Chapter 2 identified the existing methods and analysed the assumptions that underpin each method, allowing transparency when choosing methods and helping direct further research by highlighting the challenge in determining which method is optimal.

Chapters 3-6 then developed a better understanding of the existing methods by applying the guiding principles in a case study.

Several studies have assessed characteristics that affect experts' priors, or compared performance-weighted priors to unweighted ones – these studies tend to analyse results of elicitation exercises reported in databases, retrospectively. For example, Nemet et al. (2017) measured the effect of experts' characteristics and elicitation process design on the width of experts' 80% confidence interval in their judgments about future energy technologies. The studies evaluating and comparing different weighting methods are generally based on the applied exercises in the TU Delft database (Goossens, 2008b; S Lin and Cheng, 2009; Flandoli *et al.*, 2011; Colson and Cooke, 2017). Using findings from databases can limit the characteristics and weighting methods that can be compared.

The REFORM elicitation study is the first study prospectively designed to evaluate and compare weighting methods in CEDM; prospectively designing the exercise to compare different weighting methods allowed exploration of:

- Methods for capturing factors that affect experts' priors;
- Methods for scoring experts priors;
- Methods for eliciting and scoring different types of parameters.

The results of the applied work suggest that experts' priors are influenced by their substantive expertise, and that both substantive characteristics and performance-based seeds can improve the accuracy of aggregate priors.

It is important to note that generalisability, of the findings is unclear, because the study was conducted in a very specific setting. Training, conduct of elicitation (e.g. face-to-face vs remote delivery), elicited parameters and background information can all affect elicitation results, and consequently the relative importance of different factors thought to affect experts' priors. Furthermore, the methods applied in the REFORM elicitation exercise may not be applicable in other case studies; for example, when analysis is not conducted alongside a trial and so trial outcomes cannot be used as domain-specific seeds. Further research on weighting methods is required, across a range of settings, to develop guidance on the optimal method to derive weights. Nevertheless, the study provides an initial step in understanding the optimal approach to weighting.

In addition to improving the understanding of the existing weighting methods, the REFORM elicitation study has developed elicitation methods more broadly. The REFORM elicitation study developed methods for eliciting complex parameters in a remotely delivered

elicitation exercise. The elicited parameters included the rate of falls, treatment effect, and the temporal change in the treatment effect.

7.3. Limitations

This section describes the methodological challenged encountered in this thesis in turn.

Literature review

Section 2.2 conducted a literature review to address the first objective of the thesis, and identified two approaches to deriving weights, and multiple methods within each approach.

The literature review was a non-systematic BCSC. A potential caveat of the BCSC method is the reliance on authors' referencing to identify relevant publications (Hinde and Spackman, 2015). If a publication is insufficiently referenced, and it has not been cited by other publications on the topic of interest, it can lead to a 'citation island'. Any such citation islands could have been missed from the literature search. In order to minimise the risk of missing citation islands, the initial pearls were selected from a range of fields, and included general elicitation references (Cooke, 1991; P Garthwaite, J Kadane and O'Hagan, 2005; O'Hagan *et al.*, 2006).

Furthermore the search strategy was focused on applied examples in HTA. It is possible that additional weighting methods that have been applied in other fields were missed. The methods highlight the difficulty in carrying out systematic searches in elicitation due to its widespread application and varied terminology.

REFORM elicitation study protocol

When designing the REFORM elicitation exercise, methodological challenges arose from the lack of understanding of how best to elicit different types of quantities (as discussed in Chapter 1). The aim was to elicit experts' beliefs on the treatment effect of the podiatry intervention evaluated in the REFORM trial, and the temporal change in the treatment effect. A decision was made to elicit both quantities indirectly, based on practice in previous exercises that elicited the treatment effect. (Bojke *et al.*, 2010; Soares *et al.*, 2011) the treatment effect was assumed to be independent of the rate of falls and risk of fractures.

The elicitation methods required experts' beliefs on the rate of falls to be elicited in those patients who receive the intervention and those who do not. The skewed distribution of the frequency of falls made it a difficult parameter for experts to assess. There were no identified studies for deriving rates and so a novel method for eliciting rates indirectly was derived where a series of binomial distributions was elicited and conditional probabilities of different outcomes (number of falls) were assumed to be independent. It is unclear which of the two methods (direct or indirect elicitation of rates) is better and so the decision to use the indirect method was based on the results of a pilot where the participants expressed the indirect method to be more intuitive, and led to more comparable results between them. It is not clear whether this also makes it a better method.

Alternative methods for eliciting multinomial distributions exist – for example eliciting the multinomial distribution for the frequency of falls and correlation between conditional probabilities of different number of falls (Clemen, Fischer and Winkler, 2000), but the methods require extensive training and active guidance by the investigator (Bojke *et al.*, 2017) and so were not feasible in this study.

The plausibility of assumptions imposed by the elicitation methods were evaluated in section 4.5.1 and found no evidence that the assumptions were implausible.

Methods for capturing experts' characteristics

Chapter 2 highlighted the lack of understanding of which characteristics should be used as proxies for substantive expertise, perspective, normative expertise and the ability to make accurate probabilistic assessments. The REFORM elicitation study proposed using experts' role, research experience, research awareness and patients contact to reflect their substantive expertise, to use experts' statistical coherence to capture their normative expertise, and to capture their perspective using their profession, although more research in this field is required.

Methods for assessing the effect of experts' characteristics on their elicitation performance

While the REFORM elicitation study used multiple characteristics to capture experts' performance, it was not possible to assess the impact of each characteristic due to the small sample size of experts (n=41). This is a common challenge in elicitation in CEDM where the average number of experts recruited is 8.83 (Soares *et al.*, 2018).

Methods for observing the impact of different weighting methods in cost-effectiveness analysis

The model used in Chapter 6 was founded on several assumptions of uncertain clinical plausibility. Assessment of the plausibility of the stated assumptions was beyond the scope of this chapter. Nevertheless, the results presented in section 6.5 demonstrate an important point – that apparently small differences in the assessment of parameter uncertainty arising from different weighting methods, can be clinically and economically impactful, emphasising the importance of developing the methodology for deriving weights.

7.4. Further research

The impetus for research in this thesis was to understand how to use elicitation methods in a way that generate unbiased, informed estimates of uncertain quantities for use in CEDM. The findings suggest that characteristics and performance-based weights can improve the accuracy of the aggregate priors but further research is required to fully understand what the optimum elicitation methods are. Three specific research streams are discussed here that could help inform the optimum implementation of elicitation methods.

1. Develop methods for selecting experts.

This thesis suggested that priors elicited form substantive experts were more accurate than those elicited from non-substantive experts. The findings could potentially be applied to ensure only substantive experts are recruited, rather than to recruit vast pools of experts and use their experience to weight them. However, further research is required to determine what characteristics should be sought in the recruitment process.

2. Develop methods for within-elicitation validity assessments.

The REFORM elicitation study used a relatively large sample of experts and extensive analysis to determine which weighting method led to the most accurate aggregate prior in the case study. Given the many methodological uncertainties, and options for delivering elicitation, there is a need for within elicitation validity assessments to give indication of how well the exercise has done. Chapter 1 highlighted that methods for assessing internal validity exist, there are no empirical evidence-backed guidelines on how to assess validity of an elicitation exercise.

3. Experiment-based research on elicitation methods.

Given the many variables in the elicitation process (discussed in Chapter 1) the generalisability of methodological studies is often uncertain. Experiment-based research can help resolve specific methodological challenges by testing hypotheses in controlled conditions.

REFERENCES

Aspinall, W. P. and Cooke, R. (2013) 'Quantifying scientific uncertainty from expert judgement elicitation', in Rougier, J., Sparks, S., and L, H. (eds) *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge: Cambridge University Press.

Ayyub, B. M. (2001) *Elicitation of expert opinions for uncertainty and risks*. CRC Press. Available at: https://www.crcpress.com/Elicitation-of-Expert-Opinions-for-Uncertainty-and-Risks/Ayyub/p/book/9780849310874 (Accessed: 23 February 2018).

Balanowski, K. R. and Flynn, L. M. (2005) 'Effect of painful keratoses debridement on foot pain, balance and function in older adults', *Gait & Posture*, 22(4), pp. 302–307. doi: 10.1016/j.gaitpost.2004.10.006.

Bojke, L. *et al.* (2010) 'Eliciting distributions to populate decision analytic models.', *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 13(5), pp. 557–64. doi: 10.1111/j.1524-4733.2010.00709.x.

Bojke, L. *et al.* (2017) 'Informing Reimbursement Decisions Using Cost-Effectiveness Modelling: A Guide to the Process of Generating Elicited Priors to Capture Model Uncertainties', *doi.org*. Springer Nature, pp. 1–11. doi: 10.1007/s40273-017-0525-1.

Bolger, F. (2017) 'The Selection of Experts for (Probabilistic) Expert Knowledge Elicitation', in Dias, L. C., Morton, A., and Quigley, J. (eds) *Elicitation. The Science and Art of Structuring Judgement.* Cham: Springer International Publishing, pp. 393–444.

Bolger, F. and Rowe, G. (2015) 'The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight?', *Risk Analysis*, 35(1), pp. 5–11. doi: 10.1111/risa.12272.

Bowling, A. (2005) 'Mode of questionnaire administration can have serious effects on data quality', *Journal of Public Health*, 27(3), pp. 281–291. doi: 10.1093/pubmed/fdi031.

Box, G. E. P. (1979) 'Robustness in the strategy of scientific model building', in Launer, R. L. and Wilkinson, G. N. (eds) *Robustness in Statistics.* Academic Press, pp. 201–236.

Brazier, J. E. and Green, C. (2002) 'A Systematic Review of Health State Utility Values for Osteoporosis-Related Conditions', *Osteoporosis International*, 13(10), pp. 768–776. doi: 10.1007/s001980200107.

Brier (1950) 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather*

Review1, 78, pp. 1–3.

Briggs, A., Claxton, K. and Schulpher, M. (2006) *Decision Modelling for Health Economic Evaluation*. New York: Oxford University Press.

Brockhoff, K. (1975) 'The performance of forecasting groups in computer dialogue and face to face discussions', in Linstone, H. A. and Turoff, M. (eds) *The Delphi Method, Techniques and Applications*. Reading: Wesby Publishing, pp. 291–321.

Brown, A. J. and Aspinall, W. P. (2004) 'Use of expert opinion elicitation to quantify the internal erosion process in dams', in *Proceedings of the British Dam Society Conference: Long-term Benefits and performance of Dams*. Canterbury: Thomas Telford, pp. 282–297.

Budescu, D. V. and Chen, E. (2015) 'Identifying Expertise to Extract the Wisdom of Crowds', *Management Science*. INFORMS , 61(2), pp. 267–280. doi: 10.1287/mnsc.2014.1909.

Burgman, M. A. *et al.* (2011) 'Expert Status and Performance', *PLoS ONE*. Edited by A. Szolnoki. Elsevier, 6(7), p. e22998. doi: 10.1371/journal.pone.0022998.

Chaloner, K. *et al.* (1993) 'Graphical Elicitation of a Prior Distribution for a Clinical Trial', *The Statistician*. WileyRoyal Statistical Society, 42(4), p. 341. doi: 10.2307/2348469.

Chartered Society of Physiotherapy (2016) 'The falls prevention economics model'. Available at: https://www.csp.org.uk/documents/falls-prevention-economic-model.

Claxton, K. *et al.* (2005) 'Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra', *Health Economics*, 14(4), pp. 339–347. doi: 10.1002/hec.985.

Claxton, K. (2008) 'Exploring uncertainty in cost-effectiveness analysis.', *PharmacoEconomics*, 26(9), pp. 781–98. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18767898 (Accessed: 6 February 2018).

Claxton, K. *et al.* (2015) 'Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold', *Health Technology Assessment*, 19(14), pp. 1–504. doi: 10.3310/hta19140.

Claxton, K., Sculpher, M. and Drummond, M. (2002) 'A rational framework for decision making by the National Institute For Clinical Excellence (NICE)', *The Lancet*, 360(9334), pp. 711–715. doi: 10.1016/S0140-6736(02)09832-X.

Clemen, R. T. (2008) 'Comment on Cooke's classical method', *Reliability Engineering & System Safety*, pp. 760–765. doi: 10.1016/j.ress.2008.02.003.

Clemen, R. T., Fischer, G. W. and Winkler, R. L. (2000) 'Assessing Dependence: Some Experimental Results', *Management Science*. INFORMS, 46(8), pp. 1100–1115. doi: 10.1287/mnsc.46.8.1100.12023.

Clemen, R. T. and Winkler, R. L. (1999) 'Combining Probability Distributions From Experts in Risk Analysis', *Risk Analysis*. Kluwer Academic Publishers-Plenum Publishers, 19(2), pp. 187–203. doi: 10.1023/A:1006917509560.

Clemen, R. T. and Winkler, R. L. (2007) 'Aggregating Probability Distributions', in Edwards, W., Miles, R. F. J., and von Winterfeldt, D. (eds) *Advances in Decision Analysis*. Cambridge: Cambridge University Press, pp. 154–176. doi: 10.1017/CBO9780511611308.010.

Cockayne, S. *et al.* (2014) 'The REFORM study protocol: a cohort randomised controlled trial of a multifaceted podiatry intervention for the prevention of falls in older people.', *BMJ open*. British Medical Journal Publishing Group, 4(12), p. e006977. doi: 10.1136/bmjopen-2014-006977.

Colson, A. R. and Cooke, R. M. (2017) 'Cross validation for the classical model of structured expert judgment', *Reliability Engineering & System Safety*, 163, pp. 109–120. doi: 10.1016/j.ress.2017.02.003.

Colson, A. R. and Cooke, R. M. (2018) 'Expert elicitation : using the classical model to validate experts' judgments.', *Review of Environmental Economics and Policy*, 12(1), pp. 113–32. doi: 10.1093/reep/rex022.

Commonwealth Department of Health, H. a. C. S. (1992) *Guidelines for the pharmaceutical industry on preparation of submissions to the Pharmaceutical Benefit Advisory Committee*. Canberra.

Cooke, R. (1991) *Experts in uncertainty : opinion and subjective probability in science*. Oxford University Press. Available at: https://books.google.co.uk/books?hl=en&lr=&id=5nDmCwAAQBAJ&oi=fnd&pg=PR5&dq=cooke+experts+in+uncertainty&ots=5RmKXUGeK8&sig=npvdnzo47-yhxUaqcYOGvU21NUM#v=onepage&q=cooke experts in uncertainty&f=false (Accessed: 2 May 2017).

Cooke, R. (2008) 'Response to discussants', *Reliability Engineering and System Safety*, 93, pp. 775–777. doi: 10.1016/j.ress.2008.02.006.

Cooke, R. (2017) 'Elicitation in the Classical Model', in Dias, L. C., Morton, A., and Quigley, J.

(eds) *Elicitation. The Science and Art of Structuring Judgement.* Cham: Springer International Publishing.

Cooke, R., ElSaadany, S. and Huang, X. (2008) 'On the performance of social network and likelihood-based expert weighting schemes', *Reliability Engineering & System Safety*. Elsevier, 93(5), pp. 745–756. doi: 10.1016/J.RESS.2007.03.017.

Cooke, R. and Goossens, L. H. J. (1999) *'Procedures guide for structured expert judgment'. EUR 18820*. Delft, The Netherlands.

Cooke, R. and Goossens, L. H. J. (2006) *TU Delft expert judgment data base*. Delfth, The Netherlands.

Cooke, R. M. and Goossens, L. H. J. (2000) 'Procedures Guide for Structural Expert Judgement in Accident Consequence Modelling', *Radiation Protection Dosimetry*. Oxford University Press, 90(3), pp. 303–309. doi: 10.1093/oxfordjournals.rpd.a033152.

Corder, G. W. and Foreman, D. I. (2009) *Nonparametric Statistics for Non-Statisticians.* Hoboken: John Wiley & Sons.

Degroot, M. H. (1974) 'Reaching a Consensus', *Journal of the American Statistical Association*, 69(345), pp. 118–121. doi: 10.1080/01621459.1974.10480137.

Drummond, M. F. *et al.* (2015) *Methods for the Economic Evaluation of Health Care Programmes*. OUP Oxford. Available at: https://books.google.com/books?hl=en&lr=&id=yzZSCwAAQBAJ&pgis=1 (Accessed: 11 March 2016).

EFSA (2014) *Guidance on expert knowledge elicitation in foor and feed safety risk assessment*.

Eggstaff, J. W., Mazzuchi, T. A. and Sarkani, S. (2014) 'The effect of the number of seed variables on the performance of Cooke's classical model', *Reliability Engineering & System Safety*. Elsevier, 121, pp. 72–82. doi: 10.1016/J.RESS.2013.07.015.

Eldridge, S. *et al.* (2005) *Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in older people*, *Journal of Health Services Research & Policy*. Available at: https://pdfs.semanticscholar.org/9e4b/f3c75d3ce7c5296a125c99ee5b586ebb9fc0.pdf (Accessed: 10 September 2018).

Epstein, E. (1969) ‘A scoring rule system for probability forecasts of ranked categories’, *Journal of Applied Meteorology*, 8, pp. 985–987.

Ericsson, K. (2006) ‘The influence of experience and deliberate practice on the development of superior expert performance’, in Ericsson, K. A. et al. (eds) *Cambridge Handbook of Expertise and Expert Performance*. Cambridge, UK: Cambridge University Press, pp. 685–706.

Evans, M., Hastings, N. and Peacock, B. (2000) *Statistical Distributions*. 3rd ed. New York: Wiley.

Ferrell, W. R. (1985) ‘Combining Individual Judgments’, in *Behavioral Decision Making*. Boston, MA: Springer US, pp. 111–145. doi: 10.1007/978-1-4613-2391-4_6.

Ferrell, W. R. (1994) ‘Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination.’, in Wright, G. and Ayton, P. (eds) *Subjective Probability*. New York: Wiley.

Fischer, K., Lewandowski, D. and Janssen, M. P. (2013) ‘Estimating unknown parameters in haemophilia using expert judgement elicitation’, *Haemophilia*. Wiley/Blackwell (10.1111), 19(5), pp. e282–e288. doi: 10.1111/hae.12166.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1977) ‘Knowing with certainty: The appropriateness of extreme confidence.’, *Journal of Experimental Psychology: Human Perception and Performance*, 3(4), pp. 552–564. doi: 10.1037/0096-1523.3.4.552.

Flandoli, F. *et al.* (2011) ‘Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique’, *Reliability Engineering & System Safety*. Elsevier, 96(10), pp. 1292–1310. doi: 10.1016/J.RESS.2011.05.012.

Garfield, E. (2006) ‘Citation indexes for science. A new dimension in documentation through association of ideas’, *International Journal of Epidemiology*. Oxford University Press, 35(5), pp. 1123–1127. doi: 10.1093/ije/dyl189.

Garthwaite, P. H., Kadane, J. B. and O’Hagan, A. (2005) ‘Statistical Methods for Eliciting Probability Distributions’, *Journal of the American Statistical Association*, 100, pp. 680–701. Available at: http://www.stat.cmu.edu/tr/tr808/tr808.pdf (Accessed: 15 February 2016).

Garthwaite, P. H., Kadane, J. B. and O’Hagan, A. (2005) ‘Statistical Methods for Eliciting Probability Distributions’, *Journal of the American Statistical Association*. Taylor & Francis,

100(470), pp. 680–701. doi: 10.1198/016214505000000105.

Garthwaite, P., Kadane, J. and O'Hagan, A. (2005) 'Statistical Methods for Eliciting Probability Distributions', *Journal of the American Statistical Association*, 100(470), pp. 680–701. doi: 10.1198/016214505000000105.

Garthwaite, P. and O'Hagan, A. (2000) 'Quantifying expert opinion in the UK water industry: an experimental study', *The Statistician*, 49(4), pp. 455–477.

Genest, C. and McConway, K. J. (1990) 'Allocating the weights in the linear opinion pool', *Journal of Forecasting*. Wiley-Blackwell, 9(1), pp. 53–73. doi: 10.1002/for.3980090106.

Genest, C. and Zidek, J. V. (1986) 'Combining Probability Distributions: A Critique and an Annotated Bibliography', *Statistical Science*. Institute of Mathematical Statistics, pp. 114–135. doi: 10.2307/2245510.

Gigerenzer, G. (1996) 'The Psychology of Good Judgment', *Medical Decision Making*, 16(3), pp. 273–280. doi: 10.1177/0272989X9601600312.

Goossens, L. L. H. J. (2008a) 'TU Delft expert judgment data base', *Reliability Engineering & System Safety*, 93(5), pp. 657–674. doi: 10.1016/j.ress.2007.03.005.

Goossens, L. L. H. J. (2008b) 'TU Delft expert judgment data base', *Reliability Engineering & System Safety*, 93(5), pp. 657–674. doi: 10.1016/j.ress.2007.03.005.

Gosling, J. P. (2014) *Methods for eliciting expert opinion to inform health technology assessment*. Available at: https://www.mrc.ac.uk/documents/pdf/methods-for-eliciting-expert-opinion-gosling-2014/ (Accessed: 26 April 2017).

Griffin, S. C. *et al.* (2011) 'Dangerous omissions: the consequences of ignoring decision uncertainty', *Health Economics*, 20(2), pp. 212–224. doi: 10.1002/hec.1586.

Grigore, B. *et al.* (2013) 'Methods to elicit probability distributions from experts: a systematic review of reported practice in health technology assessment', *Pharmacoeconomics*. Available at: http://link.springer.com/article/10.1007/s40273-013-0092-z (Accessed: 26 April 2017).

Grigore, B. *et al.* (2016) 'A comparison of two methods for expert elicitation in health technology assessments', *BMC Medical Research Methodology*, 16(1), p. 85. doi: 10.1186/s12874-016-0186-3.

Grigore, B., Peters, J. and Hyde, C. (2016) 'A comparison of two methods for expert

elicitation in health technology assessments', *BMC Medical*. Available at: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0186-3 (Accessed: 26 April 2017).

Haakma, W. *et al.* (2014) 'Belief Elicitation to Populate Health Economic Models of Medical Diagnostic Devices in Development', *Applied Health Economics and Health Policy*. Springer International Publishing, 12(3), pp. 327–334. doi: 10.1007/s40258-014-0092-y.

Hallenbeck, C. (1920) 'FORECASTING PRECIPITATION IN PERCENTAGES OF PROBABILITY.', *Monthly Weather Review*, 48(11), pp. 645–647. doi: 10.1175/1520-0493(1920)48<645:FPIPOP>2.0.CO;2.

Hammersley, J. S., Kadous, K. and Magro, A. M. (1997) 'Cognitive and Strategic Components of the Explanation Effect', *Organizational Behavior and Human Decision Processes*. Academic Press, 70(2), pp. 149–158. doi: 10.1006/OBHD.1997.2701.

Hammitt, J. K. and Zhang, Y. (2013) 'Combining Experts' Judgments: Comparison of Algorithmic Methods Using Synthetic Data', *Risk Analysis*, 33(1). doi: 10.1111/j.1539-6924.2012.01833.x.

Hartley, D. and French, S. (2017) 'Elicitation and Calibration', in Dias, L., Morton, A., and Quigley, J. (eds) *Elicitation. The Science and Art of Structuring Judgement.* Cham, Switzerland: Springer International Publishing.

Hijmans, J. M. *et al.* (2007) 'A systematic review of the effects of shoes and other ankle or foot appliances on balance in older people and people with peripheral nervous system disorders.', *Gait & posture*, 25(2), pp. 316–23. doi: 10.1016/j.gaitpost.2006.03.010.

Hinde, S. and Spackman, E. (2015) 'Bidirectional Citation Searching to Completion: An Exploration of Literature Searching Methods', *PharmacoEconomics*. Springer International Publishing, 33(1), pp. 5–11. doi: 10.1007/s40273-014-0205-3.

Hoffmann, T. C. and Del Mar, C. (2017) 'Clinicians' Expectations of the Benefits and Harms of Treatments, Screening, and Tests', *JAMA Internal Medicine*. American Medical Association, 177(3), p. 407. doi: 10.1001/jamainternmed.2016.8254.

Iglesias, C. P. *et al.* (2016) 'Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations', *PharmacoEconomics*. Springer International Publishing, 34(11), pp. 1161–1172. doi: 10.1007/s40273-016-0425-9.

Iglesias, C. P., Manca, A. and Torgerson, D. J. (2009) 'The health-related quality of life and

cost implications of falls in elderly women.', *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, 20(6), pp. 869–78. doi: 10.1007/s00198-008-0753-5.

Jenkinson, D. (2005) 'The Elicitation of Probabilities -A Review of the Statistical Literature'.

Johanesson, P. O. (1995) *Evaluating health risks*. Cambridge, UK: Cambridge University Press.

Kadane, J. B. (1986) 'Progress Toward a More Ethical Method for Clinical Trials', *Journal of Medicine and Philosophy*. Oxford University Press, 11(4), pp. 385–404. doi: 10.1093/jmp/11.4.385.

Kadane, J. and Wolfson, L. J. (1998) 'Experiences in elicitation [Read before The Royal Statistical Society at a meeting on "Elicitation" on Wednesday, April 16th, 1997, the President, Professor A. F. M. Smith in the Chair]', *Journal of the Royal Statistical Society: Series D (The Statistician)*. Blackwell Publishers Ltd, 47(1), pp. 3–19. doi: 10.1111/1467-9884.00113.

Kahneman, D., Slovic, P. and Tversky, A. (1982) *Judgment under uncertainty : heuristics and biases*. Cambridge University Press. Available at: http://www.cambridge.org/gb/academic/subjects/psychology/cognition/judgment-under-uncertainty-heuristics-and-biases?format=PB&isbn=9780521284141 (Accessed: 21 August 2018).

Kattan, M. W. *et al.* (2016) 'The Wisdom of Crowds of Doctors', *Medical Decision Making*. SAGE PublicationsSage CA: Los Angeles, CA, 36(4), pp. 536–540. doi: 10.1177/0272989X15581615.

Kempen, G. I. J. M. *et al.* (2007) 'The Short FES-I: a shortened version of the falls efficacy scale-international to assess fear of falling', *Age and Ageing*, 37(1), pp. 45–50. doi: 10.1093/ageing/afm157.

Klarman, H. E., Francis, J. O. and Rosenthal, G. D. (1968) 'Cost Effectiveness Analysis Applied to the Treatment of Chronic Renal Disease', *Medical Care*. Lippincott Williams & Wilkins, 6(1), pp. 48–54. doi: 10.2307/3762651.

Kleinmuntz, D. N., Fennema, M. G. and Peecher, M. E. (1996) 'Conditioned Assessment of Subjective Probabilities: Identifying the Benefits of Decomposition', *Organizational*

Behavior and Human Decision Processes, 66(1), pp. 1–15. doi: 10.1006/obhd.1996.0033.

Knol, A. B. *et al.* (2010) 'The use of expert elicitation in environmental health impact assessment: a seven step procedure.', *Environmental Health*, pp. 9–19.

Kobayashi, R. *et al.* (1999) 'Effects of Toe Grasp Training for the Aged on Spontaneous Postural Sway.', *Journal of Physical Therapy Science*, 11(1), pp. 31–34. doi: 10.1589/jpts.11.31.

Koehler, J. J. (2001a) 'The Psychology of Numbers in the Courtroom: How to Make DNA Match Statistics Seem Impressive or Insufficient'. Available at: http://papers.ssrn.com/abstract=1432071 (Accessed: 11 March 2016).

Koehler, J. J. (2001b) 'When are people persuaded by DNA match statistics?', *Law and Human Behaviour*, 25, pp. 493–513.

Koepsell, T. D. *et al.* (2004) 'Footwear style and risk of falls in older adults.', *Journal of the American Geriatrics Society*, 52(9), pp. 1495–501. doi: 10.1111/j.1532-5415.2004.52412.x.

Kullback, S. and Leibler, R. A. (1951) 'On Information and Sufficiency', *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 22(1), pp. 79–86. doi: 10.1214/aoms/1177729694.

Lachman, M. E. *et al.* (1998) 'Fear of Falling and Activity Restriction: The Survey of Activities and Fear of Falling in the Elderly (SAFE)', *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*. Oxford University Press, 53B(1), pp. P43–P50. doi: 10.1093/geronb/53B.1.P43.

Leal, J. *et al.* (2007) 'Eliciting Expert Opinion for Economic Models: An Applied Example', *Value in Health*, 10(3), pp. 195–203. doi: 10.1111/j.1524-4733.2007.00169.x.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982) 'Calibration of probabilities: the state of the art to 1980', in Kahneman, D., Slovic, P., and Tversky, A. (eds) *Judgement under UNcertainty: Heuristics and Biases*. Cambridge: Cambridge University Press, pp. 306–334.

Lin, S. and Cheng, C. (2009) 'The reliability of aggregated probability judgments obtained through Cooke's classical model.', *Journal of Modelling and Management*, 4(2), pp. 149–161.

Lin, S. and Cheng, C. (2009a) 'The reliability of aggregated probability judgments obtained through Cooke's classical model', *Journal of Modelling in Management*. Emerald Group

Publishing Limited, 4(2), pp. 149–161. doi: 10.1108/17465660910973961.

Lin, S. and Cheng, C. (2009b) 'The reliability of aggregated probability judgments obtained through Cooke's classical model', *Journal of Modelling in Management*. Emerald Group Publishing Limited, 4(2), pp. 149–161. doi: 10.1108/17465660910973961.

Lord, S. R. *et al.* (2007) *Falls in Older People: Risk Factors and Strategies for Prevention*. Cambridge University Press. Available at: https://books.google.com/books?hl=en&lr=&id=1enrvVe81YgC&pgis=1 (Accessed: 12 February 2016).

Mahon, R. (2014) 'TEMPORAL UNCERTAINTY IN COST-EFFECTIVENESS DECISION MODELS:METHODS TO ADDRESS THE UNCERTAINTIES THAT ARISE WHEN THE APPROPRIATE ANALYSIS TIME HORIZON EXCEEDS THE EVIDENCE TIME HORIZON IN COST-EFFECTIVENESS DECISION MODELS AS APPLIED TO HEALTHCARE INTE'. University of York. Available at: http://etheses.whiterose.ac.uk/8268/1/Ronan Mahon Thesis %28Feb 2015%29.pdf (Accessed: 11 March 2016).

Mckenna, C. *et al.* (2009) 'Health Technology Assessment NIHR HTA programme www.hta.ac.uk Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis', *Health Technology Assessment*, 13(24). doi: 10.3310/hta13240.

Menant, J. C. *et al.* (2008) 'Effects of footwear features on balance and stepping in older people.', *Gerontology*. Karger Publishers, 54(1), pp. 18–23. doi: 10.1159/000115850.

Menant, J. C. *et al.* (2009) 'Rapid gait termination: effects of age, walking surfaces and footwear characteristics.', *Gait & posture*, 30(1), pp. 65–70. doi: 10.1016/j.gaitpost.2009.03.003.

Menz, H. B., Morris, M. E. and Lord, S. R. (2005) 'Foot and Ankle Characteristics Associated With Impaired Balance and Functional Ability in Older People', *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*. Oxford University Press, 60(12), pp. 1546–1552. doi: 10.1093/gerona/60.12.1546.

Menz, H. B., Morris, M. E. and Lord, S. R. (2006) 'Foot and ankle risk factors for falls in older people: a prospective study.', *The journals of gerontology. Series A, Biological sciences and medical sciences*. Oxford University Press, 61(8), pp. 866–70. Available at: http://biomedgerontology.oxfordjournals.org/content/61/8/866.full (Accessed: 12

February 2016).

Meyer, M. A. and Booker, J. M. (1991) *Eliciting and Analysing Expert Judgment: A Practical Guide*. London: Academic Press.

Ministry of Health (1994) *Ontario guidelines for economic analysis of pharmaceutical products.* Ontario.

Moatti, M. *et al.* (2013) 'Modeling of experts' divergent prior beliefs for a sequential phase III clinical trial', *Clinical Trials: Journal of the Society for Clinical Trials*. SAGE PublicationsSage UK: London, England, 10(4), pp. 505–514. doi: 10.1177/1740774513493528.

Montibeller, G. and von Winterfeldt, D. (2015) 'Cognitive and Motivational Biases in Decision and Risk Analysis', *Risk Analysis*, 35(7), pp. 1230–1251. doi: 10.1111/risa.12360.

Morgan, M. G. (2014) 'Use (and abuse) of expert elicitation in support of decision making for public policy.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(20), pp. 7176–84. doi: 10.1073/pnas.1319946111.

Morris, D. E., Oakley, J. E. and Crowe, J. A. (2014) 'A web-based tool for eliciting probability distributions from experts', *Environmental Modelling & Software*, 52, pp. 1–4. doi: 10.1016/j.envsoft.2013.10.010.

Mullen, P. M. (2003) 'Delphi: myths and reality', *Journal of Health Organization and Management*, 17(1), pp. 37–52. doi: 10.1108/14777260310469319.

Murphy, A. (1970) 'The ranked probability score and the probability score: A comparison', *Monthly Weather Review*, 98, pp. 917–24.

Murphy, A. (1971) 'A note on the ranked probability score', *Journal of Applied Meteorology1*, 10, pp. 155–6.

Murphy, A. H. and Murphy, A. H. (1973) 'A New Vector Partition of the Probability Score', *Journal of Applied Meteorology*, 12(4), pp. 595–600. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, A. H. and Winkler, R. L. (1977) 'Reliability of Subjective Probability Forecasts of Precipitation and Temperature', *Applied Statistics*, 26(1), p. 41. doi: 10.2307/2346866.

National Institute for Health and Care Excellence (NICE) (2011) *Diagnostics Assessment*

Programme manual. Manchester, IK. Available at: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-diagnostics-guidance/Diagnostics-assessment-programme-manual.pdf (Accessed: 15 September 2018).

National Institute for Health and Care Excellence (NICE) (2013) *Guide to Methods of Technology Appraisal*.

Nemet, G. F. Q. the E. of E. S. and E. D. on E. C. in T. J. A. F. E. T., Anadon, L. D. and Verdolini, E. (2017) 'Quantifying the Effects of Expert Selection and Elicitation Design on Experts' Confidence in Their Judgments About Future Energy Technologies', *Risk Analysis*, 37(2), pp. 315–330. doi: 10.1111/risa.12604.

Nevitt, M. C. (1989) 'Risk Factors for Recurrent Nonsyncopal Falls', *JAMA*. American Medical Association, 261(18), p. 2663. doi: 10.1001/jama.1989.03420180087036.

NHS England (2014) *Five Year Forward View*. Available at: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf (Accessed: 31 January 2018).

Nuffield Trust (2014) *Health spending per head by country of the UK*, *Health spending per head by country of the UK*. Available at: https://www.nuffieldtrust.org.uk/chart/health-spending-per-head-by-country-of-the-uk.

O'Hagan, A. *et al.* (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons. Available at: https://books.google.com/books?hl=en&lr=&id=H9KswqPWIDQC&pgis=1 (Accessed: 15 February 2016).

O'Hagan, A. *et al.* (2009) *Elicitation of IndividualsíKnowledge in Probabilistic Form*.

Pauly, M. V (1995) 'Valuing Health Benefits in Monetary Term', in Sloan, F. A. (ed.) *Valuing health care: costs, benefits and effectiveness of pharmaceurical and other medical technologies*. Cambridge, UK: Cambridge University Press.

Peterson, C. and Miller, A. (1964) 'Mode, median and mean as optimal strategies', *Journal of Experimental Psychology*, 68(4), pp. 363–367.

Philips, Z. *et al.* (2006) 'Good Practice Guidelines for Decision-Analytic Modelling in Health Technology Assessment', *PharmacoEconomics*, 24(4), pp. 355–371. doi: 10.2165/00019053-200624040-00006.

Quigley, J. *et al.* (2017) 'Elicitation in the Classical Model', in *Elicitation. The Science and Art of Structuring Judgement.2*. Cham, Switzerland: Springer International Publishing, pp. 15–37.

Rakow, T. *et al.* (2005) 'Assessing the Likelihood of an Important Clinical Outcome: New Insights from a Comparison of Clinical and Actuarial Judgment', *Medical Decision Making*. Sage PublicationsSage CA: Thousand Oaks, CA, 25(3), pp. 262–282. doi: 10.1177/0272989X05276849.

Rohrbaugh, J. (1981) 'Improving the quality of group judgment: Social judgment analysis and the nominal group technique', *Organizational Behavior and Human Performance*. Academic Press, 28(2), pp. 272–288. doi: 10.1016/0030-5073(81)90025-8.

Sabin, M. (2001) *Competence in Practice-Based Calculation: Issues for Nursing Education*. Edinburgh: Napier University. Available at: http://s3.amazonaws.com/academia.edu.documents/44227583/Competence_in_Practice-Based_Calculation20160330-18648-116ewd3.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1493741713&Signature=VS3gSZHhRYG22l3B2L9tHtW9tc4%253D&response-content-disposition=inline%25 (Accessed: 2 May 2017).

Sculpher, M. J. *et al.* (2006) 'Whither trial-based economic evaluation for health care decision making?', *Health economics*, 15(7), pp. 677–87. doi: 10.1002/hec.1093.

Sedgwick, P. (2012) 'Pearson's correlation coefficient', *BMJ*. British Medical Journal Publishing Group, 345(jul04 1), pp. e4483–e4483. doi: 10.1136/bmj.e4483.

Shabaruddin, F. H. *et al.* (2010) 'Understanding chemotherapy treatment pathways of advanced colorectal cancer patients to inform an economic evaluation in the United Kingdom', *British Journal of Cancer*. Nature Publishing Group, 103(3), pp. 315–323. doi: 10.1038/sj.bjc.6605766.

Shanteau, J. (1992) 'Competence in experts: The role of task characteristics', *Organizational Behavior and Human Decision Processes*. Academic Press, 53(2), pp. 252–266. doi: 10.1016/0749-5978(92)90064-E.

Shi-Woei Lin and Chih-Hsing Cheng (2008) 'Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities?', in *2008 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, pp. 425–429. doi:

10.1109/IEEM.2008.4737904.

Siegel-Jacobs, K. and Yates, J. F. (1996) 'Effects of Procedural and Outcome Accountability on Judgment Quality', *Organizational Behavior and Human Decision Processes*, 65(1), pp. 1–17. doi: 10.1006/obhd.1996.0001.

Silverman, W. A. (1980) *Retrolental Fibroplasia: A Modern Parable*. New York: Grune & Stratton, Inc.

Slovic, P., Monahan, J. and MacGregor, D. G. (2000) 'Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats.', *Law and HUman Behaviour*, 24, pp. 271–296.

Sniezek, J. A. (1992) 'Groups under uncertainty: An examination of confidence in group decision making', *Organizational Behavior and Human Decision Processes*. Academic Press, 52(1), pp. 124–155. doi: 10.1016/0749-5978(92)90048-C.

Soares, M. *et al.* (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*. John Wiley & Sons, Ltd, 30(19), pp. 2363–2380. doi: 10.1002/sim.4288.

Soares, M. O. *et al.* (2013) 'Methods to assess cost-effectiveness and value of further research when data are sparse: negative-pressure wound therapy for severe pressure ulcers.', *Medical decision making : an international journal of the Society for Medical Decision Making*, 33(3). doi: 10.1177/0272989X12451058.

Soares, M. O. *et al.* (2018) 'Experiences of Structured Elicitation for Model-Based Cost-Effectiveness Analyses.', *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. Elsevier, 21(6), pp. 715–723. doi: 10.1016/j.jval.2018.01.019.

Soares, M. O., Dumville, J. C. and Ashby, R. L. (2013) 'Methods to assess cost-effectiveness and value of further research when data are sparse: negative-pressure wound therapy for severe pressure ulcers', *Medical Decision*. Available at: http://journals.sagepub.com/doi/abs/10.1177/0272989X12451058 (Accessed: 26 April 2017).

Sorenson, C., Drummond, M. and Bhuiyan Khan, B. (2013) 'Medical technology as a key driver of rising health expenditure: disentangling the relationship.', *ClinicoEconomics and*

outcomes research : CEOR. Dove Press, 5, pp. 223–34. doi: 10.2147/CEOR.S39634.

Speight, P. *et al.* (2006) 'The cost-effectiveness of screening for oral cancer in primary care.', *Health Technol Assessment*, 10(14), pp. 1–144.

Sperber, D. *et al.* (2013) 'An expert on every street corner? Methods for eliciting distributions in geographically dispersed opinion pools', *Value in Health*. Available at: http://www.sciencedirect.com/science/article/pii/S1098301512041630 (Accessed: 26 April 2017).

Sperber, D. *et al.* (2013) 'An Expert on Every Street Corner? Methods for Eliciting Distributions in Geographically Dispersed Opinion Pools', *Value in Health*, 16(2), pp. 434–437. doi: 10.1016/j.jval.2012.10.011.

Spink, M. J. *et al.* (2011) 'Effectiveness of a multifaceted podiatry intervention to prevent falls in community dwelling older people with disabling foot pain: randomised controlled trial.', *BMJ (Clinical research ed.)*, 342(jun16_3), p. d3411. doi: 10.1136/bmj.d3411.

Spink, M. J., Menz, H. B. and Lord, S. R. (2008) 'Efficacy of a multifaceted podiatry intervention to improve balance and prevent falls in older people: study protocol for a randomised trial.', *BMC geriatrics*. BioMed Central, 8(1), p. 30. doi: 10.1186/1471-2318-8-30.

Stern, P. C. and Fineberg, H. V (1996) *Understanding risk: informing decisions in a democratic society*. Washington: National Academies Press. Available at: https://books.google.com/books?hl=en&lr=&id=TGydAgAAQBAJ&pgis=1 (Accessed: 11 March 2016).

Stevenson, M. D. *et al.* (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*. SAGE PublicationsSage CA: Los Angeles, CA, 29(6), pp. 678–689. doi: 10.1177/0272989X09336077.

Stone, M. (1961) 'The opinion pool', *The annals of Mathematical Statistics*, 32, pp. 1339–42.

Svedbom, A. *et al.* (2018) 'Quality of life after hip, vertebral, and distal forearm fragility fractures measured using the EQ-5D-3L, EQ-VAS, and time-trade-off: results from the ICUROS', *Quality of Life Research*, 27(3), pp. 707–716. doi: 10.1007/s11136-017-1748-5.

Tetlock, P. E. (Philip E., Gardner, D. and Richards, J. (2016) *Superforecasting : the art and science of prediction*. New York: Crown Publishing Group.

Tinetti, M. E. and Speechley, M. (1989) 'Prevention of falls among the elderly.', *The New England journal of medicine*, 320(16), pp. 1055–9. doi: 10.1056/NEJM198904203201606.

Tinetti, M. E., Speechley, M. and Ginter, S. F. (1988) 'Risk Factors for Falls among Elderly Persons Living in the Community — NEJM', *N Engl J Med*, 319, pp. 1701–1707. Available at: http://www.nejm.org/doi/full/10.1056/NEJM198812293192604 (Accessed: 12 February 2016).

Torgerson, D. (2001) 'The economics of fracture prevention', *The effective …*. Available at: https://scholar.google.co.uk/scholar?hl=en&as_sdt=0,5&q=torgerson+iglesias+the+economics+in+fracture+prevention#0 (Accessed: 12 February 2016).

Wallsten, T. S. and Budescu, D. V. (1983) 'State of the Art—Encoding Subjective Probabilities: A Psychological and Psychometric Review', *Management Science*. INFORMS , 29(2), pp. 151–173. doi: 10.1287/mnsc.29.2.151.

Watson, G. and Glaser, E. M. (2010) 'Watson-Glaser || Critical Thinking Appraisal'.

WHO (2015) *WHO | HTA Definitions*, *WHO*. World Health Organization. Available at: http://www.who.int/health-technology-assessment/about/Defining/en/ (Accessed: 4 June 2018).

Winkler, R. L. and Poses, R. M. (1993) 'Evaluating and Combining Physicians' Probabilities of Survival in an Intensive Care Unit', *Management Science*. INFORMS , 39(12), pp. 1526–1543. doi: 10.1287/mnsc.39.12.1526.

World Health Organization (2017) *Global Health Expenditure Database*. Available at: http://apps.who.int/nha/database/Select/Indicators/en (Accessed: 31 January 2018).

Yates and F., J. (1994) 'Subjective probability accuracy analysis', *Subjective probability*. Wiley, pp. 381–410. Available at: http://ci.nii.ac.jp/naid/10009704471/ (Accessed: 2 July 2017).

Yesavage, J. A., Brink, T. L. and Rose, T. L. (1982) 'Development and validation of a geriatric depression screening scale: a preliminary report', *Journal of Psychiatric Research*, 17(1), pp. 37–49.

Appendix 3.1. REFORM trial background information

Background

Falls are a major cause of morbidity and mortality in the elderly population in the UK (Torgerson, 2001; Iglesias, Manca and Torgerson, 2009) with an associated cost burden of £1.8 billion per year (Torgerson, 2001). Falls result from a combination of environmental and physiological factors (Lord *et al.*, 2007). Several studies suggest that footwear, foot pain and foot and ankle strength can affect balance and the risk of falls (Koepsell *et al.*, 2004; Menz, Morris and Lord, 2005, 2006, Menant *et al.*, 2008, 2009). Several studies have suggested that some treatments provided by podiatrists may play a role in improving balance (Kobayashi *et al.*, 1999; Balanowski and Flynn, 2005; Hijmans *et al.*, 2007), but none of these studies measured their effect on the frequency and severity of falls. Furthermore, previous studies addressed individual risk factors (footwear, foot pain or foot strength alone). Only one randomised controlled trial (RCT) has assessed the clinical effectiveness of a multifaceted podiatry intervention aiming to reduce the risk of falls by addressing all of the above mentioned risk factors (Spink, Menz and Lord, 2008; Spink *et al.*, 2011). The study was conducted in Australia and did not measure the costs of the intervention.

The REFORM trial aimed to evaluate the clinical and cost-effectiveness of a similar, multifaceted podiatry intervention, in a UK setting.

Intervention

The intervention consisted of the following four components:

- foot orthoses (insole designed to reduce pain by redistributing pressure away from foot lesions),
- a home based exercise programme aiming to stretch and strengthen the muscles of the foot and ankle. They were demonstrated by the podiatrist at the participant's initial visit and were supplemented by a DVD demonstrating the exercises and an illustrated explanatory booklet showing how to do them at home.
- Footwear assessment was carried out, and where participants do not have appropriate footwear they were provided with footwear vouchers and advice on optimal footwear.
- Falls prevention advice leaflet was sent to participants.

Outcome measures

The primary outcome measurement in the trial was the rate of falls, defined as the average number of falls patients suffer per year. Secondary outcome measurements were:

a) Proportion of patients who suffer at least one fall
b) Proportion of patients who suffer multiple falls (more than two)
c) Patient reported time to first fall during follow-up
d) Health related quality of life as measured by the EQ-5D
e) Short Falls Efficacy Scale
f) Fear of falling
g) Activity of Daily Living
h) Fracture rate
i) Health service utilisation
j) Geriatric Depressions Scale

These were all considered as potential target parameters in section 3.5.1.

Patient population

The target sample size in the trial was 890 (445 in each arm).

Participants were included in the trial if:

- they were 70 years of age and over,
- they were community dwelling,
- they have had at least one fall in the past 12 months; or one fall in the past 24 months requiring hospital attention.

Participants were excluded from the study if:

- They were known to have neuropathy.
- They were known to have a neurodegenerative disorder.
- They failed to return all monthly falls diaries over the first three month period (the pilot) or failed to return the baseline questionnaires.
- They had had a lower limb amputation (including partial foot amputation)
- They were unable to walk household distances (10 metres/32 feet) unaided.
- At recruitment they were wearing a full or 3/4 length in-shoe foot orthotic with the purpose of altering or modifying foot function in order to treat, adjust, and support various biomechanical foot disorders.

- They were known to have dementia.
- They were unable to read or speak English.
- Their usual footwear had been adapted in such a way which would not have allowed an orthotic to be fitted.

Appendix 3.2.Questions about experts' substantive expertise

Please select your role from the list by ticking the appropriate box.

□ Podiatrist

□ Occupational therapist

□ Physiotherapist

□ Orthopaedic surgeon

□ Geriatrician

□ Academia

What is your job title? Please include as much detail as possible, such as specialty, level or band.

_____

How many years have you been in this role? Please include experience at other locations.

_____

What proportion of your time do you spend working with patients at risk of falling, either helping them prevent falls or treating fall related injuries?

□ 0-10%

□ 11-30%

□ 31-50%

□ More than 50%

Are you aware of any ongoing or published research on podiatry interventions designed to reduce the risk of falls?

□ Yes

□ No

Have you been a co-author on any published research? If so, how many publications do you have? (Please note these do not have to be on a topic related to falls.)

□ Less than 3

□ 4-20

□ 21-50

□ More than 50

Have you ever been involved in writing a successful research grant proposal? If so, how many?

□ 0

□ 1-5

□ More than 5

Appendix 3.3. The non-domain seed elicited to capture experts' ability to make accurate probabilistic assessments

Figure 3.1. Question used to capture experts' normative skills.

General question

As part of our exercise we need to ask a general question, unrelated to falls. Please read through and answer carefully.

Out of 30 days in September, how many days does it rain in York, on average?

I think on average, in York, it rains at least

1

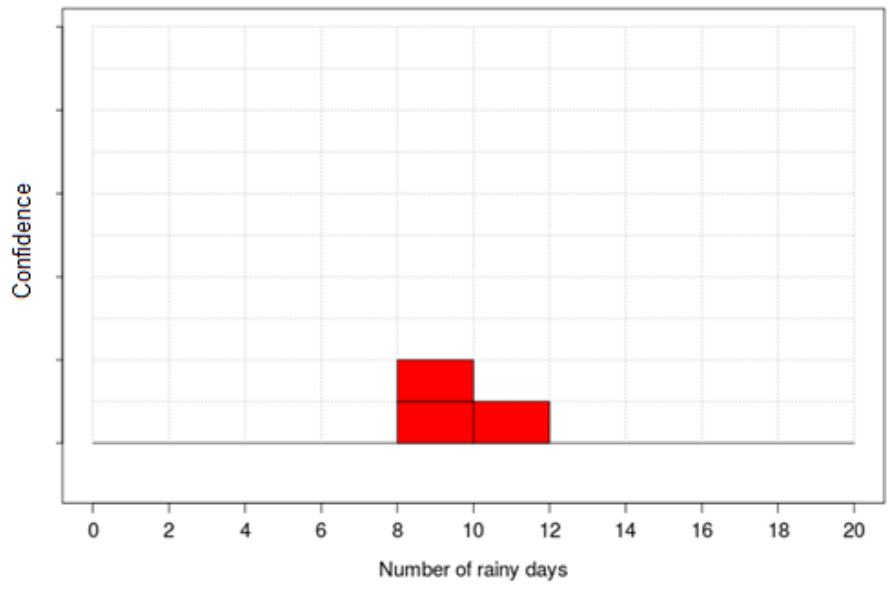
out of 30 days.
I think on average, in York, it rains no more than

17

out of 30 days in September.

Enter

Please use the grid below to indicate how certain you are about your answer. If you are unsure how to do this please refer to the 'Instructions' tab. When you are happy with your answer, click on 'Save' and scroll down for further instructions.



Save

Appendix 3.4. Questions assessing experts' inference skills

The following question was taken from a standard tool (called Watson-Glacer appraisal) used in research to gain understanding of how people process information. Please read the paragraph below to answer.

'Studies have shown that heart disease is more common among people living in the north of England than people living in the south of England. There is little if any difference, however, in rate of heart disease between northerners and southerners who have the same level of income.

'The average income of southerners in England is considerably higher than the average income of northerners.

Let's assume that the information in the paragraph is correct. The following four questions contain conclusions that some people might draw from the paragraph. For each statement please chose ONE available option that BEST describes its degree of truth or falsity, based on the paragraph above.

a) The easiest way to eliminate heart disease in England would be to raise the general standard of living.
   □This is true
   □This is probably true
   □There is insufficient information
   □This is false
   □This is probably false

b) People in high income brackets are in a better position to avoid developing heart disease than people in low income brackets.
   □This is true
   □This is probably true
   □There is insufficient information
   □This is false
   □This is probably false

c) There is a lower rate of heart disease among northerners with relatively high incomes than among northerners with much lower incomes.
   □This is true
   □This is probably true

□There is insufficient information

□This is false

□This is probably false

d) Whether northerners have high incomes or low incomes makes no difference to the likelihood of their developing heart disease.

□This is true

□This is probably true

□There is insufficient information

□This is false

□This is probably false

Appendix 3.5. Targeted search for methods for eliciting rates

Aim

As highlighted in section 3.5.2, a pragmatic, non-systematic search was conducted to identify existing methods for eliciting rates.

Methods

Two search strategies were employed:

- Searching elicitation literature identified and discussed in Chapter 1,
- Conducting a targeted database search.

The targeted database search was conducted in the three search databases recommended by the University of York: Google Scholar, Web of Knowledge, and Scopus. Details of the searches are provided in Table A3.1. Additional search terms were considered, such as 'expert' or 'expert knowledge'. An informal scoping search suggested that the additional search terms led to more references that were less relevant.

Table A3.1. Search terms used in the targeted search.

Datebase	Search terms (anywhere in text)	Additional restrictions
Google Scholar	(elicit OR eliciting OR elicitation) AND rate	English language
Web of Knowledge	elicit* AND rate	English language
Scopus	elicit* AND rate	English language

The identified references were scanned for relevance in three stages: 1) by title, 2) abstract, and 3) full text.

Results

The Google Scholar, Web of Knowledge and Scopus database searches returned 827,000, 22,275 and 32,397 results, respectively. Careful consideration of each individual reference was not feasible, instead, the first 300 results were reviewed in each.

The results of the searches are shown in Table A3.2.

Table A3.2. Results of the scoping search.

Search method		Number of citations identified at each stage of the search			
		Total	Shortlisted by title	Shortlisted by abstract	Shortlisted by full test
Citations used in Chapter 1		48	12	0	0
Databased searches	Google Scholar	300	12	2	0
	Web of Knowledge	300	7	0	0
	Scopus	300	5	0	0

No studies describing methods for eliciting rates were identified.

Appendix 3.6. Background information about REFORM trial

About the trial

It is a randomised controlled trial (RCT) aiming to evaluate the clinical and cost effectiveness of a multi-component podiatry intervention for the prevention of falls in patients over the age of 70.

Why?

There is evidence that foot and ankle problems are associated with increased risk of falls. To read the summary of evidence click here.

A recent prospective study of 176 older people indicated that **ankle flexibility, toe plantarflexor strength and plantar sensation** were significant and independent predictors of balance and functional test performance.

A 12-month follow-up of this cohort confirmed that these factors, in addition to foot pain, were significant independent predictors of falls.

A number of studies have assessed footwear in older people who have fallen and suggest that **walking barefoot or wearing stockings, increased shoe heel height and smaller sole contact area** can all increase the risk of a fall.

A number of other studies have investigated the main features of a shoe thought to affect balance, and found that **increased heel height and reduced sole hardness** have detrimental effects on balance.

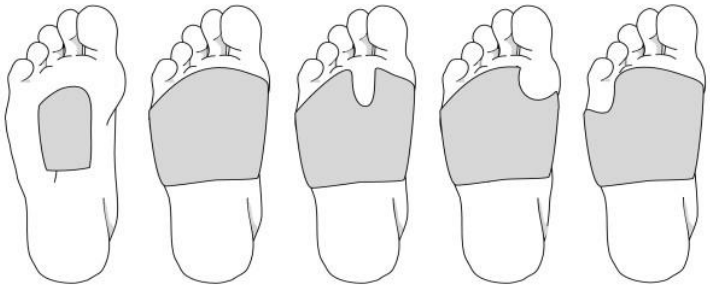
Several studies have suggested that some treatments provided by podiatrists, such as **lesion debridement, foot orthoses and foot and ankle exercises**, may play a role in improving balance.

Given the emerging evidence that foot problems and inappropriate footwear increase the risk of falls, it has been suggested that podiatry may have a role to play in falls prevention.

About the intervention

The intervention aims to address all foot and ankle related risk factors for falling. It consists of four components: foot orthoses, foot and ankle exercise programme, footwear assessment and shoe provision and a fall prevention leaflet.

- Foot orthoses are insole designed to redistribute pressure away from plantar lesions.



To read more about the orthoses click here (FOOT ORTHOSES).

- The exercise programme aiming to stretch and strengthen the muscles of the foot and ankle involves 30 minute home based exercise to be undertaken three times per week indefinitely. The exercises will be demonstrated by the podiatrist at the participant's initial visit and will be supplemented by a DVD demonstrating the exercises and an illustrated explanatory booklet showing how to do them at home. To view the details of the exercise programme click here (EXERCISE).
- Footwear assessment will be carried out. Participants who do not have appropriate footwear will be provided with footwear vouchers and advice on optimal footwear.
- Falls prevention advice leaflet ("Staying steady. Improving your strength and balance" designed by AgeUK) will be sent to participants.

Foot orthoses

Participants will be fitted with a prefabricated insole (Formthotics ™ Foot Science) manufactured from a thermoformable cross-linked closed cell polyethylene foam which will be shaped to fit the participant's foot. The orthoses will then be appropriately customised using 3mm thick PPT urethane to redistribute pressure away from any plantar lesions. The orthosis will be supplied either by the podiatrist delivering the intervention or by a manufacturer in response to a prescription from the podiatrist.

Exercise

Activity	Description	Dosage	Increments
Ankle range of motion	Sitting with knee extended. Rotate foot in clockwise direction and then anti-clockwise.	1x10 repetitions for each foot in each direction.	None.
Ankle inversion strength	Sitting, hip and ankle at 90º. Invert foot against resistive exercise band anchored by chair leg.	3x10 repetitions for each foot.	Increase resistance strength of resistive exercise band.
Ankle eversion strength	Sitting, hip and ankle at 90º. Evert foot against resistive exercise band anchored by chair leg.	3x10 repetitions for each foot.	Increase resistance strength of resistive exercise band.
Ankle dorsiflexion strength	Sitting, hip and ankle at 90º. Dorsiflex both feet to end range of motion and hold.	Hold feet in dorsiflexion for 3x10 seconds.	Increase repetitions up to maximum of 10.
Adductor hallucis stretch	Elastic band around both halluces. Move feet apart.	2x20 seconds.	None.
Toe plantarflexion strength	Place heel on plate of Archxerciser™. Place toes over spring loaded toebar. Retract bar with toes.	3x10 repetitions for each foot.	Increase distance bar is retracted.
Toe plantarflexion strength	Pick up 25mm diameter stones and place in box.	Pick up 2x20 stones for each foot.	None.

Ankle plantarflexion strength	From standing, rise up onto toes of both feet and then lower back down.	3x10 repetitions.	Increase repetitions up to maximum of 50.
Calf stretch	Standing stretch leaning against wall. Stretch knee is extended. Place support leg forward with knee flexed.	Hold stretch for 3x20 seconds on each leg.	Increase forward lean to increase stretch as required.

Data collection

They will then complete a questionnaire at 1, 3, 6 and 12 months after randomisation to collect data on compliance with the exercise programme, wearing the foot orthoses and the number of falls they had. They will keep monthly exercise and fall calendars to help them complete the questionnaire.

Follow-up questionnaires will be sent to participants in the post. Participants who provide an email address will be given the opportunity, if they prefer, to complete the questionnaire on-line.

Who will take part?

Trial participants be

- 70 years of age and over
- Community dwelling
- have had at least one fall in the past 12 months; or one fall in the past 24 months requiring hospital attention.

Participants will be excluded if:

- They are known to have neuropathy.
- They are known to have a neurodegenerative disorder.
- They fail to return all monthly falls diaries over the first three month period (the pilot) or fail to return the baseline questionnaires.
- They have had a lower limb amputation (including partial foot amputation)
- They are unable to walk household distances (10 metres/32 feet) unaided.

- They are currently wearing a full or 3/4 length in-shoe foot orthotic with the purpose of altering or modifying foot function in order to treat, adjust, and support various biomechanical foot disorders.
- They are known to have dementia.
- They are unable to read or speak English.
- They would be excluded if their usual footwear has been adapted in such a way which would not allow an orthotic to be fitted.

Appendix 3.7. Histogram technique training (Instructions tab from Figure 3.6)

How can I express my uncertainty in answers?

In order to express your uncertainty, first you will be asked to give a range of possible values: the lowest possible value, and the highest possible value. For example, what do you believe to be the average likelihood that an individuals over 50 will have a heart attack?

Let's say that the likelihood is at least 2%, and no more than 20%. We type these values in the boxes below and click on 'Enter'.

The likelihood of heart attack in people over 50 is at least

2

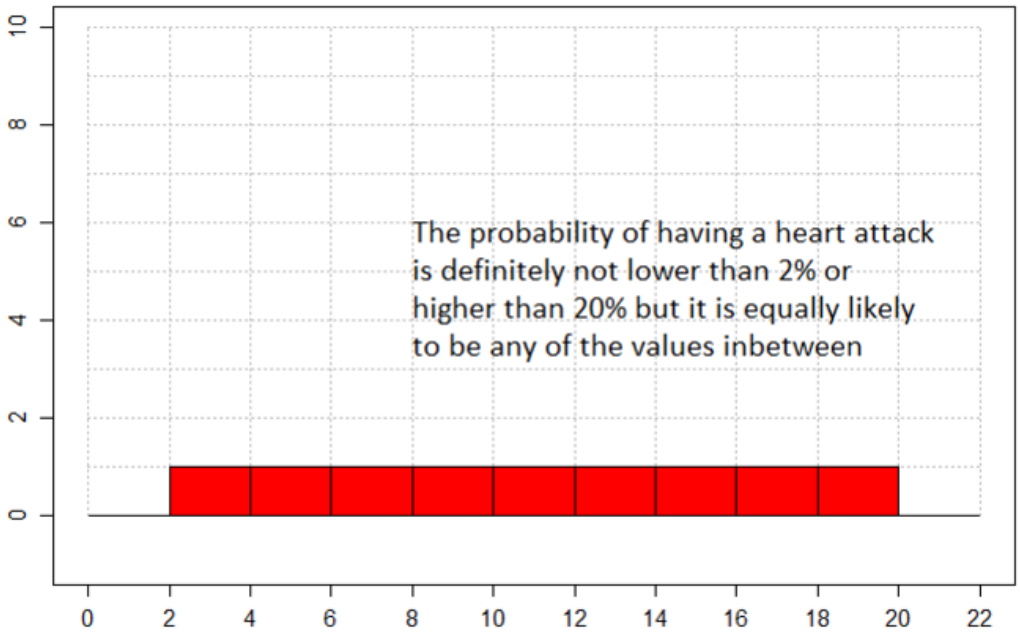
The likelihood of heart attack in people over 50 is at most

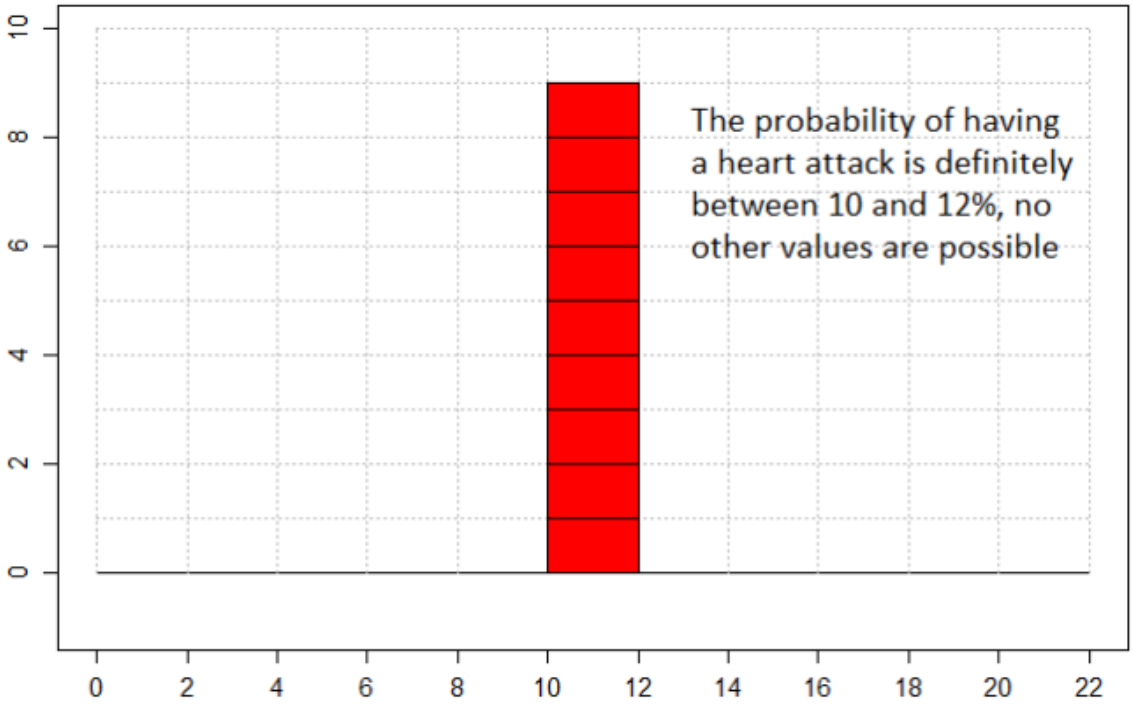
20

Enter

Then, you will be required to fill in a grid. The range of possible values (as suggested by you) will be given along the bottom of the grid. If you click on cells, that column will turn red. Use this to indicate how confident you feel that each of the values on the horizontal axis is the correct answer. The higher the red bar in the grid, the more confident you are.

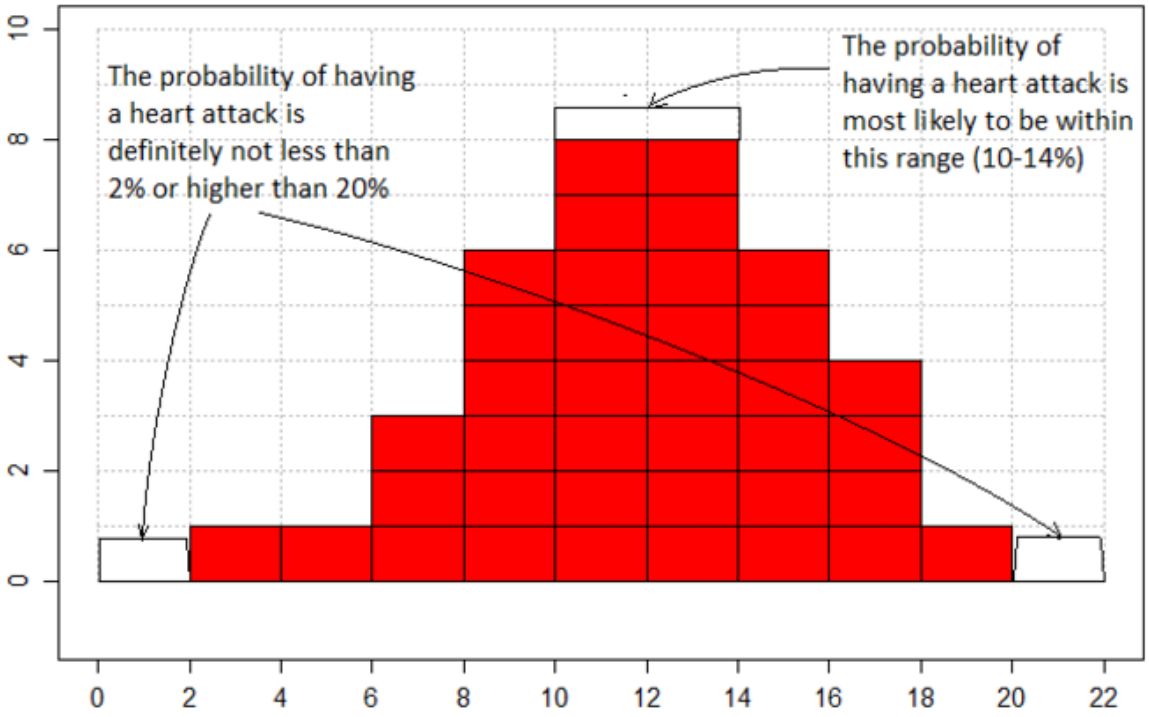
For example, if you are completely uncertain, you can use the same height of red bars across all possible values (the entire axis), as shown below.





You will need to be mindful about what your response says about the degree of certainty in your belief. For example, twice as many coloured boxes on one value means you believe that the value is twice as likely to be the answer.

Another example is shown below.



Now let's do a practice one.

Your practice question is: what do **you** believe is the average risk of heart attack in individuals over 50 years?

Note that the range on the horizontal axis of the plot will always be slightly wider than the minimum and maximum you specified. So, if you think the range is between 2 and 20, the grid will have a slightly wider range: 0 to 22. If, on second thought, you feel that it is possible that the risk is between 0 and 2, or between 20 and 22, it is possible to update your range by typing in the new values for minimum and maximum, and clicking 'Enter' again. The grid co-ordinates will update automatically.

You can change the height of each column at any point by clicking above or below the current height. If you want to erase all colour you can click on the plot, just below the horizontal axis.

You can clear the grid entirely by clicking on 'Enter' button above the grid.

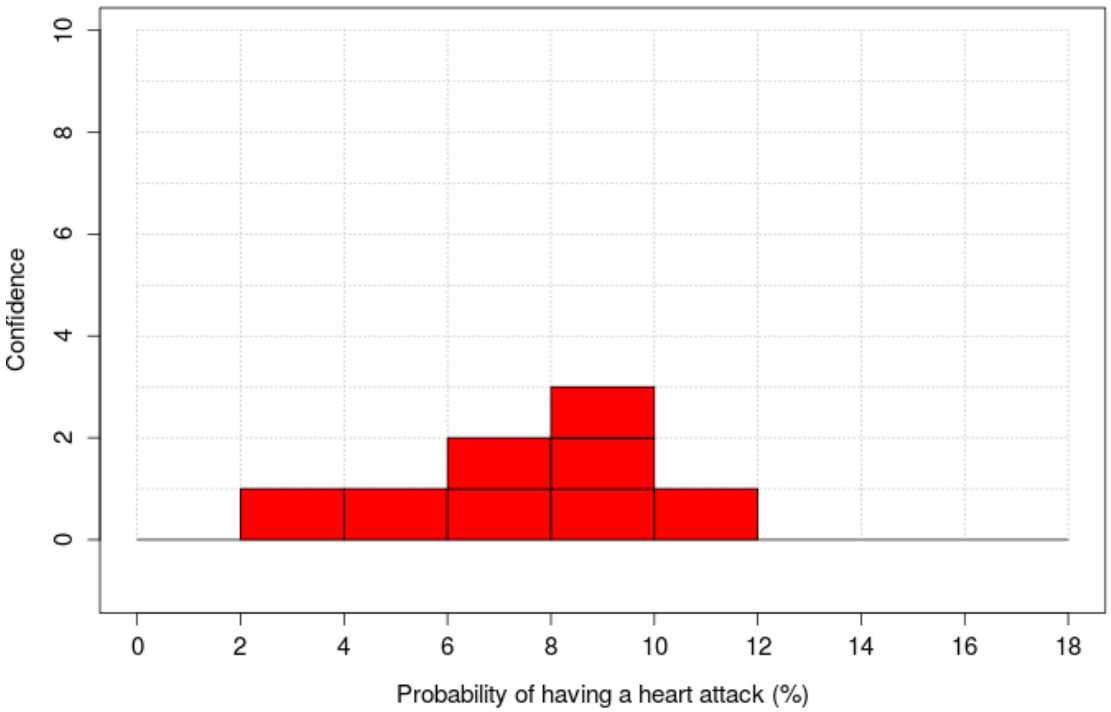
The likelihood of heart attack in people over 50 is at least

1

The likelihood of heart attack in people over 50 is at most

15

Enter



When you are satisfied that the grid expresses your opinion well, you can save it by clicking on the 'Save' button below and scroll down for further instructions.

Save

Appendix 3.8. Introduction into the elicitation exercise

Introduction | About you | Instructions | Question 1 | Question 2 | Question 3

Introduction

The study we are doing is referred to expert elicitation. This section will explain what expert elicitation is and why we are conducting it.

What is expert elicitation?

Elicitation is intended to obtain an expert's (you!) beliefs in a numerical (statistical) form - basically getting them down on paper.

Your experience with patients with foot and ankle problems and/or elderly patients in general makes you the expert here!

This does **not** mean that you are expected to know the answer to these questions. These beliefs may be things that you already have opinions on or are quite knowledgeable about but others may require some deep thinking. There are no right or wrong answers to these questions - we just want to know your opinions.

If you are unsure about (or don't know the answer to) a question you should still answer it. Just express how uncertain you are about it in your response (we will show you how to do this shortly)

Why are we doing this elicitation exercise?

You are **not being assessed** in any way during this exercise.

We are researching effectiveness of an intervention designed to prevent falls in elderly. We have no information about how effective this intervention will be long term. Without your input we have nothing to inform a decision about whether the intervention is effective in the long run.

Why me?

Your experience with elderly people, patients who have suffered falls or those with foot and ankle problems makes you the expert here.

There is a range of questions in the exercise and you may feel that some of the questions are outside your field of expertise. This is ok, please answer all questions anyway. We will ask how sure you are about your answer, and if you are not very sure at all, you can express that in your answer.

We will ask different professionals to complete this exercise and we would like everyone to answer as fully as possible.

To proceed please go to the 'About you' tab at the top of the page.