Methods for eliciting expert opinion to inform health technology assessment

John Paul Gosling[†]

4th November 2014

1 Health technology assessment and expert elicitation

Health technology is defined to be "any method used by those working in health services to promote health, prevent and treat disease, and improve rehabilitation and long-term care".[1] A key function in public health policy making is the assessment of the potential efficacy and cost-effectiveness of new health technologies. The general term used for this type of assessment is health technology assessment (HTA). When assessing new (or even existing) technologies, there are often gaps in the evidence base to help judge the efficacy and cost-effectiveness of one technology over another. In the assessment of future events or unknown outcomes, elicited opinions are often all a decision maker has for an evidence base. Coupled with this, there could be limited prospect of future data collection due to costs or physical feasibility.

Many academic publications have attempted to set out principles for HTA.[2–4] The core principles include the transparency with regards to evidence sources, reproducibility of evidence, the use of analytical techniques that are methodologically sound and a comprehensive appreciation of uncertainty. Although take up of these principles is not uniform in HTA,[5] many of the principles highlighted here should be noted when using expert judgements.

In the context of uncertainty modelling, elicitation is the process of translating someone's judgements about some uncertain quantities into something useful for a model. This will typically be a probability distribution. Expert elicitation can help us to take stock of the uncertainty about quantities of interest without the cost of data collection. We will be referring to experts throughout as people that have relevant knowledge about an HTA. These experts will have knowledge about the quantity of interest, and their opinion is of interest to the decision maker who wishes to use the HTA. Experts could be health or social care professionals, academics, patients or lay people. The wide coverage here matches with the aims of organisations such as the National Institute for Health and Care Excellence (NICE) to take into account the view of stakeholders and the wider public in HTA. The value of taking this approach is backed up by research on the importance of using lay knowledge in public health research due to the presence of extensive personal experiences.[6–8] Popay and Williams (1996) argue that both lay people and relevant professionals "have a contribution to make to understanding [of health and illness]" and that there is real merit in eliciting judgements about future health under different scenarios. Of course, the same elicitation methods will not work for every expert and consideration of the expert's abilities should always be taken into account when designing and running elicitation exercises.

[†]School of Mathematics, University of Leeds, Leeds, UK. j.p.gosling@leeds.ac.uk

In HTA, stakeholders (including patients and the wider public) are regularly questioned to help value different options and to work out the value that society places on various outcomes.[9–13] This type of elicitation can be broadly defined as value or utility elicitation and is different in nature to eliciting beliefs about model parameters and data quality. These types of elicitation exercises are used to set utility functions and/or provide economic evaluations for formal decision making processes. However, there are a number of principles in common with the expert elicitation schemes considered in this piece.

Probabilistic models are often used in HTAs to capture the effects of both variability (in responses and populations) and uncertainty.[1,2] To be useful in such models, gaps in evidence bases need to be encoded using probability distributions. The need for expert judgements is more apparent when the probabilistic models are taken within a Bayesian framework, which is becoming more popular in HTA.[14–17] Here expert judgements are used to form the basis of an analysis as "prior" distributions. There is therefore great interest in developing defensible elicitation protocols for HTA that will result in justifiable probability distributions.[18,19]

There have been a limited number of published attempts to use expert elicitation to fill evidence gaps in HTA. A recent systematic review of expert elicitation use in HTA returned only 14 articles that explained their procedures to a useful level of detail.[18] Although the number of published examples of elicitation exercises in the context of HTA are few, the examples that are currently available offer some good practical advice for potential users of the techniques and they also highlight a number of challenges worthy of future research. The relevant applications of expert elicitation to HTA will be highlighted throughout this piece.

In the next section, the key principles behind the successful design, implementation and reporting of expert elicitation techniques in general will be discussed along with how results from expert elicitation can be brought into HTA. In Section 3, several popular elicitation methods and protocols will be discussed alongside their application to HTAs. Sections 4 and 5 give a brief review of other elicitation techniques and software implementations respectively with a discussion of some of the limitations of current methods. The final section discusses the overall state of expert elicitation in HTA and some recommendations for future research.

2 The key principles of expert elicitation

Expert elicitation is not a precise science. It can be difficult for the experts to articulate their beliefs, and further complications arise due to the biases of experts and the biases created by the questioning process. The process of questioning people about their beliefs is not a new subject: it has been the focus of many studies.[20–24] However, it is recognised that these studies do not always relate to expert judgements being made to aid scientific understanding or policy making, nevertheless the facilitator of the expert elicitation exercise should be aware of the issues raised in the psychology literature and potential biases in the process, which we will discuss later in this section.

It is important to make the distinction between an elicitation method for capturing judgements about an unknown quantity and an elicitation protocol that refers to the overall process from expert recruitment to the final reporting of the elicitation results. There are

many examples of elicitation methods, but few well-established elicitation protocols. In the development of expert elicitation protocols, there are a number of key principles that often arise. For instance, it has been noted that protocol for expert elicitation should include four key stages: the recruitment and training of experts, the initial elicitation of the experts judgements, the transformation of the judgements into something useful for subsequent analyses, and the feedback of the results to the experts.[25,26]

The first stage consists of the setting up of the elicitation exercise: the clear definition of the scope of the exercise and the selection, recruitment and training of experts. In most assessments, more than one expert will be needed to build up a picture of what is known. From a performance point of view, it has been shown in several experimental studies that judgements derived from group exercises outperform individual experts.[27–30] There is little guidance in the literature about how to select experts for an elicitation exercise. Given the aforementioned NICE principles of inclusion in decision making, it seems advantageous to have experts from different backgrounds involved. It has also been recommended in the expert elicitation literature that prior distributions used in medical applications are representative of the community of experts.[31] However, because experience and background are difficult to classify, facilitators of elicitation exercises will rarely be able to cover the full range of opinions, backgrounds and experiences convincingly.[26,32]

Typically, recruitment is limited by the availability of experts and the ease at which the can be interviewed (this is worse for some behavioural techniques, which we will describe in Section 3, because the experts will be expected to be in the same room). An example of recruitment strategy of experts for HTA arose in the context of quantifying patient survival when using novel devices.[33] Here, a group of leading clinicians took part in an expert elicitation exercise. The experts were recruited because they were in the same place at the same time: the 51st annual conference of the American Society for Artificial Organs. A similar strategy was used to gauge expert opinions on the effect of warfarin as the experts were recruited from the participants of the 2008 American College of Rheumatology or American Thoracic Society meetings.[34] Of course, conferences can provide a good opportunity to get experts together, but there is a danger that the selected experts might be a biased sample because they might all be from a particular academic field and not cover all backgrounds of interest. Whatever the method is used for recruitment, there must be transparency in the process: it is important that subsequent users of the elicitation exercise know whose experience the judgements are based on and why.[26]

It will often be impossible to get all the relevant (or required) experts in the same place at the same time so there is an increasing interest in developing expert elicitation protocols for geographically-dispersed experts.[10,35,36] The internet makes such exercises feasible and cheap web conferencing even enables remote group meetings. In a recent development, a Microsoft Excel-based questionnaire was devised to capture clinicians beliefs for use in HTAs via email.[37] However, remote elicitation sessions are not favoured as often experts can struggle to perform the required tasks due to their inexperience in making the required uncertainty judgements.[38,39]

In some expert elicitation exercises, honorariums are given to the experts to compensate for their time.[40,41] This practice seems appropriate because an expert elicitation exercise can easily take an entire day if not longer especially if preparation is taken into account. However, payments beyond travel and subsistence are rarely given out to the participants

of expert elicitation exercises and, when it comes to aiding policy decisions, many see their participation as being part of their role in society or as a professional.

The training phase should be designed to familiarise the experts with the elicitation process and to give the facilitator an idea about people's ability to answer the questions and (possibly) about their level of skill at making judgements about uncertainty. Of course, having expertise in one area does not lead to expertise in giving probability judgements.[42] Often training exercises are based on an unrelated problem like road distances between two cities[43] or the probability of some upcoming newsworthy event.[44] It could be argued that such exercises give little indication about an expert's ability to answer questions relating to HTA and that the experts may not treat toy problems seriously. It is therefore worth the effort of finding a related problem to the HTA under consideration for the training exercise.

Prior to questioning the experts, it is crucial that the quantity of interest is defined unambiguously: disagreement between experts should be due to differences in experiences rather than differences in interpretation of variable definitions.[26,28] The overall aim of the questioning phase is to determine a probability distribution for use in the HTA, but it is unrealistic to expect that the experts can state their probability distributions explicitly. They will often only be able to state certain summaries of their distribution such as the mean, mode or various percentiles and these may only be elicited indirectly in some cases. It is argued in the literature on the subject that more accurate judgements are made on observable quantities and not on moments of distributions because experts are not likely to be familiar with statistical constructs and making direct judgements about them.[25,26,42]

In their review of their experiences in expert elicitation, Kadane and Wolfson (1998) highlighted four issues to be aware of when questioning experts:[42]

1. experts link their judgements to the frequency with which they can recall events (availability),

2. expert judgements can be based around some initial judgement and subsequent judgements tend to be based around that judgement (anchoring),

3. experts are known to struggle with assessing rare events which is needed to form the tails of a probability distribution (overconfidence),

4. the expert's judgements might already have been influenced by other data that are to be treated separately in subsequent analyses (hindsight bias).

The first issue can be dealt with to some extent by reminding experts of available background information and by asking experts to share their experiences. Asking experts to consider uncertainty prior to making best guesses can help to guard against anchoring and some methods have been developed to counteract this effect.[45] Overconfidence can be reduced by reminding the expert of extreme occurrences, by carefully selecting the judgements to be made and through training.[46,47] The issues of hindsight bias and solutions will be discussed later in this section in the context of using Bayesian methods. It should be noted all of these potential solutions are not fool-proof, and a skilled facilitator is needed to help identify when an expert is subject to these problems.

Many elicitation techniques depend on group meetings and discussions. Here to a get a group consensus that is reflective of the expertise present each individual's opinion must be

pooled together in a way in which everyone is satisfied. There are another number of issues that arise from these situations. A facilitator's role is to actively engage the entire group and to guard against domination of the proceedings by a small subset of the experts.[26] Experience suggests that questioning phases of elicitation exercises can be more successful if the facilitator is on hand to guide the experts through the questioning stage. In an assessment of genotoxic carcinogens, experts were subjected two rounds of elicitation: one was conducted remotely and the other as a facilitated group at the same location. In this exercise, the experts were far happier with the results of the group meetings and several experts failed to complete the quantitative part in the first round.[39]

Of course, identifying and attempting to control for biases is congruent with NICE's guidance on collecting evidence for HTA.[1] Further discussion of these psychological issues from an expert elicitation standpoint can be found in several key texts.[24,26,42]

It is important that the expert judgements are made (and recorded) in such a way that there is no ambiguity in subsequent interpretation. Some elicitation studies have been based around verbal descriptions of likely outcomes and associated uncertainties,[48,49] and the difficulties arise with consistency between experts and translation to something that is statistically useful. For example, saying that something is likely may mean around a 80% chance to one person and closer to 50% for another.

Once judgements about the quantity of interest have been elicited, the facilitator must convert them into something useful for the HTA; in most applications, this means conversion to a probability distribution. Most of the applications of elicitation employ parametric techniques. In a parametric technique, the facilitator fits the elicited summaries to a distribution that is a member of some specified parametric family (for instance, a normal or a beta distribution).[26] The elicited summaries from the questioning phase do not identify the expert's distribution uniquely: the facilitator must use their judgement to fit an appropriate probability distribution. Of course, facilitators are required to be flexible in the fitting process. The facilitator will not always be able to predetermine the distributional forms, and, if there are many parameters to be considered in one exercise, the facilitator should be prepared to use many different distributional forms. For example, in an elicitation exercise regarding ulceration treatment, beta and various transformed normal distributions were used to model the beliefs of the experts about the ten parameters.[41]

The final stage is in many ways the most important part of the elicitation process: it is essential that the fitted distribution is shared with the experts for comment. The results of the third stage should be presented to the experts so that they have the final say on whether the results fit with their beliefs. This is where the assumption of the fitted distribution being a valid representation of the experts' beliefs is judged. It is common to feedback different summaries from the distribution than the ones originally elicited.[43] If the experts do not agree with the results, then they should be able to revise or add to their earlier judgements, and this should be repeated until the experts are satisfied. Here the facilitator is trying to find a representation that the experts are willing to accept as their own (a satisficing distribution in the context of distribution elicitation[50]). It is conceivable that any number of distributions would be accepted by the experts as their distribution and attempts have been made to model this additional uncertainty.[51,52]

Due to the subjective nature of elicitation, transparency in the process is important for defensible results. It is recommended that a record should be made of any interviews

or workshops.[18,26] Such records should include details of experts present at the meeting, a summary of each expert's relevant expertise and declarations of interest obtained from them at the start of the elicitation process. The declarations of interest should not be used as grounds for exclusion from the elicitation exercise, but as another aid to transparency. In some elicitation exercises, there will be enough experts to ensure that the questioned experts have not got a stake in the results of the analysis,[19] but it should be noted that it will be common for experts to be stakeholders in the HTA.

Looking at the principles from a HTA perspective, a case study about eliciting expert judgements about bowel cancer treatment provisions highlighted several "points of good practice in eliciting and using expert judgement".[16] Distilling that list down, we find that the authors agree that expert training and participation in refining the elicitation process is important and that capturing assumptions and recording the steps in the elicitation exercise are great aids for users of the elicitation results. Very recent work on evaluating the cost effectiveness of e-prescribing systems showcases the embedding of the key principles described in this section into a HTA elicitation protocol.[19] Considerations are made about appropriate training for the experts, providing opportunities for feedback and revision of judgements and video recording of expert sessions has been considered.

As probabilistic analysis forms an integral part of HTA,[1,2] the most common use of expert-derived distributions will be in uncertainty propagation through some model of the impacts of the health technology under consideration[53,54] and to inform sensitivity analyses of those models.[55] For example, this is demonstrated in the HTA literature in the economic evaluation of DNA testing,[44] in the design of clinical trials[56] and in the evaluation of methods for detecting cervical cancer recurrence.[57] Also relevant to HTA is the inclusion of elicitation results about uncertainty in functions. A recent study of surgeons' beliefs about operation times as experience increases for two surgical techniques led to uncertainty about the functional response being modelled.[58] Part of this elicitation exercise had surgeons making direct judgements on the parameters of a curve (time taken for first procedure and number of procedures when operation time will plateau). The surgeons gave judgements about medians and interquartile ranges for the parameters; however, only a aggregation mean was used in subsequent analysis instead of the propagating all of the uncertainty.

Bayesian methods offer a natural and rational way of incorporating expert knowledge into an analysis through the use of informative prior distributions.[59,60] Bayesian methods start with uncertainty about some quantity encoded as a prior distribution, and this distribution is updated after the receipt of new data. It is important to consider hindsight bias when using expert judgements to inform Bayesian analyses. If the data that is being used to update the prior distributions has been seen by the experts, the Bayesian approach is not really valid. In this case, the analyst must consider whether the experts will have correctly synthesised the data in their judgements or if more independent data is required. Expert knowledge has been applied in HTA using Bayesian methods for the development of medical devices,[61] to inform future clinical trials[62,63] and to assess the effects of drugs[34] to name a few. Expert judgements are also required in Bayesian weight-of-evidence procedures where the influence of data sources on the quantity of interest need to be elicited.[64] Such considerations for weight-of-evidence are not limited to Bayesian approaches and there are many other less formal uses of expert judgements for weight-of-evidence.[65]

Other quantitative techniques exist for the incorporation of expert knowledge. However, these techniques are not always compatible with the probabilistic assessments required for HTA by NICE.[1] For instance, Dempster Shafer theory is the basis of an expert-driven multi-criteria decision tool[66] and fuzzy logic has been used to help model uncertainty in project evaluation with inputs from domain experts.[67]

Away from quantification, experts can also be used in an indirect way to sense check models, analyses and data sources.[68,69] Here judgements tend to be more *ad hoc* and used to exclude components.[70] A recent study into errors in HTA models reveals the utility of experts for such identification and refinement processes.[71]

In this section, we have covered some of the key principles of expert elicitation at great lengths and have avoided practical issues. There is no perfect elicitation method or protocol; therefore, it is important to understand the desirable features of a defensible elicitation and the potential limitations prior to designing an elicitation exercise. An in depth discussion of these principles is given in several seminal works on expert elicitation[26,72,73] and a discussion from a policy perspective is given in a recent US Environment Protection Agency white paper.[74] In the next section, we move on to a discussion of specific methods, some practicalities and the methods' application in HTA.

3 Popular methods in expert elicitation

There are a number of methods that are widely used to elicit expert beliefs about single values. As already stated, experts are rarely asked (or able to) to specify their complete distribution (for example, a log-normal distribution with given mean and variance). However, the roulette (or histogram) method is becoming popular. In this method, the experts are asked to build up a probability density by allocating "chips" that represent blocks of probability to predefined intervals for the parameter of interest.[41,44,75,76] Successful implementation of this approach depends heavily on a good appreciation of probability density functions and can lead to poorly assessed distributional tails. A compromise is to ask the expert to make direct probability judgements or judgements about percentiles.[43]

A simple alternative method is to ask for a best estimate and a range with a specified probability (usually 90 % or 95%) of the true value being covered by the range.[56] The bisection method is similar as the expert is asked for the median of their distribution and then the quartiles in an indirect and unambiguous manner.[77] Such methods are prone to biases due to anchoring around the initial estimate. To help counteract this, the double bisection was developed to force the experts to think more carefully about their judgements.[45] However, it is unclear what effect the burden of the extra judgements would have if the expert was making judgement about many quantities. Another strategy is to remind the experts of extreme circumstances prior to eliciting information about the range or quartiles, but this can have the effect of over inflating the judged uncertainty.[47] There are of course many more methods for eliciting beliefs about single values, and many are reviewed in O'Hagan *et al.* (2006).[26]

Once judgements have been made the facilitator must chose how to fit an appropriate probability distribution. This can be done in many different ways. One approach is to use a least squares method to match the 'closest' distribution to the summaries specified by the expert.[43] This has the desirable feature of being able to handle more judgements

than there are parameters for the chosen distribution. A simple approach is to use a method of moments where the elicited summaries are matched to distribution parameters using known theoretical relationships,[41] but this can be difficult if the number of elicited judgements exceeds the number of distribution parameters. If a roulette or histogram method has been employed, some facilitators chose to abandon the parametric approach and directly sample from the histogram (which can be thought of as a mixture of non-overlapping uniform distributions).[76] Again, this is the facilitator's choice and care should be taken to assess the potential impact of such choices.

Whatever elicitation method is used, it is likely that the facilitator will be faced with the problem of combining judgements or fitted distributions from multiple experts to reach a consensus distribution. Schemes for aggregating the judgements of several experts to form a consensus distribution can broadly be categorised as one of three approaches: mathematical, Bayesian or behavioural. A recent, more comprehensive review of aggregation techniques can be found in French (2011).[78]

Mathematical aggregation schemes usually involve the elicitation of each expert's judgements individually and the use of some mathematical rule to combine them. The simplest form of this is the linear opinion pool[79] where the individuals' probability distributions are summed with equal weight being given to each expert. Although giving equal weight to each expert seems democratic, it may lead to undesirable consequences such as favouring one school of thought or the inclusion of an individual who may not be able to contribute much. Research into appropriate weighting schemes is ongoing[80,81] and arguments persist that a democratic technique for producing consensus distributions is the only viable method.[82] A recent elicitation exercise where mathematical aggregation was used had 18 experts making judgements about the sensitivity and specificity of diagnosis devices.[83] Here, the experts' individual judgements were weighted using a score based on years of experience and the frequency of carrying out diagnoses using similar technology. The scores are easily reproducible, but there are serious questions about experience being accumulated in a linear fashion and the relative importance of general experience over specific experience of applying a particular method.

A more principled approach to mathematical aggregation is for the facilitator to model the expert's judgements in a Bayesian fashion: typically, the facilitator will have very weak prior information on the quantity of interest and the expert judgements will be treated as data to update the facilitator's beliefs.[84–86] Just like in mathematical aggregation, the facilitator must carefully consider the influence that each expert will have on the consensus distribution. The Bayesian approach is very similar to the random effects meta-analysis approach that has been applied in the context of active psoriatic arthritis treatment: experts are treated as inputs into a statistical model to produce an overall distribution.[76]

Many applications of expert elicitation techniques avoid mathematical aggregation altogether in order to elicit the consensus distribution directly from the entire group.[47,87] Such methods require the all participating experts to be present for the entire elicitation exercise. These types of approaches can be broadly classified as behavioural approaches.[78] In these approaches, it is of paramount importance to have a skilled facilitator who can handle group dynamics.[88]

A recent publication by the European Food Safety Authority (EFSA)[89] highlighted three expert elicitation protocols for use in their assessments and gave clear recommen-

dations on their use. The three protocols are Cooke's classical method,[73] the Sheffield elicitation framework (SHELF)[90] and a novel, modified Delphi scheme that incorporates SHELF principles. The first two are well-established methods for eliciting and combining expert judgements. The modified Delphi scheme extends the SHELF method in such a way that it could be useful when it is difficult to arrange a meeting with all the relevant experts present.

SHELF provides a framework for capturing information about an elicitation exercise including the experts' backgrounds and potential conflicts of interest and any reasoning or key sources of information that underpins expert judgements. This is done through a comprehensive set of questions that are designed to cover everything a user of the elicitation exercise results needs to know before using them (including results on any training exercises). SHELF is set up to cover a variety of univariate elicitation techniques including the roulette, bisection and range methods mentioned earlier in this section and uses least squares fitting to model the experts' probability distributions.

A SHELF exercise is designed to be carried out by a group led by a facilitator. The individual experts are asked to make their own quantitative judgements after the quantity of interest has been discussed and then a group consensus distribution is suggested by the facilitator using linear opinion pooling with equal weights. Both the individual fitted distributions and the consensus distribution are displayed to the group and discussions are encouraged as to whether the consensus distribution is valid. This final step is a behavioural approach to aggregation that uses mathematical aggregation as a point of departure. Throughout the experts, have opportunities to contribute and to revise their judgements, but a skilled facilitator is needed to avoid the problems of group interactions.

SHELF has been developed with the key principles described in Section 2 in mind and has associated software to help facilitators without the required statistical expertise. We will discuss the associated software and documentation for the implementation of SHELF in Section 5. SHELF has been implemented for the elicitation of expert beliefs for veterinary treatments[91] and, in the context of HTA, to help design clinical trials.[17] In a recent exercise to inform decisions around the treatment of persistent cervical cancer, a similar scheme was used where experts were asked to declare their background and relevant expertise prior to making judgements.[57] In that exercise, a type of roulette method was used to elicit individual judgements with a paper based exercise with no group interaction for the formal elicitation part and mathematical aggregation was done without the experts intervention.

The linear opinion pool and the subsequent behavioural aggregation stage in SHELF gives rise to questions about whether it is desirable to equally weight the experts in the shared group distribution and what effect dominant personalities could have on reaching the consensus distribution. Cooke's classical method was designed to try to circumnavigate these kinds of problems, to make expert elicitation more reproducible and to take account of the experts' ability to make judgements about uncertainty.[73,92]

The classical method is an elicitation protocol that attempts to use expert performance on a set of related elicitation questions (called seed questions) where answers are known to weigh the experts' contribution to a opinion pool (this has been extended to where answers have a well-established uncertainty associated with them[76]). The guidance on the number of seed questions varies, but most applications have around five to help avoid fatigue in the questioning process.[93] Each expert records their own answers to questions about the

seed quantities and the quantities of interest although the facilitator may organise a group meeting to go over the aims of the exercise and to discuss background knowledge. The judgements about the quantities of interest are often made using the best estimate and range technique, but facilitators are not bound by this, and fitting can proceed using least-squares (or similar) fitting. The judgements are combined through a linear opinion pool with weights determined by the experts' performance on the seed questions and the level of certainty displayed in their answers. The idea behind the latter is that the more certain someone is the more expertise they have. Of course, it is still important for this method that the facilitator guards against typical overconfidence. This method has been applied in HTA for the assessment of treatments for active psoriatic arthritis.[76]

Proponents argue that the classical method discourages over- and under-confidence in expert judgements and that the resulting weightings are not only practical: they are also theoretically sound.[94] Opponents of the method argue that producing seed questions that are informative enough could often be impossible and that performance in making judgements on a handful of seed questions might not be representative of expert performance.[82] Despite these contrasting views, there have been many applications of these techniques in high impact policy areas[95–97] and the facilitators of such exercises claim that experts buy in to the process.[93,94] Others have suggested alternative ways of weighting expert opinion based around procedures for scoring judgements based on self-ratings, peer-ratings or facilitator ratings.[86,98,99]

The final protocol mentioned in the EFSA guidance on expert elicitation is the modified Delphi method. The Delphi method is well-known technique for capturing expert opinion.[100–102] A Delphi exercise involves involves several rounds of questionnaires where each expert has access to the opinions of the other participants. A recent priority setting exercise for methodological research into clinical trials used the Delphi method to elicit judgements from 41 experts.[103] The aim of this exercise was to identify and rank areas for future methodological research. This type of technique has little relevance for finding values for quantitative models, but the results of this exercise shows that experts in healthcare are willing to engage with a Delphi-style exercise.

In order to make the Delphi technique more relevant to probabilistic assessments, it has been suggested that the SHELF method could be embedded within a Delphi-like process.[89] Here the idea is to use the SHELF process to capture all the relevant information about the elicitation exercise whilst avoiding the need to meet. The experts are asked to make their judgements separately and then have several rounds of revisions based on the other expert's judgement and the fitted consensus distribution. Like SHELF, the relative simplicity of technique makes it relatively easy to produce software (see Section 5), but given the remote nature of these exercises and lack of facilitator interaction, great care must go into software design.

4 Other methods in expert elicitation

The EFSA guidance on expert elicitation[89] focussed on just three elicitation protocols because there has been limited effort on developing generic protocols that will be fit for capturing probabilistic judgements in a wide range of applications. However, there are numerous examples of *ad hoc* methods for modelling expert knowledge probabilistically and

of methods being developed to capture expert knowledge where probability distributions are not the focus.

One area where several such techniques have been used is in elicitation of judgements about dependencies. In their guidelines for technology assessment, NICE state that "evidence about the extent of correlation between individual parameters should be carefully considered and reflected in the probabilistic analysis".[1] Despite this, there is evidence to suggest that even experienced facilitators of expert elicitation exercises are reluctant to attempt to elicit multivariate probability structures due to the complexity and the lack of understanding from the experts.[38] Where multivariate structure is incorporated in subsequent analyses, independence is often assumed to avoid this issue (for example, see the fitted joint distributions from an assessment of left-ventricular assist devices[33]). Of course, there are cases where problems can be structured to get independence (or at least conditional independence), which can significantly reduce the number of judgements needed about dependencies.[64,104]

There are a few cases where expert elicitation have been used for model parameters are constrained. For instance, when they must sum to 100%.[105,106] There has been more success in eliciting correlation structures in the context of Bayes linear methods, which do not rely on the specification of full probability distributions.[64,107,108] However, research into assessing dependencies between parameters of interest provides little consistent guidance as to what procedure is best or what is realistic in terms of making expert judgements.[109–111]

A more indirect route can be used to elicit multivariate structure from experts. The probabilistic inversion method uses rankings of multi-attribute items to infer multivariate dependencies.[112–114] This type of method has some desirable properties: it shares features with multidimensional criteria analysis and it is believed that experts are more reliable in providing rankings rather than making direct judgements about variables. Similar principles are used in analytic hierarchy process, which is a structured way for experts to consider multiple attributes in a pairwise comparison scheme leading directly to a decision.[115,116]

Sometimes it is unrealistic to expect the experts to make quantitative judgements about all the relevant uncertainties. When considering impact of technologies and possible potential future technologies, the decision maker is faced with some uncertainties that are known and some that can never be determined.[117] It can be false to claim that everything is quantifiable and in some situations more inclusive, qualitative approaches are required. In particular, methods like scenario methods, horizon scanning, focus and dissensus groups and multicriteria mapping are encouraged to help explore the problem and its potential impacts.[118] This type of approach was used for the HTA for the older part of the population where 4304 stakeholders were (indirectly) questioned about requirements fro future technologies.[119] Another approach to handling uncertainties that cannot be quantified is to list them and qualitatively assess their potential impact.[120] Such an approach can aid transparency because the experts are forced to highlight areas that they have not considered quantitatively and this can be helpful for directing future research.

Another popular qualitative method to capture expert knowledge is the nominal group technique.[102,121,122] This technique is used to identify issues and potential solutions and results in a ranking of the issues and solutions. It is widely used because it is easy to administer, produces results quickly and is designed to give everyone an equal voice. It was recently employed in the context of personalised medicine where 47 experts were questioned

about economic evaluation.[123] The downside to this method is that its strict protocol is not conducive to eliciting knowledge about values and associated uncertainties. However, some of the key features would be useful for eliciting judgement about values: having all experts contribute and allowing the group to share ideas. Other examples of group consensus techniques that share principles with probabilistic expert elicitation methods are given in a number of reviews.[124,125]

5 Existing software tools and resources

Given the amount of research that has been done on expert elicitation and the number of applications, there are surprisingly few resources readily available for applications that are fit for purpose in HTA. This is due in part to the tailored nature of an expert elicitation exercise. Despite this, using software in expert elicitation exercises is seen as beneficial because there is strong evidence to suggest that experts benefit from having instantaneous feedback that software programs can provide.[126,127] There have been some notable attempts to provide software for expert elicitation exercises, and we will briefly review them in this section.

In Section 3, the elicitation protocol behind SHELF was described. This protocol has been implemented through the provision of generic report templates and R functions[90] (see `www.tonyohagan.co.uk/shelf/`). The report templates for capturing information about the elicitation exercise are provided in both portable document and Microsoft Word formats. R is a widely used and free R software environment for statistical computing.[128] In order to implement the SHELF protocol no knowledge of R is required: firstly, because there is a user manual with worked examples for SHELF and, secondly, because SHELF takes advantage of a graphical user interface within R. Although this is computer-based, the expectation is that interaction with the software is done as a group to aid the behavioural aggregation stage.

As already discussed, it is not always possible to get the experts together at the same time. However, internet connectivity allows for online applications to be developed that can aid the remote capture of expert judgements. The R program behind the SHELF implementation has been modified for remote use in the online MATCH uncertainty elicitation tool[129] (see `optics.eee.nottingham.ac.uk/match/uncertainty.php`). A similar protocol to SHELF has also been implemented as part of the UncertWeb project[130] (see `elicitator.uncertweb.org/`). In this implementation, facilities are provided to capture all the briefing documents for an elicitation exercise, to invite experts to participate and to keep track of the progress of all the experts during the exercise.

Remote expert elicitation exercises are also routinely done via email. Outside of probabilistic expert elicitation exercises software has been developed for the application of the Delphi method and there is scope for adaptation to probabilistic applications in the future.[36,131,132] These could be especially valuable to facilitators looking to apply a behavioural aggregation method. SHELF has been adapted to an Microsoft Excel spreadsheet for several different applications and has been piloted for capturing disease prevalence[37] and for food safety assessments (which is in fact a modified Delphi technique).[89] Other spreadsheet-based elicitation exercises have been used to capture expert judgements in the context of HTA.[44,76,83] The flexibility of modern spreadsheet software means that the nec-

essary information can be easily recorded and instantaneous feedback can be provided in bespoke ways with little knowledge of graphical user interface development. Also, health experts are likely to be familiar with spreadsheets and they (and the facilitators) may even use them for some of their probabilistic modelling. However, it is important that the experts buy-in to an elicitation exercise and a long spreadsheet-based questionnaire may not be ideal.

Cooke's classical method has been implemented in the EXCALIBUR package for Windows (see `www.lighttwist.net/wp/excalibur`). It is designed for the facilitator of an expert elicitation exercise to input experts' judgements and combine their assessments based on equal weights, user weights or expert performance-based weights. Although the software is dated, it provides a number of diagnostic tools that are of use to the facilitator: for example, the facilitator can see how sensitive the consensus distribution is to the choice of experts and choice of seed questions.

Designers of elicitation software often like to have the experts interacting with graphical interfaces and getting instantaneous feedback.[130,133,134] Such an approach tallies with the principles of giving experts ownership of the exercise results because they will see their judgements being translated on screen and the principle of feedback of fitting consequences as discussed in Section 2.

A 2011 report into expert elicitation software by the Dutch National Institute for Public Health and the Environment identified 65 pieces of software with some elicitation functionality.[135] All had aspects that could be used to capture judgements from a single expert for a single expert, almost all could be used to capture and aggregate judgements from multiple experts, but little attention was paid to the capture of an entire expert elicitation exercise. For univariate elicitation problems, there should be few obstacles to building a multi-purpose software program to record and perform the necessary calculations for distribution fitting and aggregation. For multivariate elicitation problems, the difficulty is that there are no techniques for capturing assessments that would be considered to be valid across all applications. However, both the Prior Elicitation Graphical Software and Unicorn offer partial solutions to capturing knowledge about multivariate parameters (see `statistics.open.ac.uk/elicitation` and `www.lighttwist.net/wp/unicorn-download` respectively). It has also been asserted that expert elicitation exercises are more successful when elicitation software is tailored to the problem in hand especially in terms of giving problem-specific feedback.[127,136] In the context of HTA, such an interactive software program has been used to evaluate bowel cancer service.[16]

6 Discussion and recommendations

In the context of HTA, Leal *et al.* (2007) concluded that "there is clearly a gap in the literature between theoretical elicitation techniques and tools that can be used in applied decision-analytic models".[44] Expert elicitation methods seem to exist for most (univariate) HTA applications, but clear guidance on how to carry out an elicitation exercise from inception to reporting seems to be missing for HTA. One issue, as stated by Grigore *et al.* (2013), is that evaluation of expert elicitation protocols for HTA is done in an *ad hoc* manner.[18] This means that it is difficult for potential users of the protocols to judge the utility of elicitation protocols. An important step in improving this is to encourage

(or force) facilitators of expert elicitation exercises to keep comprehensive records of the exercises. This will aid both transparency and defensibility for the problem in hand and also help in trying to standardise and improve elicitation applications in HTA.

There continue to be concerns about the overall validity of expert elicitation exercises and the ability of the experts to make useful judgements.[137] Results from an expert elicitation exercise are a snap shot of opinions at a moment in time. As experts gain experience in their area of expertise and in making judgements about uncertainty, there performance will change. It has been noted that weather forecasters become very well calibrated,[26] but for most HTA relevant elicitation exercises the variables being judged might never be realised and feedback in terms of long-term evidence is far from instantaneous. As such, validation of expert elicitation exercises is problematic. There have been some notable attempts to demonstrate the performance of expert judgements, but there is evidence that routine judgements that might be considered in such studies are not the same as judgements made about policy-relevant quantities.[138]

There is also little attention given to the situation where experts are giving conflicting judgements. It is clearly not preferable to average across wildly contradictory judgements and using bimodal distributions may result in analyses that no one person believes in,[32] but would it be acceptable for an HTA to have several conclusions each based on judgements from different schools of thought?

On the more technical side of elicitation, it is clear that more research is needed in to the accurate elicitation of beliefs about dependencies and multivariate model parameters.[110,139] This is important for HTA as it is recommended that dependencies are properly considered in the assessment. Also, there is a gap for rapid expert elicitation protocols. The ideal situation of having a pool of experts at a decision makers finger tips ready to engage in making expert judgements has yet to be realised.[130] In fact, many of the protocols and methods described in this piece require a significant amount of time to implement and significant input from the experts (although there is a suggestion that some of the methods can be applied in emergency situations[35,94]). Further research is required into the reliability of rapidly-implemented elicitation exercises, but having some prior ideas about reliable elicitation protocols could aid this.

For all of these problems, the answer to the question of which area is most profitable for research now lies in the interpretation of expert elicitation results by users of HTAs. Discussions about expectations for elicitation exercises between experts in the process of elicitation, health technology assessors and policy makers could lead to the standardisation of expert elicitation in HTA and relevant methodological developments. This could be achieved by giving extensive, prescriptive guidelines that are aimed at producing expert elicitation exercises that are fit for purpose. At the very least, such a dialogue could help produce the criteria from which a elicitation exercise can be judged.

References

[1] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. *NICE: process and methods guides*, 2013.

[2] Karl Claxton, Mark Sculpher, Chris McCabe, Andrew Briggs, Ron Akehurst, Martin Buxton, John Brazier, and Anthony O'Hagan. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health economics*, 14(4):339–347, 2005.

[3] Karl Claxton. Exploring uncertainty in cost-effectiveness analysis. *Pharmacoeconomics*, 26(9):781–798, 2008.

[4] Michael F Drummond, J Sanford Schwartz, Bengt Jönsson, Bryan R Luce, Peter J Neumann, Uwe Siebert, and Sean D Sullivan. Key principles for the improved conduct of health technology assessments for resource allocation decisions. *International journal of technology assessment in health care*, 24(03):244–258, 2008.

[5] Peter J Neumann, Michael F Drummond, Bengt Jönsson, Bryan R Luce, J Sanford Schwartz, Uwe Siebert, and Sean D Sullivan. Are key principles for improved health technology assessment supported and used by health technology assessment organizations? *International journal of technology assessment in health care*, 26(01):71–78, 2010.

[6] Jennie Popay and Gareth Williams. Public health research and lay knowledge. *Social science & medicine*, 42(5):759–768, 1996.

[7] Christine Putland, Fran E Baum, and Anna M Ziersch. From causes to solutions-insights from lay knowledge about health inequalities. *BMC public health*, 11(1):67, 2011.

[8] Jill Thompson, Paul Bissell, Cindy Cooper, Chris J Armitage, and Rosemary Barber. Credibility and the professionalizedlay expert: Reflections on the dilemmas and opportunities of public involvement in health research. *Health:*, page 1363459312441008, 2012.

[9] Melvin M Mark and R Lance Shotland. Stakeholder-based evaluation and value judgments. *Evaluation Review*, 9(5):605–626, 1985.

[10] Ken Stein, Matthew Dyer, Tania Crabb, Ruairidh Milne, Alison Round, Julie Ratcliffe, and John Brazier. A pilot internet value of health panel: recruitment, participation and compliance. *Health and quality of life outcomes*, 4(90), 2006.

[11] Yvonne Bombard, Julia Abelson, Dorina Simeonov, and Francois-Pierre Gauvin. Eliciting ethical and social values in health technology assessment: A participatory approach. *Social science & medicine*, 73(1):135–144, 2011.

[12] Shepley Orr, Jonathan Wolff, and Stephen Morris. What values should count in HTA for new medicines under value based pricing in the UK. In *Health Economists Study Group Conference, Bangor University, Bangor*, 2011.

[13] Michael Drummond, Rosanna Tarricone, and Aleksandra Torbica. Assessing the added value of health technologies: reconciling different perspectives. *Value in Health*, 16(1):S7–S13, 2013.

[14] David J Spiegelhalter, Jonathan P Myles, David R Jones, and Keith R Abrams. Methods in health service research: an introduction to Bayesian methods in health technology assessment. *BMJ: British Medical Journal*, 319(7208):508, 1999.

[15] Alan J Girling, Richard J Lilford, David A Braunholtz, and Wayne R Gillett. Sample-size calculations for trials that inform individual treatment decisions: a true-choiceapproach. *Clinical Trials*, 4(1):15–24, 2007.

[16] Paul H Garthwaite, James B Chilcott, David J Jenkinson, and Paul Tappenden. Use of expert knowledge in evaluating costs and benefits of alternative service provisions: A case study. *International journal of technology assessment in health care*, 24(03):350–357, 2008.

[17] Nelson Kinnersley and Simon Day. Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. *Pharmaceutical statistics*, 12(2):104–113, 2013.

[18] Bogdan Grigore, Jaime Peters, Christopher Hyde, and Ken Stein. Methods to elicit probability distributions from experts: A systematic review of reported practice in health technology assessment. *PharmacoEconomics*, 31(11):991–1003, 2013.

[19] Richard J Lilford, Alan J Girling, Aziz Sheikh, Jamie J Coleman, Peter J Chilton, Samantha L Burn, David J Jenkinson, Laurence Blake, and Karla Hemming. Protocol for evaluation of the cost-effectiveness of ePrescribing systems and candidate prototype for other related health information technologies. *BMC health services research*, 14(1):314, 2014.

[20] Paul Slovic. Psychological study of human judgment: Implications for investment decision making. *The Journal of Finance*, 27(4):779–799, 1972.

[21] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[22] Robin M Hogarth. Cognitive processes and the assessment of subjective probability distributions. *Journal of the American statistical Association*, 70(350):271–289, 1975.

[23] George Ed Wright and Peter Ed Ayton. *Subjective probability.* John Wiley & Sons, 1994.

[24] Mary Kynn. The heuristics and biases bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1):239–264, 2008.

[25] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[26] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities.* John Wiley & Sons, 2006.

[27] Robert L Winkler. Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, 66(336):675–685, 1971.

[28] Lawrence D Phillips. Group elicitation of probability distributions: are many heads better than one? *Decision Science and Technology*, pages 313–330, 1999.

[29] J Scott Armstrong. Combining forecasts. *Principles of forecasting*, pages 417–439, 2001.

[30] Shi-Woei Lin and Chih-Hsing Cheng. The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management*, 4(2):149–161, 2009.

[31] Joseph B Kadane. Progress toward a more ethical method for clinical trials. *Journal of Medicine and Philosophy*, 11(4):385–404, 1986.

[32] Robert T Clemen and Robert L Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

[33] Alan J Girling, Guy Freeman, Jason P Gordon, Philip Poole-Wilson, David A Scott, and Richard J Lilford. Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy. *International journal of technology assessment in health care*, 23(02):269–277, 2007.

[34] Sindhu R Johnson, John T Granton, George A Tomlinson, Haddas A Grosbein, Gillian A Hawker, and Brian M Feldman. Effect of warfarin on survival in scleroderma-associated pulmonary arterial hypertension (SSc-PAH) and idiopathic PAH. belief elicitation for Bayesian priors. *The Journal of rheumatology*, 38(3):462–469, 2011.

[35] David Mendonça, Robert Rush, and William A Wallace. Timely knowledge elicitation from geographically separate, mobile experts during emergency response. *Safety Science*, 35(1):193–208, 2000.

[36] Siddhartha Dalal, Dmitry Khodyakov, Ramesh Srinivasan, Susan Straus, and John Adams. Expertlens: A system for eliciting opinions from a large pool of non-collocated experts with diverse knowledge. *Technological Forecasting and Social Change*, 78(8):1426–1444, 2011.

[37] Daniel Sperber, Duncan Mortimer, Paula Lorgelly, and David Berlowitz. An expert on every street corner? Methods for eliciting distributions in geographically dispersed opinion pools. *Value in Health*, 16(2):434–437, 2013.

[38] Kathryn Chaloner and Frank S Rhame. Quantifying and documenting prior beliefs in clinical trials. *Statistics in medicine*, 20(4):581–600, 2001.

[39] Alan Boobis, Villie Flari, John Paul Gosling, Andy Hart, Peter Craig, Lesley Rushton, and Ehi Idahosa-Taylor. Interpretation of the margin of exposure for genotoxic carcinogens–elicitation of expert knowledge about the form of the dose response curve at human relevant exposures. *Food and Chemical Toxicology*, 57:106–118, 2013.

[40] M Granger Morgan, Peter J Adams, and David W Keith. Elicitation of expert judgments of aerosol forcing. *Climatic Change*, 75(1-2):195–214, 2006.

[41] Marta O Soares, Laura Bojke, Jo Dumville, Cynthia Iglesias, Nicky Cullum, and Karl Claxton. 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*, 30(19):2363–2380, 2011.

[42] Joseph Kadane and Lara J Wolfson. Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.

[43] Anthony O'Hagan. Eliciting expert beliefs in substantial practical applications. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):21–35, 1998.

[44] José Leal, Sarah Wordsworth, Rosa Legood, and Edward Blair. Eliciting expert opinion for economic models: an applied example. *Value in Health*, 10(3):195–203, 2007.

[45] Paul H Garthwaite and James M Dickey. Double- and single-bisection methods for subjective probability assessment in a location-scale family. *Journal of econometrics*, 29(1):149–163, 1985.

[46] Andrew Speirs-Bridge, Fiona Fidler, Marissa McBride, Louisa Flander, Geoff Cumming, and Mark Burgman. Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, 30(3):512–523, 2010.

[47] John Paul Gosling, Andy Hart, David C Mouat, Mirzet Sabirovic, Simon Scanlan, and Alick Simmons. Quantifying experts uncertainty about the future cost of exotic diseases. *Risk Analysis*, 32:881–93, 2011.

[48] Silja Renooij and Cilia Witteman. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22(3):169–194, 1999.

[49] Jose AA Andrade and John Paul Gosling. Predicting rainy seasons: quantifying the beliefs of prophets. *Journal of Applied Statistics*, 38(1):183–193, 2011.

[50] Robert L Winkler. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association*, 62(319):776–800, 1967.

[51] John Paul Gosling, Jeremy E Oakley, and Anthony O'Hagan. Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Analysis*, 2(693-718), 2007.

[52] Jeremy E Oakley and Anthony O'Hagan. Uncertainty in prior elicitations: a nonparametric approach. *Biometrika*, 94(2):427–441, 2007.

[53] Stan Kaplan. 'Expert information' versus 'expert opinions'. another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliability Engineering & System Safety*, 35(1):61–72, 1992.

[54] Tim Bedford and Roger Cooke. *Probabilistic risk analysis: foundations and methods*. Cambridge University Press, 2001.

[55] Andrea Saltelli, Karen Chan, and E Marian Scott. *Sensitivity analysis*. Wiley, 2000.

[56] Matt D Stevenson, Jeremy E Oakley, Myfanwy Lloyd Jones, Alan Brennan, Juliet E Compston, Eugene V McCloskey, and Peter L Selby. 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*, 2009.

[57] C Meads, P Auguste, C Davenport, S Małysiak, S Sundar, M Kowalska, A Zapalska, P Guest, S Thangaratinam, P Martin-Hirsch, et al. Positron emission tomography/computerised tomography imaging in detecting and managing recurrent cervical cancer: systematic review of evidence, elicitation of subjective probabilities and economic modelling. *Health Technol Assess*, 17(12):1–323, 2013.

[58] Jonathan A Cook, Craig R Ramsay, Andrew J Carr, and Jonathan L Rees. A questionnaire elicitation of surgeons belief about learning within a surgical trial. *PloS one*, 7(11):e49178, 2012.

[59] William M Bolstad. *Introduction to Bayesian statistics*. John Wiley & Sons, 2007.

[60] Peter M Lee. *Bayesian statistics: an introduction*. John Wiley & Sons, 2012.

[61] Laura Vallejo-Torres, Lotte MG Steuten, Martin J Buxton, Alan J Girling, Richard J Lilford, and Terry Young. Integrating health economics modeling in the product development cycle of medical devices: a Bayesian approach. *International journal of technology assessment in health care*, 24(04):459–464, 2008.

[62] David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles. Bayesian approaches to clinical trials and health-care evaluation. 2004.

[63] Lisa Hampson, John Whitehead, Despina Eleftheriou, Catherine Tudur-Smith, Rachel Jones, David Jayne, Helen Hickey, and Paul Brogan. Elicitation of expert prior opinion: application to the mypan trial in childhood polyarteritis nodosa. *Pediatric Rheumatology*, 12(1):1–1, 2014.

[64] John Paul Gosling, Andy Hart, Helen Owen, Michael Davies, Jin Li, and Cameron MacKay. A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. *Bayesian Analysis*, 8(1):169–186, 2013.

[65] Douglas L Weed. Weight of evidence: a review of concept and methods. *Risk Analysis*, 25(6):1545–1557, 2005.

[66] Malcolm Beynon, Darren Cosker, and David Marshall. An expert system for multi-criteria decision making using dempster shafer theory. *Expert Systems with Applications*, 20(4):357–367, 2001.

[67] FT Dweiri and MM Kablan. Using fuzzy decision making for the evaluation of the project management internal efficiency. *Decision Support Systems*, 42(2):712–726, 2006.

[68] Robert G Sargent. Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation*, pages 130–143. Winter Simulation Conference, 2005.

[69] Jeroen P Van Der Sluijs, Matthieu Craye, Silvio Funtowicz, Penny Kloprogge, Jerry Ravetz, and James Risbey. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the nusap system. *Risk analysis*, 25(2):481–492, 2005.

[70] Dwayne Boyers, Xueli Jia, Mark Crowther, David James Jenkinson, Cynthia Mary Fraser, and Graham Mowatt. Eltrombopag for the treatment of chronic idiopathic (immune) thrombocytopenic purpura (ITP): A single technology appraisal. *Health Technology Assessment*, 2010.

[71] J Chilcott, P Tappenden, S Paisley, A Rawdin, M Johnson, and E Kaltenthaler. Choice and judgement in developing models for health technology assessment; a qualitative study. 2010.

[72] M Granger Morgan, Max Henrion, M Small, et al. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press Cambridge, 1990.

[73] Roger M Cooke. *Experts in uncertainty: opinion and subjective probability in science*. 1991.

[74] USEPA. Expert elicitation task force white paper. *Science and Technology Policy Council, U.S. Environmental Protection Agency*, 2011.

[75] Claire McKenna, C McDaid, S Suekarran, N Hawkins, K Claxton, K Light, M Chester, J Cleland, N Woolacott, and M Sculpher. Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis. 2009.

[76] Laura Bojke, Karl Claxton, Yolanda Bravo-Vergel, Mark Sculpher, Stephen Palmer, and Keith Abrams. Eliciting distributions to populate decision analytic models. *Value in Health*, 13(5):557–564, 2010.

[77] Howard Raiffa. *Decision analysis: introductory lectures on choices under uncertainty.* Addison-Wesley, 1968.

[78] Simon French. Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1):181–206, 2011.

[79] Mervyn Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32(4):1339–1342, 1961.

[80] James K Hammitt and Yifan Zhang. Combining experts judgments: Comparison of algorithmic methods using synthetic data. *Risk Analysis*, 33(1):109–120, 2013.

[81] Ani Guerdjikova and Klaus Nehring. Weighing experts, weighing sources: The diversity value. *Unpublished manuscript, Théma, Université de Cergy-Pontoise*, 2014.

[82] Fergus Bolger and Gene Rowe. The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 2014.

[83] Wieke Haakma. Expert elicitation to populate early health economic models of medical diagnostic devices in development. Master's thesis, University of Twente, 2011.

[84] Mike West. Modelling expert opinion. *Bayesian statistics*, 3:493–508, 1988.

[85] MP Wiper and LI Pettit. On improving a model for combining experts forecasts. *Bayesian Statistics 5 (eds. JM Bernardo et al.)*, pages 809–813, 1996.

[86] Isabelle Albert, Sophie Donnet, Chantal Guihenneuc-Jouyaux, Samantha Low-Choy, Kerrie Mengersen, and Judith Rousseau. Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7(3):503–532, 2012.

[87] Aaron R Dewispelare, L Tandy Herren, and Robert T Clemen. The use of probability elicitation in the high-level nuclear waste regulation program. *International Journal of Forecasting*, 11(1):5–24, 1995.

[88] Simon French, John Maule, and Nadia Papamichail. *Decision behaviour, analysis and support.* Cambridge University Press, 2009.

[89] European Food Safety Authority. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, 12(6), 2014.

[90] SHELF: the Sheffield elicitation framework (version 2.0). School of Mathematics and Statistics, University of Sheffield, Sheffield.

[91] HM Higgins, Ian L Dryden, and Martin J Green. A Bayesian elicitation of veterinary beliefs regarding systemic dry cow therapy: Variation and importance for clinical trial design. *Preventive veterinary medicine*, 106(2):87–96, 2012.

[92] Roger Cooke, Max Mendel, and Wim Thijs. Calibration and information in expert resolution; a classical approach. *Automatica*, 24(1):87–93, 1988.

[93] Roger M Cooke and Louis HJ Goossens. TU Delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5):657–674, 2008.

[94] Willy Aspinall. A route to more tractable expert advice. *Nature*, 463(7279):294–295, 2010.

[95] Roger M Cooke and Louis H J Goossens. Procedures guide for structural expert judgement in accident consequence modelling. *Radiation Protection Dosimetry*, 90(3):303–309, 2000.

[96] Roger M Cooke and Karen A Slijkhuis. Expert judgment in the uncertainty analysis of dike ring failure frequency. *Case Studies in Reliability and Maintenance*, pages 331–350, 2003.

[97] Julie JCH Ryan, Thomas A Mazzuchi, Daniel J Ryan, Juliana Lopez de la Cruz, and Roger Cooke. Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research*, 39(4):774–784, 2012.

[98] José M Marques, Elisabete M Robalo, and Susana A Rocha. Ingroup bias and the 'black sheep' effect: Assessing the impact of social identification and perceived variability on group judgements. *European Journal of Social Psychology*, 22(4):331–352, 1992.

[99] Mark A Burgman, Marissa McBride, Raquel Ashton, Andrew Speirs-Bridge, Louisa Flander, Bonnie Wintle, Fiona Fidler, Libby Rumpff, and Charles Twardy. Expert status and performance. *PLoS One*, 6(7):e22998, 2011.

[100] Norman Dalkey. An experimental study of group opinion: the Delphi method. *Futures*, 1(5):408–426, 1969.

[101] Gene Rowe and George Wright. Expert opinions in forecasting: the role of the Delphi technique. *Principles of forecasting*, pages 125–144, 2001.

[102] Jeremy Jones and Duncan Hunter. Qualitative research: consensus methods for medical and health services research. *Bmj*, 311(7001):376–380, 1995.

[103] Catrin Tudur Smith, Helen Hickey, Mike Clarke, Jane Blazeby, and Paula Williamson. The trials methodological research agenda: results from a priority setting exercise. *Trials*, 15(1):32, 2014.

[104] Peter S Craig, Michael Goldstein, Allan Seheult, and James Smith. Constructing partial prior specifications for models of complex physical systems. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):37–53, 1998.

[105] Fadlalla G Elfadaly and Paul H Garthwaite. Eliciting Dirichlet and Connor-Mosimann prior distributions for multinomial models. *Test*, 22(4):628–646, 2013.

[106] Rita Esther Zapata-Vázquez, Anthony O'Hagan, and Leonardo Soares Bastos. Eliciting expert judgements about a set of proportions. *Journal of Applied Statistics*, 41(9):1919–1933, 2014.

[107] Malcolm Farrow. Practical building of subjective covariance structures for large complicated systems. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(4):553–573, 2003.

[108] John Quigley, Kevin J Wilson, Lesley Walls, and Tim Bedford. A Bayes linear Bayes method for estimation of correlated event rates. *Risk Analysis*, 33(12):2209–2224, 2013.

[109] Robert T Clemen, Gregory W Fischer, and Robert L Winkler. Assessing dependence: Some experimental results. *Management Science*, 46(8):1100–1115, 2000.

[110] Alireza Daneshkhah and JE Oakley. Eliciting multivariate probability distributions. *Rethinking risk measurement and reporting*, 2010.

[111] M Revie, T Bedford, and L Walls. Evaluation of elicitation methods to quantify Bayes linear models. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 224(4):322–332, 2010.

[112] C Du, Dorota Kurowicka, and Roger M Cooke. Techniques for generic probabilistic inversion. *Computational statistics & data analysis*, 50(5):1164–1187, 2006.

[113] Rabin Neslo, Fiorenza Micheli, Carrie V Kappel, Kimberly A Selkoe, Benjamin S Halpern, and Roger M Cooke. Modeling stakeholder preferences with probabilistic inversion. *Real-Time and Deliberative Decision Making*, pages 265–284, 2009.

[114] Villie Flari, Qasim Chaudhry, Rabin Neslo, and Roger Cooke. Expert judgment based multi-criteria decision model to address uncertainties in risk assessment of nanotechnology-enabled food products. *Journal of Nanoparticle Research*, 13(5):1813–1831, 2011.

[115] Ernest H Forman and Saul I Gass. The analytic hierarchy process-an exposition. *Operations research*, 49(4):469–486, 2001.

[116] L Pecchia, F Crispino, and SP Morgan. A software tool to support the health technology assessment (HTA) and the user need elicitation of medical devices via the analytic hierarchy process (AHP). In *The International Conference on Health Informatics*, pages 292–295. Springer International Publishing, 2014.

[117] Denis Loveridge and Ozcan Saritas. Ignorance and uncertainty: influences on future-oriented technology analysis. *Technology Analysis & Strategic Management*, 24(8):753–767, 2012.

[118] Andy Stirling. Keep it complex. *Nature*, 468(7327):1029–1031, 2010.

[119] Maarten J Ijzerman and Lotte MG Steuten. Early assessment of medical technologies to inform product development and market access. *Applied health economics and health policy*, 9(5):331–347, 2011.

[120] Andy Hart, John Paul Gosling, and Peter Craig. A simple, structured approach to assessing uncertainties that are not part of a quantitative assessment. Technical report, The Food and Environment Research Agency., 2010.

[121] Andrew H Van de Ven and Andre L Delbecq. The nominal group as a research instrument for exploratory health studies. *American Journal of Public Health*, 62(3):337–342, 1972.

[122] Morris Gallagher, Tim Hares, John Spencer, Colin Bradshaw, and Ian Webb. The nominal group technique: a research tool for general practice? *Family Practice*, 10(1):76–81, 1993.

[123] Wolf Rogowski, Katherine Payne, Petra Schnell-Inderst, Andrea Manca, Ursula Rochau, Beate Jahn, Oguzhan Alagoz, Reiner Leidl, and Uwe Siebert. Concepts of personalizationin personalized medicine: Implications for economic evaluation. *PharmacoEconomics*, pages 1–11, 2014.

[124] Arlene Fink, Jacqueline Kosecoff, Mark Chassin, and Robert H Brook. Consensus methods: characteristics and guidelines for use. *American journal of public health*, 74(9):979–983, 1984.

[125] Fanny Bourrée, Philippe Michel, and Louis Rachid Salmi. Consensus methods: review of original methods and their main alternatives used in public health. *Revue d'epidemiologie et de sante publique*, 56(6):e13–e21, 2008.

[126] Kathryn Chaloner, Timothy Church, Thomas A Louis, and John P Matts. Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, 42:341–353, 1993.

[127] Allan James, Samantha Low Choy, and Kerrie Mengersen. Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25(1):129–145, 2010.

[128] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[129] David E Morris, Jeremy E Oakley, and John A Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4, 2014.

[130] Lucy Bastin, Matthew Williams, John P Gosling, Phuong Truong, Dan Cornford, Gerard Heuvelink, and Frederic Achard. Web based expert elicitation of uncertainties in environmental model inputs. *European Geosciences Union General Assembly 2011*, page 5384, 2011.

[131] Murray Turoff and Starr Roxanne Hiltz. Computer based Delphi processes. *Gazing into the oracle: The Delphi method and its application to social policy and public health*, pages 56–85, 1996.

[132] Charlotte Rietbergen, Rolf HH Groenwold, Herbert JA Hoijtink, Karl GM Moons, and Irene Klugkist. Expert elicitation of study weights for Bayesian analysis and meta-analysis. *Journal of Mixed Methods Research*, page 1558689814553850, 2014.

[133] Joseph B Kadane, James M Dickey, Robert L Winkler, Wayne S Smith, and Stephen C Peters. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854, 1980.

[134] Samantha Low-Choy, Allan James, Justine Murray, and Kerrie Mengersen. Elicitator: a user-friendly, interactive tool to support scenario-based elicitation of expert knowledge. *Expert Knowledge and Its Application in Landscape Ecology*, pages 39–67, 2012.

[135] JLA Devilee and AB Knol. Software to support expert elicitation: An exploratory study of existing software packages. *RIVM letter report 630003001*, 2011.

[136] Ian Vernon, Michael Goldstein, Richard G Bower, et al. Galaxy formation: a bayesian uncertainty analysis. *Bayesian Analysis*, 5(4):619–669, 2010.

[137] M Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20):7176–7184, 2014.

[138] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[139] Anthony O'Hagan. Research in elicitation. In S. K. Upadhyay, U. Singh, and D. K. Dey, editors, *Bayesian Statistics and its Applications*. Anamaya, New Delhi, 2006.