**Informing reimbursement decisions using cost-effectiveness modelling: a guide to the process of generating elicited priors to capture model uncertainties**

**Laura Bojke, Bogdan Grigore, Dina Jankovic, Jaime Peters, Marta Soares, Ken Stein**

## Abstract

In informing decisions, utilising health technology assessment, expert elicitation can provide valuable information, particularly where there is a less developed evidence-base at the point of market access.

In these circumstances formal methods to elicit expert judgements are preferred to improve the accountability and transparency of the decision making process, help reduce bias and the use of heuristics, and also provide a structure which allows uncertainty to be expressed.

Expert elicitation is the process of transforming the subjective and implicit knowledge of experts into their quantifiable expressions. The use of expert elicitation in health technology assessment is gaining momentum, and there is particular interest in its application to diagnostics, medical devices and complex interventions such as in public health or social care.

Compared with the gathering of experimental evidence, elicitation constitutes a reasonably low cost source of evidence. Given its inherent subject nature, the potential biases in elicited evidence cannot be ignored and, due to its infancy in HTA, there is little guidance to the analyst wishing to conduct a formal elicitation exercise. This article attempts to summarise the stages of designing and conducting an expert elicitation, drawing on key literature and examples most of which is not in HTA. In addition we critique their applicability to HTA, given its distinguishing features.

There are a number of issues that the analyst should be mindful of, in particular the need to characterise appropriately the uncertainty associated with model inputs and the fact that there are often numerous parameters required, not all of which can be defined using the same quantities. This increases the need for the elicitation task to be as straightforward as possible for the expert to complete.

## Key points

1. Expert elicitation can provide valuable information, particularly where evidence is missing, where it may not be as well developed or limited.
2. The potential biases in elicited evidence cannot be ignored and, due to its infancy in HTA, there is little guidance to the analyst wishing to conduct a formal elicitation exercise.
3. There are a number of issues that the analyst should be mindful of, in particular the need to characterise appropriately the uncertainty associated with model inputs and the fact that there are often numerous parameters required, not all of which can be defined using the same quantities.

**Compliance with Ethical Standards**

No funding was received for the work carried out for this manuscript.

Laura Bojke was supported in the preparation/submission of this paper by the HEOM Theme of the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH). [www.clahrc-yh.nir.ac.uk](http://www.clahrc-yh.nir.ac.uk). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health.

Laura Bojke, Bogdan Grigore, Dina Jankovic, Jaime Peters, Marta Soares and Ken Stein all have no conflicts of interest.

**Acknowledgements**

Laura Bojke, Bogdan Grigore and Dina Jankovic were primarily responsible for drafting the manuscript. Jaime Peters, Marta Soares and Ken Stein contributed towards writing and commented on various versions of the manuscript.

## Introduction

In the face of constrained budgets, unavoidable decisions about the use of health care interventions have to be made. Decision makers seeking to maximise health for their given budget should use the best available information on effectiveness and cost-effectiveness, and for such they may procure a process of gathering and combining existing evidence in this context called Health Technology Assessment (HTA).

Unfortunately, there is almost always a degree of uncertainty associated with health care decisions. Evidence to inform a decision may be incomplete or in some situations completely missing. In these circumstances a decision can be deferred until evidence becomes available; however in some areas, such as rare diseases, the required evidence may never become available. Decisions often cannot be deferred without consequences such as loss in patient benefits and higher incurred costs, and as such decision makers will often reach a decision using uncertain information.

In informing decisions, utilising HTA, expert elicitation can provide valuable information, particularly where evidence is missing, where it may not be as well developed (e.g. diagnostics, medical devices, early access to medicines scheme or public health) or limited (insufficient, not very relevant, contradictory and/or flawed). Here, formal methods to elicit expert judgements are preferred to improve the accountability and transparency of the decision making process, to help mitigate against biases and the use of heuristics, and also to provide a structure which allows uncertainty to be expressed.

Elicitation is the process of transforming the subjective and implicit knowledge of experts into their quantifiable expressions. There are many possible uses for elicitation in HTA (Box 1 provides some examples but is not an exhaustive list), which may or may not require elicitation of expert opinion as probability distributions. For instance, for items 1 and 2 in Box1, it would not be appropriate to consider eliciting distributions, however for items 3, 4 and 5 it would be important. The exact form of validation will govern if probability assessments are appropriate for item 6.

**Box 1: Uses of elicitation in cost-effectiveness modelling**

• Generating an appropriate set of comparators.

• Identifying appropriate patient pathways and relevant events.

• Describing parameters and their associated uncertainty.

• Quantifying the extent of bias, or improving generalizability from one context to another.

• Characterizing structural uncertainties either through generating differential weights for scenarios or by eliciting distributions of parameterized uncertainties.

• Validating or calibrating model estimates

Expert elicitation is a relatively new technique in HTA and there are few examples of its use[1], although momentum is growing. The focus of the literature thus far has been primarily on generating initial estimates of parameters and their associated uncertainty[2, 3]. A recent review suggested that, despite the increased use of such techniques, there are no common standards for elicitation in HTA[1], nor are there standards for reporting[4]. This article attempts to summarise the steps in designing and conducting an expert elicitation and highlights where there are methodological uncertainties in HTA, specifically used to inform cost-effectiveness analyses. Based on their typical chronology, activities in elicitation can be divided in three stages[5]: pre-elicitation (including the design of the exercise and expert selection), elicitation (the conduct of the exercise) and post-elicitation (including the synthesizing of data from multiple experts and assessment of the exercise). In describing these stages we have sought to highlight where choosing alternative approaches could lead to different outcomes and where appropriate we use relevant examples from both HTA and other disciplines to do this.

## Pre-elicitation

Decisions on what and how to elicit unknown quantities should be determined by the intended purpose of the exercise and how the quantitative expressions will be used. While much is context specific, there are a number of issues to consider, and these can be categorised as: whose beliefs, what and how to elicit, and eliciting complex parameters (e.g. correlation).

*Whose beliefs?*

The literature does not offer a widely agreed definition of an expert. For some authors, ‘expert’ merely indicates those persons from whom opinion is recorded, without implying any particular nature or extent of expertise[6], while others have attempted to identify certain characteristics that could be used to select experts. For instance, Hora and von Winterfeldt [7] give six criteria for experts: a) tangible evidence of expertise; b) reputation; c) availability and willingness to participate; d) understanding of the general problem area; e) impartiality; f) lack of an economic or personal stake in the potential findings. While some of the requirements for experts could be easily checked (e.g. conflicts of interest, availability), others may be more difficult to confirm (e.g. impartiality).

It is generally agreed that any participant in an elicitation exercise should be a “substantive expert in the particular area”[8]. Such experts could be identified from a number of sources ranging from citations in peer reviewed articles to membership of professional societies. However, the issue of whether an expert should possess any particular elicitation skills (e.g. the statistical knowledge to accurately express their belief as a distribution, sometimes referred to as normative expertise), is less clear and will depend on the complexity of the task. Experts with normative knowledge may be able to express quantities such as population moments or parameters of statistical distributions. Experts with less developed normative skills may still be able to provide reasonable estimates of observable quantities, such as proportions[9], using their clinical knowledge.

Besides identifying the relevant group of experts, there is also the issue of how many experts to include in an exercise. Generally, multiple experts can provide ‘better’ information than a single expert [10, 11] and can overcome individual biases [12]; however, the appropriate number of experts may depend on many factors. Ideally, heterogeneity among experts (i.e. experts with different roles or perspectives) is desirable [10], as it prevents the over-emphasis of similar opinions. The predictive power of experts also increases when more experts are added [13], although with diminishing marginal returns. Some authors recommend that at least four experts are included [14], others suggest that between six and twelve experts should be included [15]. Availability of experts is, however, a critical factor [15] and the number of opinions that could be gathered in a typical HTA may be limited. Elicitation exercises reported so far in HTA have used numbers of experts ranging from one to twenty three[2]. While a recommendation on the number of experts to include is not within the purpose of this paper, expert selection should be considered as early as feasible in the elicitation exercise.

### *What to elicit?*

Many previous elicitation exercises have sought to elicit probabilities or occurrence of events; however costs, quality of life weights and relative effectiveness can also be elicited[3] as well as their associated uncertainty. When considering what to elicit, two factors should be taken into account: what is required in the analysis and what is reasonable to ask of experts. For many experts in HTA, that do not possess normative skills and instead rely on substantive practical experience, it may not be advantageous to elicit un-observables[16], such as the hazard. In addition, many experts also find it cognitively challenging to express moments of a distribution (except possibly the first moment, the mean)or coefficients for covariates[16]. Little empirical evidence exists on the most appropriate way to elicit a treatment effect, however it can be hypothesised that if non-normative experts are unable to give informed estimates on parameters, such as odds/relative risk/hazard ratios, it may be more appropriate to elicit the probability of an event in patients who receive the intervention and those who do not, separately, then derive the treatment effect separately from the elicitation task.

Once the analyst has decided what parameters to elicit, the quantities used to express these need to be determined. In some instances there are alternative ways to present the same parameter. For example, when eliciting a transition probability, experts can be asked to indicate their beliefs regarding the probability itself, the time required for x% patients to experience the event or the proportion of patients which would have had the event after y amount of time. In other words, conditional on particular assumptions, evidence on each of these quantities can inform the same parameter. In selecting which is most appropriate, there may be a need to reflect the format of other parameters in the model that may be used jointly with the elicited parameters. Also, where multiple parameters are to be elicited, the analyst may find it advantageous to promote homogeneity on the quantities used, avoiding, for example, seeking judgements on transition probabilities by using proportions of patients for some parameters and the time required for x% patients having had the event for others.

### *How to elicit?*

After choosing which quantities to elicit, the expert needs to be able to express their degree of uncertainty for these quantities. While non-numerical expressions of uncertain quantities can be coded as probabilities[17], this is largely impractical because of the many linguistic and semantic confusions that can occur [18]. Single numerical estimates have been used extensively in decision making [8], however probability distributions are increasingly elicited from experts. The latter avoids attributing unreasonably high confidence to the expert’s estimate and, at the same time provides adequate parameter specification for current decision modelling developments [19].

A probability distribution is usually elicited by asking experts to specify their beliefs for a number of summaries (e.g. probabilities, shape features, moments) characterising their uncertainty surrounding the quantity of interest [6]. These summaries can then be used to describe a distribution that represents the expert’s opinion. Ideally, the focus should be on eliciting summaries with which the experts are familiar, and which could be used to provide instant feedback.

The number of summaries to elicit should be sufficient that a distribution can be built with a degree of precision, but limit the burden on the expert[6, 17]. For instance, experts can be asked to elicit credible intervals directly[20] (the range of values that an expert believes possible within a specified degree of credibility, usually 95%) or other percentiles of the distribution. Variable interval methods can be used, where percentiles are pre-specified and the expert is asked to indicate intervals of values in accordance with his/her beliefs about the particular parameter. Alternatively, the fixed interval method, which is also based on percentiles, requires the analyst to specify a set of intervals that a specific quantity *X* can be contained within. The expert then gives the probability that *X* lies within each interval. One such method, that has been previously used in HTA, is the “four complementary interval” method [20-22], where experts are asked to assign their degree of belief (probability) to 4 intervals, for example 20-25, 25-30, 30-35 and 35-40 events. Experts have expressed some preference for this method, due to it being both easy to engage with and offering a good representation of beliefs [20, 21].

Another method that has been applied in HTA is the histogram technique (also referred to as probability grid, “chips and bins” or trial roulette) [9, 20, 21, 23-25]. This is a graphical derivation of the fixed interval method where the expert is presented with possible values (or ranges of values) of the quantity of interest, displayed in a frequency chart on which he/she is asked to place a given number of crosses (or chips) in the intervals (or bins). Histograms are appealing to even the most non-technical of experts because a probability distribution can be built intuitively without extensive statistical knowledge [26] (see Box 2 for a previous application of this method by Soares et al, 2011[9]). Many more such methods have been described in the elicitation literature[6], however only a few have been used in HTA[2].

In addition to the methods used, decisions also need to be made on the format of the elicitation. The elicitation can be undertaken on a face-to-face basis with one or more experts and a facilitator. Alternatively the task can be conducted at distance, for example by email or via the internet. 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[27-29]. Interview guides are also easier to develop than surveys as questions can be adapted to elicit more relevant responses. 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. On the other hand, arranging individual face-to-face elicitations with experts is time intensive [29]. Grigore et al[2] found that the median amount of time from experts agreeing to participate and the actual face-to-face meeting taking place was approximately 3 months[21]. If a group face-to-face elicitation is to take place, resources are needed to get all of the experts together at the same time and venue, and if distance elicitation is to be done, resource needs to be invested in a tool which experts can use themselves with minimal guidance. There are a number of ‘off the shelf’ tools available for expert elicitation, including the MATCH Uncertainty Elicitation Tool[30]. This and other freely available tools have been summarised elsewhere[31].

What proportion of patients do you think would have a grade 3 reference ulcer (rather than a grade 4 reference ulcer)?

You have inserted 18 crosses in this grid, please insert 3 more crosses.

Please include a total of 21 crosses

**Return to the previous screen**

**Clear grid**

**Box 2 – Application of the histogram method**

In an empirical application uncertain quantities were elicited to inform a cost effectiveness model of negative pressure wound therapy for severe pressure ulceration. 23 nurses elicited 18 uncertain quantities as probabilities. A common scale was used (from zero to 100). A snapshot of the instrument used to display the questions is represented in Figure 1.

**Figure 1: Example of elicitation questionnaire**For each uncertain quantity, individual experts were asked to place 21 crosses on a grid defined to have 21x21 cells (Figure 2). For ease, the possible values the quantity may take were discretized (i.e. 0, 5, 10, ., 100). By placing the 21 crosses in the grid, the expert is allocates a probability mass to each of the possible values. The expert can express full certainty for a value by placing all of the crosses in a single column. By attributing one cross to each possible value expert expresses full uncertainty.

**Figure 2: Graphic set up for the data capture histogram**

Think of UK patients with at least one debrided grade 3 or 4 pressure ulcer (greater than 5 cm2 in area).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | **x** |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | **x** |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | **x** |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | **x** | **x** | **x** |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | **x** | **x** | **x** | **x** | **x** |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | **x** | **x** | **x** | **x** | **x** | **x** | **x** |  |  |  |  |  |  |  |  |
| 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100 |

### *Eliciting complex parameters*

Complex parameters, frequently used in cost-effectiveness modelling, include joint and conditional quantities, regression parameters, correlation and transitions defining a multistate model (e.g. a Markov model). Perhaps the most common challenge arises with parameters that are interdependent (i.e. correlated), where a joint distribution may need to be elicited[32]. To model dependence, the analyst can either assess the model’s sensitivity to variations in the correlation coefficient, or estimate the correlation as part of the elicitation exercise. There are a number of methods to elicit correlations[6]; however, applied studies have failed to achieve consensus on which is the most appropriate. Methods include verbal descriptions of likely strength of correlation, direct assessment and the specification of a percentile for quantity *X* contingent on a specified percentile for quantity *Y*[33]. However, the task of eliciting probability distributions conditional upon other probability distributions is likely to be too cognitively difficult for many experts. In these circumstances it may be appropriate to use a 'second best' solution and to elicit distributions conditional upon means or best guesses. This was the approach used by Soares et al[9], where experts were first asked to record the probability (and uncertainty) of a patient's pressure ulcer being healed when they received treatment with hydrocolloid dressing (HC). For experts who believed the effectiveness of other treatments was different from HC, the distribution for the treatment effect was elicited, conditional to the value assumed known for HC.

## Conducting the elicitation exercise

### *Understanding bias and the impact of heuristics*

It can be useful to understand how experts estimate unknown quantities, in particular the influence of a number of psychological shortcuts[34] (referred to as *heuristics)*, which may make the elicitation task more difficult to design and results subject to bias. These heuristics are the principles or methods experts use to enable them to formulate their judgements, particularly in the assessment of probabilities. Since the first such cognitive strategies were defined by Tversky and Kahneman, 1981[35] many more have been described[36]. Examples of biases in elicitation are described in Box 3. This is not intended to be an exhaustive list, instead a recent review of the issues and de-biasing strategies can be found in Montibeller and von Winterfeldt, 2015[37].

These heuristics can be useful in formulating beliefs, but can sometimes lead to systematic errors. In addition to cognitive heuristics, bias arises from the expert’s motivation, operational experience, adherence to a particular hypothesis (i.e. confirmation bias) and interaction with peers (e.g. groupthink). Also, not all sources of bias are associated with the expert. Framing of the questions can also introduce bias: for instance, subjective preference for a health intervention over its alternatives may change when the effectiveness is expressed as mortality rather than survival [35].

The analyst should be aware of the potential impact of such biases when eliciting experts’ beliefs and should take appropriate measures to mitigate against them[36] by designing the elicitation task appropriately[38]. For instance, avoiding sample numbers on which the expert may anchor responses, framing questions in a neutral way, checking for consistency etc. [38]. Recommendations for de-biasing are available in the literature[39, 40], however, even when measures are taken to control expected biases, the direction and extent of true bias is likely to be unknown.

**Box 3 - Examples of biases in elicitation**

Biases in elicitation can include:

* Biases associated with experts
  + Motivation biases: e.g. when experts have an incentive (for example, financial) to reach a certain conclusion
  + Cognitive biases: these commonly involve the use of heuristics, or ‘rules of thumb’, to help reach decisions, solve problems or form judgements quickly. Examples are:

*Conjunction fallacy:* when the probability of conjunction (combined) events is judged to be more likely than either of its constituents.

*Availability:* where easy to recall events (like natural disasters) are judged to have a high probabilities of occurring.

*Hindsight bias:* the tendency to overestimate the predictability of past events

*Anchoring effect*: the tendency to rely on an anchor value that does not provide any information about the actual value.

* Biases associated with elicitation methods
  + Structuring elicitation questions: biases may arise from how the question is framed, for example, if relevant events have been omitted experts are less likely to consider these in replying. But biases can also occur when scales are used; for example, *contraction bias* occurs when the full range of a scale has not been presented to the expert
  + Elicitation medium (e.g. interview or email survey) or aggregation method: as an example, experts in group meetings (typically conducted when consensus aggregation methods are applied) tend to adopt a stronger position often resulting in overconfident statements.
  + Fitting probability distributions: the encoding of the summaries elicited from the expert as a distribution usually implies assumptions referring to the shape of the distribution that may differ from what the expert intended. For example an expert, when fitting distributions to his/her own beliefs, may be driven by familiar probability distribution shapes, in particular bell shaped curves (*familiarity bias*).

Whilst the literature suggests that biases cannot be completely avoided, it is good practice to be aware of possible biases and to employ strategies to mitigate against these (de-biasing) in both designing and conducting the elicitation exercise.

### *Explaining the concept of uncertainty*

Eliciting measures of uncertainty can be complicated, particularly as one wants to ensure that quantities reflect uncertainty in the expected value rather than variability or heterogeneity. This is in part driven by the format of the exercise but training is also paramount especially where experts have limited experience of elicitation[28, 39] and/or there is a reason to believe that they poses limited normative skills[6].

In training experts it can also be useful to present contrasting examples of uncertainty and variability to help the expert understand the key distinctions. Visual aids (such as the histogram described above) can be useful for the elicitation exercise and can help to reduce the burden on experts. Experts will often respond more consistently and confidently to questions and give more accurate assessments if they are familiar with the purpose and methods used for the elicitation exercise.[41] Frequent feedback should also be given to the expert during the elicitation process and, if possible, experts should be allowed to revise their judgements. Processes such as Delphi[42], that include an explicit feedback process, are relevant to consider here.

## Post-elicitation

There are a number of methodological choices regarding how to use the experts’ beliefs. Some of these, such as approaches to aggregating the beliefs of several experts, imply consequences for the elicitation task itself as well as its design.

### *Synthesising multiple elicited beliefs*

When judgements from several experts are required it is often desirable to obtain a unique distribution that reflects the judgements of them all[43]. Methods to achieve this fall into two categories, behavioural and mathematical.

Behavioural approaches have typically focussed on achieving a single distribution for the parameter of interest. A group of experts is asked jointly to elicit its beliefs as if it was a single expert and experts are encouraged to interact in order to achieve a level of agreement for a particular parameter. Thus synthesis becomes part of the elicitation process, avoiding the need for quantitative synthesis of individual opinions post elicitation.

Several phenomena associated with group interaction may lead to biased results such as: dominance of strong-minded or strident individuals[28], groupthink (desire of quickly reaching agreement at the expense of a reasonable outcome), socially reinforced irrelevance (i.e. taboos) or group reinforced bias due to common background of group members[44]. There are a number of behavioural aggregation techniques and they employ various techniques to alleviate such phenomena.  In the Nominal Group Technique[45] individuals first express their own beliefs to the group before updating these on the basis of group discussion (interaction). The discussion is facilitated either by an expert on the topic or a credible non-expert. Values (distributions) are then prioritised and/or ranked by the group. Probably the best known consensus approach is the Delphi technique[42], which has been frequently applied to decision making in healthcare[46]. It distinguishes itself from conventional group approaches in that it incorporates controlled feedback and does not require a consensus to be achieved, nor does it postulate that the true answer has been determined if consensus is reached [47].

Social dynamics are however not the only concern with behavioural methods. For those that focus on consensus, doing so may not be easily achieved and, in some circumstances, there may not be a value that all experts are willing to agree on[48]. Perhaps most importantly, the focus on achieving consensus means that some behavioural approaches, such as the nominal group, do not consider the inherent uncertainty in experts’ beliefs about a parameter. In addition, there is a tendency for the group to be over confident when reaching consensus about an unknown parameter[49].

Mathematical approaches to synthesising multiple beliefs do not attempt to generate a consensus. Rather, they focus on combining individual beliefs to generate a single distribution using mathematical techniques. Subsequent aggregation of individual experts’ estimates into a single distribution is the preferred approach in applied studies [43], however, some studies have also used individual experts’ assessments as separate scenarios, which can be useful to explore. Synthesis of data from multiple experts often involves two steps: (a) fitting probability distributions, and (b) combining probability distributions.

### 

### *Fitting probability distributions*

Fitting probability distributions to elicited beliefs can either be undertaken by the analyst post elicitation or by asking the experts to assess fitting as part of the elicitation exercise[36]. The method of fitting of probability distributions will often be governed by the intended purpose of the elicited evidence. In particular if the distributions are to be used in Bayesian updating, i.e. they will be updated with additional data (sample information), or they will be used to represent uncertainty regarding an unknown parameter without the use of additional data[36].

In Bayesian updating, parametric distributions can be fitted if one believes that an expert's estimates can be represented in such a way. The choice of parametric distribution may be driven by the need to ensure conjugate distributions (i.e. distributions from the same statistical family), which is advantageous for analytical simplicity. However, the development of computational methods has made it possible to choose non-conjugate distributions. Non-parametric methods can also be used. These do not assume that the data structure can be specified *a priori*; in effect, they have an unknown distribution.

### *Combining probability distributions*

There are two main methods for combining probability distributions: weighted combination and Bayesian approaches [43]. Weighted combination is referred to as *opinion pooling*, specifically linear opinion pooling or logarithmic opinion pooling. If *p(θ)* is the probability distribution for unknown parameter *θ*, in linear pooling experts’ probabilities are aggregated using the simple linear combination: , where *w*i represents a weight assigned to expert *i*. In logarithmic opinion pooling, averaging is undertaken using a multiplicative averaging method. These two methods can differ greatly, with the logarithmic method typically producing a narrower distribution for the parameter implying less uncertainty in the estimate.

An example of the use of linear pooling is described by White et al[50], who elicited expert opinion on treatment effects and the interaction between three trials. Experts were asked to assign a *weight of belief* (up to 100) to intervals of annual event rates. Experts’ weights were then combined by taking the arithmetic mean of individual assessments (linear pooling with equal weighting of experts). Weights can be derived from experts’ professional experience. A recent example by Shabaruddin et al[51] used the mean number of relevant patients seen to derive a weighting for each expert. This was then used to generate weighted means in the linear pooling. Alternatively, weights can be based on experts’ elicitation performance. Performance based weighting approaches are described in Genest et al[52] and these are discussed below (assessing the elicitation process) .

More recently, there has been a move to using Bayesian models for combining probabilities[53]. Aggregation in a Bayesian model uses the experts’ probability assessments to update the decision makers’ own prior beliefs about an uncertain parameter. These methods have not yet been applied in HTA and the need for the decision makers input is difficult to implement in practice.

### *Interdependence of experts*

Regardless of the method used to combine experts’ probability distributions, an additional level of complexity is introduced when the assumption that experts provide independent beliefs is not sustainable[54].

A degree of interdependence is likely if experts are chosen from a given professional organization or are basing their beliefs on shared experience or information. Although interdependence between experts is difficult to empirically test, if it is suspected, models must be used to consider any correlation. Mathematical aggregation techniques such as opinion pooling do not account for such dependencies and instead joint distributions must be used, incorporating the covariance matrix for the experts’ assessments. A recent study used a clustering approach to capture the extent of interdependence between experts’ beliefs[55]. For a particular parameter, experts’ best estimates were grouped together into a variable number of clusters using Ward's minimum variance criterion. This was then used to determine if estimates should be treated as one group or as two or more clusters.

## Assessing the elicitation process

A number of techniques are available to determine “how well” the elicitation has done and in particular to establish the value of the elicited beliefs.

Two general properties have been described to evaluate the elicited priors[56]: validity, referring to the accuracy with which the opinion of one person is encoded in a probabilistic form, and reliability, referring to the reproducibility of the result[56]. A number of components can be used to measure these properties: internal consistency, overfitting, fitness for purpose, scoring rules and calibration.

Kadane and Wolfson[16] propose that beyond such measures, practicality should be used as a criterion in designing elicitation and in evaluating its coherence . More recently, Johnson et al[39] expand on the work of Kadane and Wolfson by describing four measurement science criteria relevant for elicitation: validity, reliability, responsiveness (the ability to detect a meaningful change in belief when this occurs) and feasibility. Below we describe a number of approaches to assess the elicitation process, which address such criteria.

### *Internal consistency*

The issue of internal consistency is particularly relevant when eliciting probabilities. An expert's assessment of one (or more) unknown parameters should be consistent with the laws of probability[36], for example if P(A) = 0.2, P(B) = 0.3 and P(A or B) =0.4, these probabilities are non-coherent. Achieving coherence may, however, involve more complex reasoning and, in the presence of such complexity, either incoherent judgements are transformed for further use or the exercise is constructed in order to minimise or eliminate incoherence[16]. Qualitative feedback can also be useful in assessing internal consistency[6]. Any discrepancies can be fed back to the experts and appropriate adjustments to assessments made. Another way to check for consistency is overfitting, where more estimates are elicited from the expert than are minimally necessary to build the probability distribution, then a distribution is fitted to reconcile all assessments[6].

### *Fitness for purpose*

Inevitably some degree of imprecision will remain in the elicited beliefs and the fitted distributions. Sensitivity analysis can, therefore, be useful in determining if the ultimate results of the analysis change if alternative (but also plausible given the expert’s knowledge) distributions are used. A commonly used sensitivity analysis in a Bayesian framework is to explore alternative prior distributions (see Kuhnert et al[40] for an example). If the model results do not change appreciably in sensitivity analysis, then the distributions can be said to represent the experts’ knowledge and thus are fit for purpose.

### *Scoring rules*

For parameters that are known or subsequently become known to analysts, i.e. those performing the elicitation or analysing its results, comparisons can be made between elicited distributions and those known distributions (sometimes referred to as seeds[57]). This provides the opportunity to assess the ‘closeness’ of the elicited and actual distributions. This ‘scoring rule’ then attaches a reward (a score) to an expert using some measure of performance, with those gaining higher scores being regarded as ‘better’.

There are various scoring rules described in the literature. The majority use some measure of calibration, the distance between elicited probabilities and observed outcomes. There is more than one method for quantifying calibration, including but not exclusive to Brier’s probability scores [58], Kullback–Leibler divergence [57], and linear and spherical calibration scores [6].

Scoring rules can also incorporate other measures of performance. For example, Cooke’s Classical Model [57] combines calibration and experts’ confidence into overall scores, arguing that if two experts are equally accurate, the one who is more confident should be scored higher. Murphy [59]combines calibration, the variance of the observed parameter and resolution into scores, where resolution is the measure of how well the expert discriminates between events with high and low relative frequencies. Yates’s scoring rule [60] takes into account bias (calibration), slope (comparable to Murphy’s resolution) and scatter (random error in experts’ judgment). It is not clear which scoring rule is optimal.

Scoring has been suggested to be a useful feedback tool for training experts [61, 62]. Experts’ scores on seed questions can be used as a predictor of their performance for unknown parameters[63]. When experts are found to consistently overestimate or underestimate probabilities of events, these can be used to adjust estimates of future unknown quantities (post-hoc adjustments). They can also be useful in identifying experts with skills valuable to elicitation, however the credibility of the derived score may be affected by the seed question and performance against known distributions may not be a good predictor of performance against unknown parameters.

In the example by Chaloner[64] elicitation was used to inform a model looking at the intermediate results of a randomized trial. On completion of the trial, comparisons were made between elicited estimates and those based on actual data. It was concluded that the elicitation exercise, although producing some thought-provoking results, did not necessarily predict trial outcomes with much accuracy. Although not done explicitly as part of the exercise, it would have been possible to score and weight experts' beliefs retrospectively, with a view to improve prediction.

## Conclusions

The use of formally elicited evidence to parameterise HTA decision models has yet to filter into standard practice, including, for example the reference case for evaluation by the National Institute for Health and Clinical Excellence (NICE)[65]. However, it has huge potential particularly in areas where there is a less developed evidence-base at the point of market access, such as diagnostics (the co-dependent technology process in Australia and the Diagnostics Assessment Programme in the UK), medical devices and complex interventions such as in public health or social care. The evidence base for these products is often less developed as the demonstration of efficacy or performance is not subject to the same regulatory requirements as pharmaceutical products.

Compared with the collection of experimental evidence, elicitation also constitutes a reasonably low cost source of evidence, although the time and resources needed should not be overlooked. However, the potential biases in elicited evidence cannot be ignored and, due to its infancy in HTA, there is little guidance to the analyst wishing to conduct a formal elicitation exercise in this area.

This paper has attempted to summarise the stages of elicitation and the methodological choices that an analyst will face when designing and conducting a formal elicitation exercise in HTA. In doing so, it attempts to draw together some of the vast literature on elicitation, most of which is not in HTA.

There are a number of issues that an analyst working in HTA should be mindful of. In particular is the reality that, in HTA, there are often numerous parameters required, not all of which can be defined using the same quantities. This increases the need for the elicitation task to be as straightforward as possible for the expert to complete.

There are numerous methodological issues that need to be resolved when applying elicitation methods to HTA decision analysis. The need to characterise appropriately the uncertainty associated with model parameters necessitates the use of probabilistic assessments of quantities. Distinguishing between uncertainty and variability may not be instinctive for many health care experts, and further research is required to explore which elicitation methods and training prompts encourage the expression of uncertainty. It is also not clear if behavioural or mathematical approaches perform better in HTA and the extent to which this is driven by the specific task. For instance there may be some circumstances in which experts are only able to formulate their beliefs about particular quantities as a group, for example when the parameters relate to a new technology.

In choosing to use more complex methods of elicitation it is also important to note that the complexity of many HTA decision models and the need to capture experts’ beliefs, as inputs into these, creates a tension between generating unbiased elicited beliefs and populating a decision model with usable parameters. However where experimental evidence is sparse and difficult to collect, such as for emerging technologies, the need to explore the added value of elicited evidence seems particularly pressing.References

1. Hunger T, et al., *Using expert opinion in health technology assessment: a guideline review.* International Journal of Technology Assessment in Health Care, 2016. **32**(3): p. 131-9.

2. Grigore B, et al., *Methods to elicit probability distributions from experts: a systematic review of reported practice in health technology assessment.* Pharmacoeconomics, 2013. **31**(11): p. 991-1003.

3. Hadorn D, et al., *Use of expert knowledge elicitation to estimate parameters in health economic decision models.* Int J Technol Assess Health Care, 2014. **30**(4): p. 461-8.

4. Iglesias CP, et al., *Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations.* Pharmacoeconomics, 2016

5. Hart A, et al., *Training Course on Steering an Expert Knowledge Elicitation.* EFSA Journal, 2016.

6. O’Hagan A, et al., *Uncertain judgements: eliciting experts’ probabilities*, ed. J.W. Sons. 2006.

7. Hora SC and Von Winterfeldt D, *Nuclear waste and future societies: A look into the deep future.* Technological Forecasting and Social Change 1997. **56**(2): p. 155-70.

8. Jenkinson, D., *The elicitation of probabilities: A review of the statistical literature. Bayesian Elicitation of Experts' Probabilities (BEEP)*. 2005: Sheffield University.

9. Soares MO, et al., *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, 2011. **30**(19): p. 2363-2380.

10. Clemen RT and Winkler RL, *Aggregating probability distributions*, in *Advances in decision analysis: From foundations to applications*, Edwards W, Miles Jr R, and Von Winterfeldt D, Editors. 2007, Cambridge University Press: Cambridge, UK. p. 154-176.

11. French, S., *Aggregating expert judgement.* Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 2011. **105**(1): p. 181-206.

12. Johnson, S.R., et al., *Methods to elicit beliefs for Bayesian priors: a systematic review.* Journal of Clinical Epidemiology, 2010. **63**(4): p. 355-369.

13. Kattan, M.W., et al., *The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones.* Medical Decision Making, 2016. **36**(4): p. 536-540.

14. Cooke, R.M. and L.H.J. Goossens, *Expert judgement elicitation for risk assessments of critical infrastructures.* Journal of Risk Research, 2004. **7**(6): p. 643-656.

15. Knol AB, et al., *The use of expert elicitation in environmental health impact assessment: a seven step procedure.* Environmental health : a global access science source, 2010. **9**: p. 19.

16. Kadane J and Wolfson LJ, *Experiences in elicitation.* Journal of the Royal Statistical Society: Series D (The Statistician), 1998. **47**(1): p. 3-19.

17. Renooij S and Witteman C, *Talking probabilities: communicating probabilistic information with words and numbers.* International Journal of Approximate Reasoning 1999. **22**(3): p. 169-194.

18. Bruine de Bruin W, et al., *What Number is “Fifty-Fifty”?: Redistributing Excessive 50% Responses in Elicited Probabilities.* Risk Analysis, 2002. **22**(4): p. 713-23.

19. Claxton K, et al., *Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra.* Health economics, 2005. **14**(4): p. 339-47.

20. Leal J, W.S., Legood R, Blair E,, *Eliciting expert opinion for economic models: an applied example.* Value Health, 2007. **10**(3): p. 195-203.

21. Grigore B, et al., *A comparison of two methods for expert elicitation in health technology assessments.* BMC Med Res Methodol, 2016.

22. Pibouleau L and Chevret S, *An Internet-based Method To Elicit Experts’ Beliefs For Bayesian Priors: A Case Study In Intracranial Stent Evaluation.* International Journal of Technology Assessment in Health Care, 2014. **30**(4): p. 1-8.

23. Bojke L, et al., *Eliciting distributions to populate decision analytic models.* Value in Health, 2010. **13**(5): p. 557-64.

24. McKenna C, et al., *Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic evaluation*, -. Health Technology Assessment, Editor. 2009.

25. Speight PM, et al., *The cost-effectiveness of screening for oral cancer in primary care.*, H.t. assessment, Editor. 2006.

26. Van Noortwijk JM, et al., *Expert judgment in maintenance optimization.* Reliability, IEEE Transactions on, 1992. **41**(3): p. 427-32.

27. Bowling A, *Mode of questionnaire administration can have serious effects on data quality.* Journal of Public Health, 2005. **27**: p. 281-91.

28. Knol AB, et al., *The use of expert elicitation in environmental health impact assessment: a seven step procedure.* Environmental health : a global access science source 2010. **9**(19).

29. Baker E, et al., *Facing the experts: survey mode and expert elicitation.* Fondazione Eni Enrico Mattei, 2014. Nora di Lavoro.

30. Morris DE, Oakley JE, and Crowe JA, *A web-based tool for eliciting probability distributions from experts.* Environmental Modelling & Software, 2014. **52**: p. 1-4.

31. Expert Judgement Network 2016; Available from: <http://www.expertsinuncertainty.net/Software/tabid/4149/Default.aspx>.

32. Elfadaly FG and Garthwaite PH, *Eliciting Dirichlet and Connor–Mosimann prior distributions for multinomial models.* TEST, 2013. **22**(4): p. 628-46.

33. Garthwaite PH, et al., *Use of expert knowledge in evaluating costs and benefits of alternative service provisions: a case study.* International Journal of Technology Assessment in Health Care 2008. **24**(3): p. 350-357.

34. WikiBooks. *Cognitive Science: An Introduction/Biases and Reasoning Heuristics*. Cognitive Science: An Introduction 2016 3/5/2017]; Available from: <https://en.wikibooks.org/wiki/Cognitive_Science:_An_Introduction/Biases_and_Reasoning_Heuristics>.

35. Tversky A and Kahneman D, *The framing of decisions and the psychology of choice.* science 211, 1981. **4481**: p. 453-458.

36. Garthwaite PH, Kadane JB, and O'Hagan A, *Statistical methods for eliciting probability distributions.* Journal of the American Statistical Association, 2005. **100**(470): p. 680-701.

37. Montibeller G and von Winterfeldt D, *Cognative and motivational biases in decison and risk analysis.* Risk analysis, 2015. **35**(7).

38. Kynn M, *The ‘heuristics and biases’ bias in expert elicitation.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2008. **171**(1): p. 239-264.

39. Johnson SR, et al., *Methods to elicit beliefs for Bayesian priors: a systematic review.* Journal of Clinical Epidemiology, 2010. **63**(4): p. 355-369.

40. Kuhnert PM, Martin TG, and Griffiths SP, *A guide to eliciting and using expert knowledge in Bayesian ecological models.* Ecology Letters, 2010 **13**: p. 900-914.

41. Mullin TM, *Understanding and supporting the process of probabilistic estimation*. 1986, Carnegie-Mellon University.

42. Dalkey N and Helmer O, *An Experimental Application of the Delphi Method to the use of experts.* Management Science, 1963. **9**(3): p. 458–467.

43. Clemen RT and Winkler RL, *Aggregating probability distributions*, in *Advances in decision analysis: From foundations to applications*, R.M.J.a.D.V.W. W. Edwards, Editor. 2007, Cambridge University Press: Cambridge, UK.

44. Ayyub B, *Elicitation of Expert Opinions for Uncertainty and Risks.* 2001.

45. Rohrbaugh J, *Improving the quality of group judgment: Social judgment analysis and the nominal group technique.* Organizational Behavior and Human Performance, 1981. **28**(2): p. 272-88.

46. Sullivan W and Payne K, *The appropriate elicitation of expert opinion in economic models.* Pharmacoeconomics., 2011. **29**(6): p. 455-9.

47. Keeney S, McKenna H, and Hasson F, *The Delphi Technique in Nursing and Health Research*. 2010: John Wiley & Sons. 208.

48. Myers DG and Lamm H, *The polarizing effect of group discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence.* American Scientist, 1975. **63**(3): p. 297-303.

49. Sniezek JA, *Groups under uncertainty: An examination of confidence in group decision making.* Organizational Behavior and Human Decision Processes, 1992. **52**(1): p. 124-155.

50. White IR, Pocock SJ, and Wang D, *Eliciting and using expert opinions about influence of patient characteristics on treatment effects: a Bayesian analysis of the CHARM trials.* Statistics in Medicine, 2005. **24**(24).

51. Shabaruddin, F., R. Elliott, J. W. Valle, W. Newman and K. Payne,, *Understanding chemotherapy treatment pathways of advanced colorectal cancer patients to inform an economic evaluation in the United Kingdom.* British journal of cancer 2010. **103**(3): p. 315-323.

52. Genest C and McConway KJ, *Allocating the weights in the linear opinion pool.* Journal of Forecasting, 1990. **9**(1): p. 53-73.

53. Moatti M, et al., *Modeling of experts' divergent prior beliefs for a sequential phase III clinical trial.* Clin Trials. **10**(4): p. 2013.

54. Wilson KJ, *An investigation of dependence in expert judgement studies with multiple experts.* International Journal of Forecasting, 2017. **33**(1): p. 325-336.

55. Hoelzer K, et al., *Structured Expert Elicitation About Listeria monocytogenes Cross-Contamination in the Environment of Retail Deli Operations in the United State.*  Risk Analysis, 2012. **32**(7): p. 1139-1156.

56. Wallsten TS and Budescu DV, *State of the art-Encoding subjective probabilities: A psychological and psychometric review.* Management science 1983. **29**(2): p. 151-173.

57. Cooke R.M, *Experts in Uncertainty*. 1991: Oxford University Press.

58. Brier, G.W., *Verification of forecasts expressed in terms of probability.* Monthly weather review, 1950. **78**(1): p. 1-3.

59. Murphy, A.H., *A new vector partition of the probability score.* Journal of Applied Meteorology, 1973. **12**(4): p. 595-600.

60. Yates, J.F., *Subjective probability analysis*, in *Subjective Probability*, G. Wright and P. Ayton, Editors. 1994, John Wiley: London. p. 382-410.

61. Remus, W., M. O'Conner, and K. Griggs, *Does Feedback Improve the Accuracy of Recurrent Judgmental Forecasts?* Organizational Behavior and Human Decision Processes, 1996. **66**(1): p. 22-30.

62. Subbotin, V., *Outcome Feedback Effects on Under- and Overconfident Judgments (General Knowledge Tasks).* Organizational Behavior and Human Decision Processes, 1996. **66**(3): p. 268-276.

63. Bolger F and Rowe G, *The Aggregation of Expert Judgment: Do Good Things Come to ThoseWho Weight? .* Risk Analysis, 2015. **35**(1).

64. Chaloner K, et al., *Graphical elicitation of a prior distribution for a clinical trial.* The Statistician 1993. **42**: p. 341-353.

65. National Institute for Health and Clinical Excellence, *Guide to the methods of technology appraisal* 2013.