Structured expert elicitation for healthcare decision making: A practical guide

James Horscroft, Lumanity; Dawn Lee, PenTAG, The University of Exeter; Dina Jankovic, The University of York; Marta Soares, The University of York; Laura Bojke, The University of York.

Contents

This document

This document is based on the protocol for structured expert elicitation (SEE) funded by the Medical Research Council (MRC) written by Bojke et al.[1] and aims to provide a practical guide to SEE in the context of healthcare decision making. It was written as a collaboration between Lumanity, which regularly conducts SEE to support health technology assessment (HTA) and market access strategic planning, and the Centre for Health Economics at the University of York, which authored the original MRC protocol. The MRC protocol was developed based on a systematic review of elicitation methods and the practical challenges of implementing them in a healthcare decision making setting.

Choosing your approach

Which quantities of interest?

Often in health economic modelling, several model parameters will be associated with significant uncertainty. However, these will have varying levels of impact on the overall uncertainty of the model. Budgets, timelines, and the cognitive burden of the elicitation on experts will often limit the number of quantities of interest that can be elicited in a single SEE. Therefore, it is important to carefully prioritize your research questions, based on the following:

1. Expected impact: After considering empirical data of relevance and the potential for additional data collection, identify key remaining uncertainties that are likely to impact on decision making (either modelled and not modelled)
2. Feasibility: It may not be feasible to elicit certain quantities and be confident that these represent experts' beliefs e.g. non-observable or complex quantities. For example, health state utility values are not encountered in day-to-day clinical practice, and the long-term survival of patients receiving a novel intervention may be very difficult for experts to predict

For some model inputs it may be desirable to elicit multiple quantities of interest. There may be multiple patient subgroups and/or comparators for which model inputs will need to be elicited separately. There may also be a need to look at multiple time points for the same parameter, for example, long-term survival may require estimates at multiple timepoints to explore how hazards change over time.

How many quantities of interest?

The maximum number of quantities that can be included in a single SEE will depend on the approach. As a general rule of thumb, no more than two hours should be spent on the elicitation exercise for remote sessions. Where a group consensus workshop is used to aggregate judgements, this limits the number of quantities of interest to three or four, while more can be included if a simple method (e.g. bisection) is used to elicit individual judgements and resulting probability distributions are aggregated mathematically.

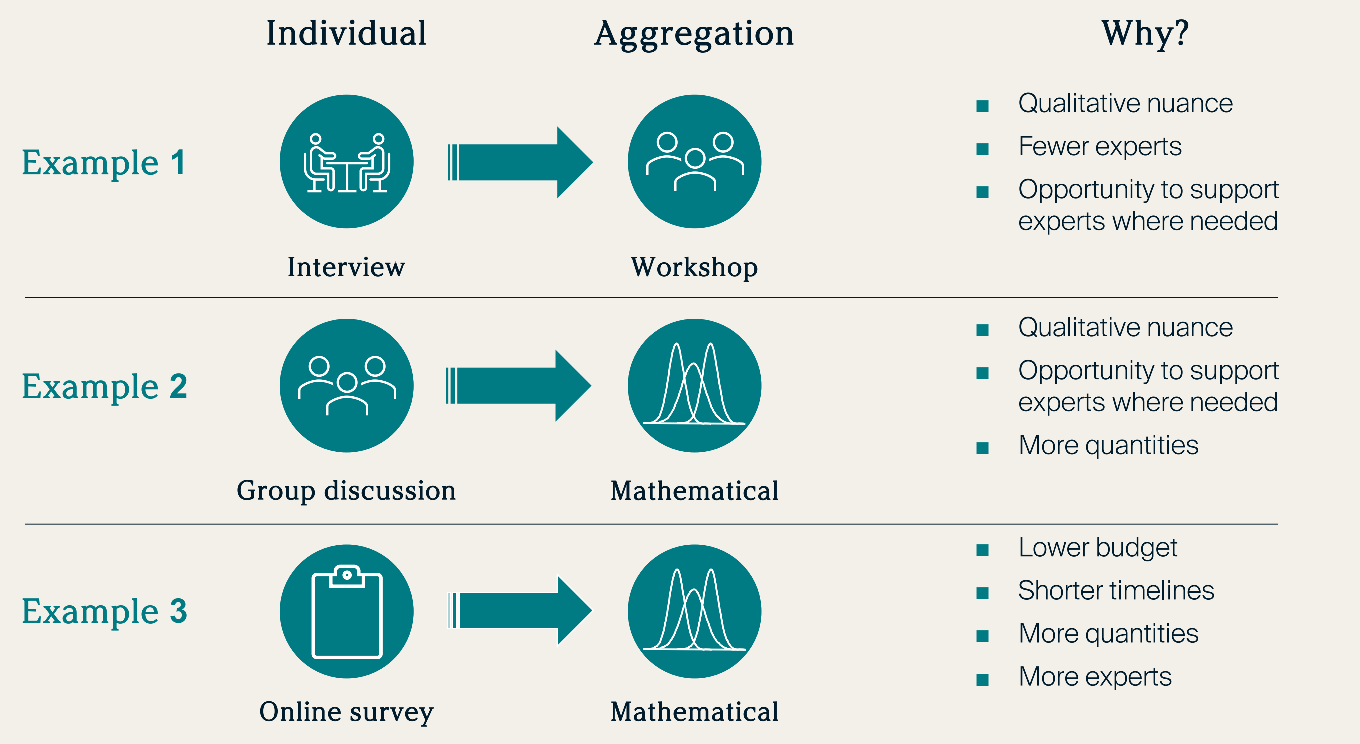
Approach to elicitation

For SEE that captures uncertainty, it is advised that judgements are first elicited individually, ideally through live facilitated elicitation (either in a group setting or via individual interviews). A remote survey approach may be preferred under certain circumstances, though this may lead to lower quality responses due to the lack of interaction with peers and facilitators. Aggregation using mathematical methods (such as linear opinion pooling) is preferred, but behavioural methods (such as a consensus workshop) may be applicable under certain circumstances.1 It is also possible to combine these two methods; this is particularly useful when a consensus cannot be reached in behavioural aggregation.

Guidelines offer conflicting recommendations on the approach and method to aggregate judgements. The MRC protocol (Sections 5.5.2 and 9.3.12) advises that mathematical and behavioural aggregation perform similarly in terms of 'accuracy' including representation of uncertainty on the basis of the evidence available.[1]

Taken together, there are several combinations of elicitation/aggregation to be considered, and a few examples are shown in Figure 1.

Figure 1
Example approaches to elicitation



The following should be considered when deciding on an approach:

1. **Budget and timelines**. Smaller budgets and shorter timelines favour use of fewer experts, or online surveys and mathematical aggregation, as this minimizes the number of touchpoints with the experts
2. **Number of experts.** Workshops are not suitable for large numbers of experts, so mathematical aggregation is advised if the expert pool is large
3. **Number of quantities of interest.** Interviews and workshops are not well-suited to large numbers of quantities of interest
4. **Value of qualitative nuance.** For quantities of interest for which the empirical evidence is severely lacking despite being critical for decision making, the qualitative discussion surrounding experts' judgements can be highly valuable as this can inform the narrative that supports the model and demonstrates the value of a product. Group interaction may also be valuable in bringing together the clinical community and provides an opportunity to discuss wider aspects of a HTA submission with clinicians. This benefit needs balancing with the need for an experienced facilitator in order to avoid the introduction of bias

Writing your protocol

It is essential to write a protocol before initiating an SEE study, such that methodological choices can be recorded and justified. We recommend using the reference protocol for HTA provided by Bojke et al.[1] (Table 1) as a basis for the choices you make, and justifying any deviations from this protocol that are required in the context of the specific study.

Table 1
A reference protocol for HTA

Element	Reference methods suggested
Experts	1. Recruitment will be driven by the context; however, the SEE should pursue diversity, representing the full range of valid experts beliefs. Experts should be willing to participate. 2. Focus on gathering substantive expertise or experience. Normative skills can be developed during the training session as part of the SEE. 3. Minimize and record conflicts of interest among the experts. Include experts external to the SEE task, i.e. not those involved in developing the task. 4. At least five experts should be included in the SEE.
Quantities elicited	1. Simple observable quantities should be elicited where possible; ratios or complex parameters such as regression coefficients should not be elicited directly. 2. Dependence between variables should be captured in SEE. Expressing dependent variables in terms of independent variables is preferable when experts do not have strong normative skills. 3. Wording should be clear, and quantities should be decomposed where this means a better fit with experts mental models.
Approach to elicitation	1. Beliefs should be elicited from experts individually, even if a group interaction follows. 2. Although interaction between experts can be structured through face-to-face sessions, constraints in HCDM, such as a lack of experienced facilitators, will usually mean that this will take place via a Delphi style remote process. 3. Between-expert variation should be explored explicitly.
Method	Both VIM or FIM work well; however, decision makers should aim for consistency across applications.
Aggregation	1. Statistical distributions should be fitted to experts' individually elicited judgements. 2. Following fitting, a summary of the individual distributions should be obtained using linear pooling with equal weighting of experts. 3. Any adjustments applied should be to improve coherence and consistency, not reduce variability. Internal and external review can be used to assess validity.
Delivery	1. Face-to-face where possible to allow a facilitator to deliver training to the expert. 2. Feedback to experts should be given during the SEE. Following feedback, experts should be given an opportunity to revise their distributions, either during or after a SEE session.
Training & piloting	1. Training is crucial and should focus on avoiding bias and expressing uncertainty. 2. Piloting should be undertaken.
Rationales & documentation	1. Rationales for how the experts made their judgements should be collected post SEE. 2. All methodological choices for the SEE must be documented and justified.

Key: FIM, fixed interval methods; HCDM, healthcare decision making; SEE, structured expert elicitation; VIM, variable interval methods.

Preparation

Quantities to elicit

In the context of healthcare decision making, the following parameter types commonly form the basis of quantities of interest:

- Simple probabilities (e.g. probability of infection)
- Conditional probabilities (e.g. probability of testing positive conditional on having the disease)
- Transition probabilities between health states (e.g. probability of disease progression)
- Time-to-event data (e.g. survival)
- Probability of repeated events (e.g. hypoglycaemic episodes in diabetes)
- Mean/median values (e.g. dosing, healthcare resource use)

Regardless of the parameter types, quantities of interest should be defined based on the following rules:

- Quantities should have a single, unknown value to ensure responses reflect subjective uncertainty surrounding that value, as opposed to variability within a dataset (e.g. 'mean daily dose of treatment X administered to population Y' has a single value, while 'daily dose of treatment X administered to population Y' contains variability as different patients may receive different doses)
- Quantities should be observable, where possible
- Quantities should be written in neutral, unambiguous, comprehensible language
- Where multiple quantities of interest are interdependent, this could be handled by expressing dependent variables in terms of independent variables, or by eliciting conditional probabilities. For more complex problems dependence elicitation techniques can be used.[2, 3]

Note that quantities of interest relating to clinical outcomes on a novel therapy may be challenging for experts to answer, unless they have experience with the treatment or analogous treatments beyond the trial (e.g. in an off-label or compassionate use setting).

Example wording for quantities of interest is shown in Figure 2.

Figure 2
Example quantities of interest

Example 1

Consider that all eligible patients in Population A receive Treatment X. What proportion of patients will be alive 5 years after treatment initiation?

Example 2

Consider that all eligible patients in Population A receive Treatment X. What is the mean dose (mg/kg/day) of Treatment X received by patients for the duration of treatment in clinical practice in Geography Z?

Example 3

Consider that all eligible patients in Population A receive Treatment X. What proportion of patients will experience Event Y in the first year of treatment?

Method for elicitation

Most elicitation approaches can be classified into the following two categories:

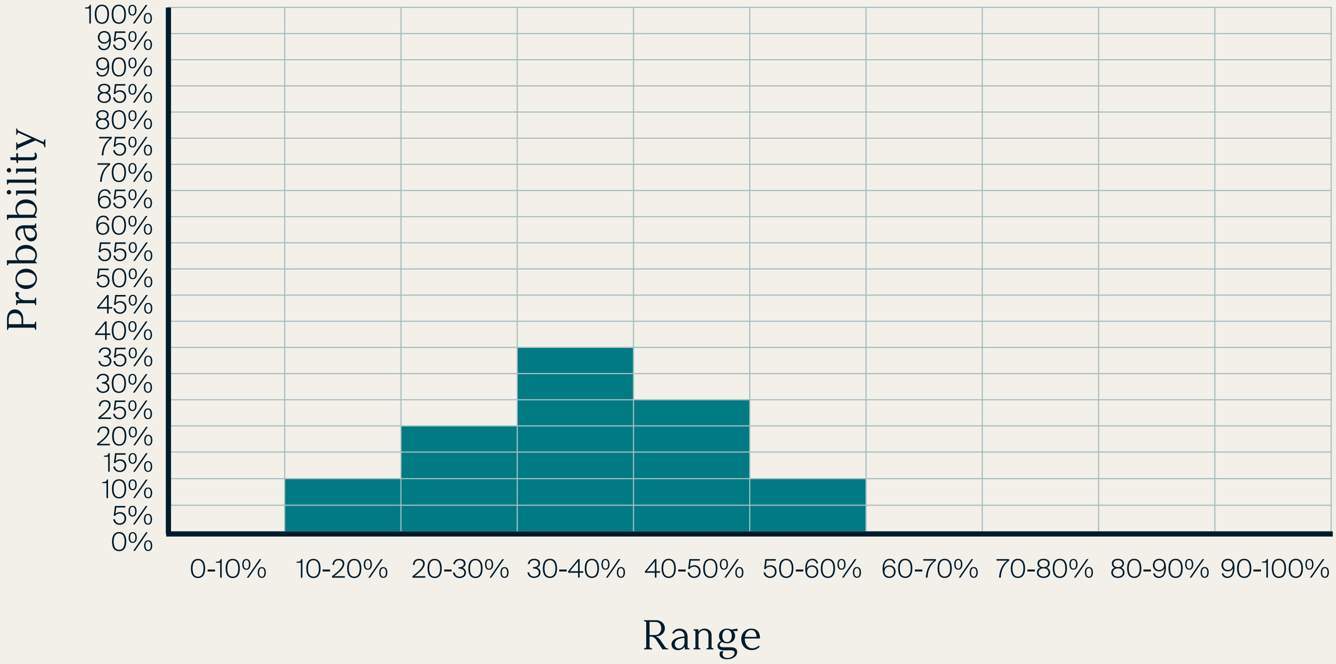
- **Fixed interval methods (FIM):** experts are provided with fixed ranges of possible values of the quantity of interest, and the experts provide the probability that the true value falls within each range (e.g. chips and bins)
- **Variable interval methods (VIM):** experts are provided with fixed percentiles, and the experts provide the ranges of possible values of the quantity of interest (e.g. quartiles)

There is a lack of empirical evidence on which method works better for healthcare decision making, and both methods have been used in this context.[1] FIM is generally preferred by experts and is more intuitive, but there may be a tendency for experts to focus on the shape of the histogram rather than the probabilities they are expressing. FIM is also more challenging in a group consensus context, given the large number of 'chips' (see Figure 3 below) to be placed per quantity of interest, each requiring discussion and agreement.

Any form of FIM or VIM can be selected, but consistency is recommended across all quantities of interest. Regardless of approach, the elicitation should begin by asking experts for their plausible limits (i.e. a range whereby it may be theoretically possible but is extremely unlikely for the true value to be below and above the lower and upper limits, respectively). This minimizes the risk of anchoring to a median estimate and thus overconfidence.

The FIM most used in SEE for healthcare decision making is the 'roulette' or 'chips and bins' method. In this method, experts are provided with a grid that divides the expert's plausible range into intervals and are asked to construct a histogram, with each 'chip' representing a unit of probability (e.g. 20 chips worth 5%) placed into one of the intervals or 'bins'. An example of a completed histogram is shown in Figure 3. Tools for conducting a chips and bins exercise (in both R and Excel) can be found here.

Figure 3
Example quantities of interest



A commonly used VIM method is the 'quartiles' method, in which experts are asked to provide values of the quantities of interest at the median (50th percentile) and the lower and upper quartiles (25th and 75th percentile). These values will be between the upper and lower limit, and their exact placement will be dependent on the amount of confidence the expert has in the true value being in each part of the plausible range.

A summary of commonly used elicitation methods is provided in Table 2.

Figure 8
'Quartiles' or 'bisection' method

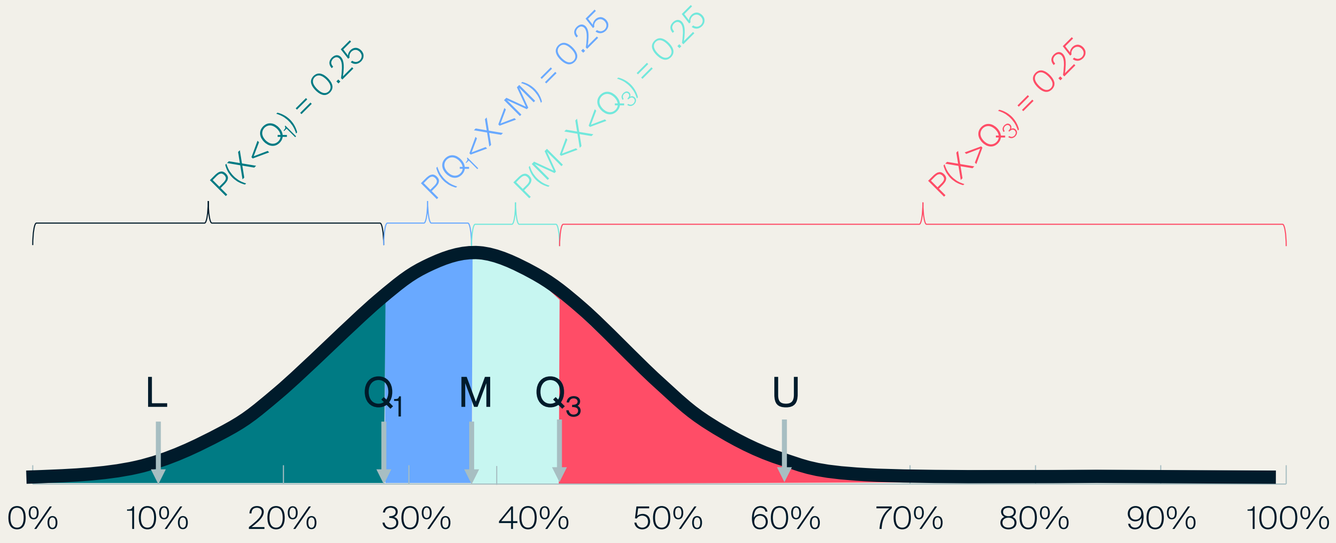


Table 2
Elicitation methods

Method	Type	Judgements elicited
Bisection	VIM	50th percentile
Quartiles	VIM	25th, 50th, and 75th percentiles
Tertiles	VIM	33rd, 50th, and 67th percentiles
Chips and bins	FIM	Probabilities associated with equal intervals of the quantity of interest (e.g. 10 intervals between lower and upper plausible limit)
Probability	FIM	Probabilities associated with three set intervals of the quantity of interest (e.g. X<X1; X>X2, X1<X<X2)

Key: FIM, fixed interval methods; VIM, variable interval methods.

Expert recruitment

How many experts?

The target sample size for an SEE exercise should be greater than five wherever possible and will depend on the method of aggregation and the scarcity of experts on the subject. Behavioural aggregation via a consensus workshop is not practical with a large number of experts, while there is no theoretical upper limit for mathematical aggregation. If a consensus workshop is desired, experience shows that this works best with five to eight experts.

Where there is a scarcity of experts (e.g. there may only be a handful of genuine experts on quantities of interest relating to ultra-rare diseases), a sample size of five may not be achievable (nor desirable). While efforts should be made to obtain a reasonable sample, the quantity of experts should not overly compromise the quality of their expertise.

Recruitment criteria

To minimize the risk of bias, inclusion and exclusion criteria should be developed as part of the protocol. These criteria should focus primarily on the level of substantive experience of the expert, and examples of how this can be measured include:

- Reputation in the field (e.g. number and quality of publications, referrals from other experts)
- Relevant experience (e.g. number of patients under their care, years of experience treating the patient group, involvement in key clinical trials)

Other aspects that recruitment criteria may specify include:

- Willingness to participate
- Availability to participate
- Absence of specific personal and financial conflicts of interest

A criterion related to normative expertise (the expert's ability to accurately assess and clearly communicate their beliefs in probabilistic form[1]) may also be included, though substantive experts often lack normative skills, hence the need for training. Furthermore, while adaptive skills (the expert's ability to adapt their knowledge to new situations for which they do not have prior experience[1]) may be desirable where outcomes relating to a new intervention are being explored, these are challenging to measure objectively.

Care should be taken when developing the criteria that diversity of opinion and range of expertise is not reduced unnecessarily. This will ensure that the result provides the best representation of the current state of knowledge, as variation between experts will be reflected as uncertainty in the aggregated distribution.

Identifying experts

The recruitment method should consider the relationship between the target sample size and the estimated total number of experts who meet the recruitment criteria, as well as pragmatic considerations such as timelines and budget. In some cases, it may be possible to contact the entire eligible expert pool (e.g. where patient care is organized in a small number of known specialist centres). Key Opinion Leader (KOL) mapping, a comprehensive process for identifying, ranking and profiling clinical experts, is a robust way of prioritizing experts for recruitment where this is not practical or feasible. In other cases, less robust recruitment strategies may be necessary, such as convenience sampling (recruiting experts known to the project team) and peer nomination (requesting recommendations from known experts), though these both carry the risk of homogeneity, and thus not capturing the full range of plausible opinions.

Expert preparation

Experts should be provided with training before the elicitation itself. This provides an opportunity to introduce the approach, as most experts are unlikely to have participated in such an exercise before, and the quantities of interest, to ensure the experts have the necessary context to provide informed judgements. In addition, although the evidence is limited, there are some suggestions from the literature that training can reduce biases such as anchoring, confirmation bias and overconfidence, and may help with expert motivation.

The content of the training will be driven by the specific requirements of the SEE exercise, but the following core elements should be included:

- Description of what is required from experts
- Outline of process
- Description of the quantities of interest and discussion/resolution of any ambiguity
- Description of how dependence between quantities of interest will be handled
- Example and practice questions
- Common biases and heuristics
- Assumptions and definitions used

Additional content should be considered based on context, and may include, for example, a description of performance measurement (e.g. if expert judgements will be weighted)

In addition to training, it is good practice to provide an evidence dossier to the experts before the elicitation, with the goal of contextualizing the questions being asked and comprehensively summarizing the evidence relating to the quantities of interest. This minimizes the risk of availability bias, where people make judgements based on the evidence that they can quickly bring to mind, ignoring less memorable evidence.[4] Creating a single image that summarizes all relevant data in a visual way can further mitigate availability bias.

Execution of elicitation

Terminology

Regardless of whether the elicitation is being conducted in writing (e.g. a survey) or in spoken word (e.g. an interview), care should be taken around terminology. Questions should be worded such that it is clear to the expert that their personal, subjective probability is being sought. For example, SHELF guidance recommends using terms such as 'your upper plausible limit' and 'your median'.[4] Where the quantity of interest is a probability of an event occurring (e.g. 1-year survival), it may be helpful to refer to this as a proportion (e.g. proportion of patients alive after 1 year), rather than a probability (e.g. probability of survival in Year 1), to avoid confusion with the experts' subjective probability that a certain value is true.

Rationales

For each judgement an expert provides, it is essential to also seek their rationale for those judgements. This can be used to (i) assess the validity of the elicited beliefs, and (ii) understand between-expert variation. Where a group-level elicitation will follow, it may be useful to explicitly explore the rationales provided in the individual-level elicitation to help experts move towards a consensus distribution.

Fitting distributions

Once experts have provided their judgements via either a FIM such as chips and bins, or a VIM such as the quartiles method, a probability distribution should be fitted to the elicited data. There is little evidence to indicate the best way of doing this, but options include the following:

1. A pre-selected statistical function (e.g. the beta distribution) is fitted to the expert's judgements
2. The best-fitting distribution from a pre-selected list of statistical functions is fitted to the expert's judgements

Feedback and refinement

Feeding back results to experts and allowing them to refine their judgements is a useful strategy for avoiding misunderstanding, increasing precision and minimizing bias. There are two key stages in the elicitation process where this can be done:

1. Following elicitation of initial judgements
2. Following fitting of a probability distribution

Following elicitation of initial judgements, the experts' judgements can be put back to them in a different way to test the judgement. For example, if the chips and bins method is used, the implied median can be estimated and fed back to the experts to check if this is reflective of their beliefs. Similarly, if the quartiles method is used, the experts could be asked if it is reasonable to believe that it is equally likely that the true value falls within the interquartile range as outside of it.

The fitted probability distribution(s) may also be fed back to the experts for validation and/or selection. Where experts do not feel that the fitted probability distribution is reflective of their beliefs, they may choose to adjust their initial judgements or select a different probability distribution. However, this is challenging for experts with limited normative skills, and works best in an interview or group workshop setting.

We recommend always including some form of feedback and refinement of initial judgements and applying discretion as to whether probability distributions are fed back to the experts. Where possible, experts should be shown the outcome of their answers visually; both on the quantity of interest and any end outcomes (for example if survival at a certain timepoint is being elicited and used to guide lifetime survival projections the impact on the final projection should be shown). Regardless of approach, it is important to make this explicit to the experts and document the process accurately.

Aggregation

There are two methods available to produce a pooled distribution reflecting all of the experts' views:

- Behavioural: grouping together individual experts to generate a consensus or independent rational observer's summary
- Mathematical: combining the beliefs of individual experts using a mathematical rule, such as linear opinion pooling

Guidance on selecting the most appropriate methodology for aggregation is provided on Page 3.

Mathematical aggregation

Mathematical aggregation methods fall into two general approaches: Bayesian combination methods, and opinion pooling methods. Opinion pooling is the method most commonly used and takes an average of the distributions from each expert. The most common approach is linear pooling assigning each expert equal weighting.

The MRC protocol advises the use of linear pooling with equal weights for mathematical aggregation for simplicity and due to a lack of research on how to generate appropriate weights. Further research on more complex methods is required to determine where/if these may be more appropriate. Consideration of more complex weighting methods is, however, advised where experts represent different disciplines, contribute different perspectives on the elicited quantities or are considered likely to have major differences in accuracy and therefore considerable heterogeneity is anticipated.

Behavioural aggregation

Behavioural aggregation methods aim to allow a group of experts to form an aggregated distribution through agreement using a structured communication or discussion framework, rather than a mathematical formula.

The benefits of this approach are that:

- The resulting distribution has a sensible practical interpretation, as the experts have agreed on it
- A consensus distribution avoids cases where the experts have major disagreements with the aggregated distributions, as the final distribution would be agreed upon by the group

The main drawback is that an experienced facilitator is required to minimize biases associated with group interaction, and even then these biases cannot be completely avoided. In addition, experts may not always be able to form a consensus as a group (in which case the facilitator should look to understand the number of different prevalent hypotheses and create separate aggregations per hypothesis for use either separately or to be aggregated mathematically).

There is not enough evidence to support any particular method of behavioural aggregation. The SHELF resources provide guidance on production of a single distribution that represents how a 'rational independent observer' would summarize the range of expert opinions.[4] SHELF recommends experts are instead asked to provide a prior for a rational impartial observer, who has observed their discussion and all of their evidence as a neutral viewpoint which aims to help experts to avoid bias, personal investment or interpersonal difficulties.

Assessing the expected accuracy of expert judgements

Guidelines differ in their definitions of validity and discussion of how the concept can be operationalized in an elicitation. There are a few different methods that can be used:

- Qualitative feedback and coherence testing
- Use of seed questions to assess the accuracy of judgements
- Internal and external peer review

These are discussed in more detail in the MRC protocol (Section 5.6 and 9.3.17).[1]

Qualitative feedback and coherence testing

This is the simplest method of validation and is recommended for all SEE. Experts should be asked for the rationale for the answers given, any difficulties completing the exercise and how easy to understand the wording of the questions is. The answers should be used to confirm that the questions have been understood.

Additional methods to test for coherence include:

- Comparison across elicited probabilities for coherence (e.g. checking 3-year vs 5-year survival)
- Overfitting (asking for one more summary than is needed)
- Use of different elicitation methods and comparison of results

Use of seed questions and calibration

'Calibration' or 'seed' questions (questions where the answer is known) can be used to assess the accuracy of elicited judgements using scoring rules that compare the elicited assessments with known answers, though there is limited guidance on developing meaningful seed questions for healthcare decision making.[1] The most common strictly proper scoring rule used in SEE is that of Cooke's Classical Model, which has been used to elicit judgements and measure calibration in over 100 expert panels.[5]

- Multiple seed questions are needed to assess the accuracy of elicited probability distributions
- Seed questions must be closely related to the target questions but unknown to the experts participating in the elicitation
- Seed questions are used to assess an expert's skill in quantifying uncertainty, so they should not just be a test of the expert's ability to recall established facts or familiar quantities
- Seed questions commonly come from four sources: future measurements, unpublished measurements, unfamiliar information from standard datasets, or combining or comparing different datasets

The answers to these questions can be used within the aggregation step to weight expert responses according to performance, and also provide an opportunity to check experts' understanding of the task. This is not, however, currently recommended as more research is needed on the usefulness and methodology for this in a healthcare setting.

Use within economic analyses

The output from the elicitation exercise will come in one of two formats:

- From behavioural aggregation or an individual expert: a fitted distribution type with parameters, e.g. a beta distribution with parameters alpha and beta
- From mathematical aggregation: a set of samples from the pooled cumulative distribution function calculated by weighting (usually linearly) the individual fitted distributions[a]

Discrete and continuous variables

In your economic model you can implement the output from discrete and continuous variables deterministically simply by calculating the mean of your fitted distribution from behavioural aggregation or pooled cumulative distribution function from mathematical aggregation.

In order to implement probabilistically you simply take a sufficiently large number of random samples from either the fitted distribution from behavioural aggregation or from the pooled cumulative distribution function from mathematical aggregation.

Time-to-event data

Implementation is less straightforward for parameters which are to be used to inform time-to-event analysis (for example datapoints elicited for overall survival). The literature is not conclusive on how to integrate inputs elicited from inputs with empirical evidence for use within economic analysis. Potential methods with a few example case study references are provided below and include:

1. Use of expert-elicited values as priors within a Bayesian analysis. For example:

   a. Use of elicited survival proportions to define modelled survival parameters (and the correlation between these)[6, 7]

   b. Introduction of expert opinions on the form of the treatment effect (e.g. trend in the hazard ratio).[8] Example:

$$HR(\mathrm{t})\text{~}N\left(\frac{h_{1,\mathrm{RCT}}(t)}{h_{0,\mathrm{RCT}}(t)}, 0.1^2\right),\ t = 6 \ldots 35\ \text{years}$$

   c. Combination with other types of external evidence, including basic techniques such as omission of implausible parameters sets, ensuring that prior survival functions are monotonically decreasing and that prior estimates of the population mean are bounded and/or external datasets such as general population lifetables (as an upper bound) and observational data sources

2. Conversion of the elicited survival to discrete interval hazards for each timepoint following trial end which can then be pooled with the original data set from the trial either prior to or post survival model fitting using a variety of methods including weighted model averaging[9]
3. One package under development which allows the incorporation of outputs from expert elicitation into survival analysis as well as other external data sources is survextrap in R: http://chjackson.github.io/survextrap

Whilst expert elicitation exercises can be used as a form of validation and to guide model selection based on trial data alone and has been used for this purpose in many prior submissions, this type of retrospective use may not achieve its intended objective and is an inefficient use of information.[7]

Further research is still required to develop practical examples of the use of experts' beliefs in survival models and how these should be used either within a Bayesian framework to formulate prior model probabilities or within a classical framework to produce a preferred combined estimate for survival.

In the interim, we would recommend that the methodologies provided in the three reference papers cited above provide good examples of how such data can be used.

[a] You can use a package such as the SHELF package in R to fit distributions and to perform linear aggregation using commands:

- fitdist: fit distributions to the elicited values per expert or from behavioural aggregation
- Plinearpool, qlinearpool and rlinearpool – provides probabilities, quantiles and samples from a (weighted) linear pool

Reporting

Clear reporting of the methods used for expert elicitation or expert opinion (quantitative) is needed from study planning to conduct. All methodological choices for the SEE must be documented and justified. You are likely to need to use the report as a supporting reference or appendix for future HTA submissions. A detailed list of items that should be reported is provided in Table 3. Additionally, you may have other elements critical to your study that need reporting.

Table 3
Elicitation methods

Criterion	Description	Note
Research rationale	The need for using an expert elicitation exercise should be described	This should ideally include some reference to the design and conduct of systematic reviews to identify key input parameters for the decision–analytic model and a statement confirming that these reviews did not identify data relevant for the model-based economic analysis as specified
Research problem	All uncertain quantities (model input parameters) that will be elicited should be described	In some instances, there may be a substantial number of uncertain quantities required, and a degree of 'preselection' will have occurred to identify a relevant subset. Clear justification for model parameters identified as key for the decision problem needs to be provided
Measurement of uncertain quantities	The rationale for the measure type and method of encoding of each uncertain quantity elicited should be described	The measurement type of uncertain quantities can be (but not limited to): scalar quantities (i.e. numbers); proportions (e.g. probabilities); ratios (e.g. odds, hazard); risk (e.g. relative); rate (e.g. mortality), etc. Some measures are easier to understand and elicit than others; thus, it is important to fully justify the selection of any measurement type The approach used for measurement may include either FIM (e.g. Roulette or chips and bins method), VIM (e.g. Bisection, Quantiles, plausible probabilities etc.), hybrid fixed-variable interval methods, summary statistics or other approaches to encode expert judgements
Definition of an expert	The nature of the expert population should be described to clearly state what topic of expertise they represent and why	It is unlikely that a single expert will be sufficient, and it is generally necessary to elicit judgement from a group of experts that were selected to represent the views of a larger population
Number of experts	The selection criteria, the number of experts approached and the final number of experts who provide expert judgement should be reported	Selection criteria need to be described in detail. There should be clear and specific pre-defined criteria used to identify how experts were selected and if/how their elicited quantities were used
Conflicts of Interest	The declaration of potential conflict(s) of interest from each expert whose opinion was sought	Minimize and record conflicts of interest among the experts. Include experts external to the structured expert elicitation task (i.e. not those involved in developing the task)
Preparation	There should be clear reference made to a protocol that describes the design and conduct of the elicitation exercise	None
Piloting	It should be clearly reported if the elicitation exercise process was piloted, and a summary of any modifications should be made	The selection and number of experts used in the piloting process should be reported. Key aspects that may have required modification include: selection of experts; measure type and number of uncertain quantities to be elicited; training exercise; framing of the elicitation question; method of aggregation
Data collection	The approach to collect the data should be reported	Data can be collected from individual experts or a group/s of experts. Collecting data from individual experts means that a mathematical aggregation process may need to be used. Collecting data from group(s) of experts means that behavioural aggregation methods may be used
Administration	The mode of administering the elicitation exercise should be reported	Elicitation exercises can be conducted face-to-face or via the telephone and/or computer. In some situations, it may be feasible to collect the data using a self-administered online or postal survey. Interactions or discussion between experts may be permitted and occur prior, during or after group elicitations (given opportunity for revision)
Revisions and further rational	The use of estimate revisions, interaction and discussions between experts and providing further rational and justifications	This should detail whether elicitations were one-shot or if iterative adjustments were allowed (i.e. a circling draft elicitation report was used within a group), and whether the elicitation was confidential to the researcher or derived within a group. The additional feedback may include written or oral descriptions of the experts rational for their judgements, any discussion on the possibility of further research and any further thoughts on the process of elicitation

Criterion	Description	Note
Training	The use of training materials should be reported and made available	This may include training materials sent to the experts and/or training in the use of probabilities and nature of distributions. A description of what is required from experts, along with how results will be used, example and practise questions and a review of assumptions made. This document may need to provide an explanation of efforts made to prevent influencing experts' knowledge and judgement, including the effect of heuristics and bias. In practice, this recommendation will require a copy of the elicitation exercise to be included, which is likely to be presented as electronic supplementary material
The exercise	The number and framing of questions used in the exercise should be reported and made available	This will require a copy of the elicitation exercise to be included, which is likely to be presented as electronic supplementary material
Data aggregation	The type of aggregation method (mathematical or behavioural) should be reported together with a description of the method or process used to aggregate the data	Mathematical aggregation (relevant when data were collected from multiple individual experts) can be conducted using a range of methods, for example: Bayesian methods; opinion pooling; Cooke's method. Behavioural aggregation (relevant when data were collected from group[s] of experts) can be conducted using processes such as, for example: Delphi or Nominal Group technique
Measures of performance for data aggregation	The processes followed to estimate measures of performance (calibration/ information) for data aggregation need to be fully described	Calibration is the process of measuring the performance of experts by comparing their judgement with a 'seed parameter' (parameter whose true values are known or can be found within the duration of a study). Calibration scores represent the probability that any differences between expert's probabilities and observed values of 'seed parameters' might have arisen by chance. Information represents the degree to which an expert's distribution is concentrated, relative to some user-selected background measure
Ethical issues	The ethical issues for the expert sample and research community should be described	The use of expert elicitation should acknowledge the issues of ethical responsibility, anonymity, reliability and validity in an ongoing manner throughout the data collection and aggregation process
Presentation of results	The individual, and aggregated, point estimate(s) and distribution for each uncertain quantity (quantities) should be presented	The units of measurement should be clear, and attention should be paid to the style of presentation that may benefit from the use of figures rather than relying on a tabular format
Interpretation of results	The interpretation of uncertain quantities elicited should be presented together with a description of how the results will be used in the model-based economic analysis	This should include an explanation of how the reader should interpret the results. It should be recognized that the number and type of experts used will affect the results obtained. The interpretation of results should comment on the degree of uncertainty observed

Key: FIM, fixed interval methods; VIM, variable interval methods.

References: Iglesias et al.[10]; MRC protocol[1]; NICE health technology evaluations manual[11]

Practical considerations

Project team and roles

Conducting an SEE exercise and incorporating the results into an economic analysis requires a team with a broad skillset. Key technical skills are summarized in Table 4.

Table 4
Technical skills needed for SEE

Technical skill	Experience level required	Scenario
Selecting optimal approach and defining the protocol for elicitation	High	Always needed
Summarizing the evidence base and report writing	Mid	Always needed
Coordinating experts and scheduling meetings	Low	Always needed
Leading interviews	Mid	Only needed for studies involving interviews
Facilitation of training and group workshops	High	Usually needed
Generating probability distributions using available tools	Low	Always needed
Mathematical aggregation using available tools	Low	Only needed for mathematical aggregation studies
Qualitative analysis	Mid	Usually needed
Incorporation of simple inputs into an economic model	Mid	Usually needed
Incorporation of complex inputs into an economic model (e.g. time-to-event data)	High	Not always needed

In our experience, these responsibilities can be shared among a project team of three to six people. Example roles and responsibilities are shown below in Table 5.

Table 5
Roles and responsibilities

Role	Experience level required	Responsibilities
Project Manager	Low	Coordinating experts and scheduling meetings
Health Economist and/or Statistician	Mid–High	Generating probability distributions using R, mathematical aggregation using R, incorporation of inputs into statistical or economic model
Research Associate	Low	Summarizing the evidence base and report writing, qualitative analysis
Lead/Facilitator/Advisor	High	Selecting optimal approach to elicitation, Leading interviews, facilitation of training and group workshops

Ethics and compliance requirements

It is essential to ensure that research is conducted in an ethical manner that is compliant with industry regulations. Most pharmaceutical and biotechnology companies have compliance approval processes for this purpose. The following guidance relevant to the design of SEE exercises from the Association of the British Pharmaceutical Industry's Code of Practice 2021[12] is provided as an example, but it is important to consider the relevant local guidance:

1. Where health professionals are used as consultants and advisors:

   a. A written contract should be agreed in advance

   b. A legitimate need for the serve must be clearly identified

c. The criteria for selection must be directly related to the need

d. The number of contracted individuals and the extent of the service must not be greater than reasonably necessary

e. The contracting company must maintain records and make appropriate use of the service provided

f. The hiring of the contracted party to provide the relevant service must not be an inducement to prescribe, supply, administer, recommend, buy or sell any medicine

g. The remuneration for the services must be reasonable and reflect the fair market value of the services provided

2. Materials and activities must not be disguised promotion
3. Companies must document and publicly disclose certain transfers of value made directly or indirectly to health professionals, other relevant decision makers and healthcare organizations
4. Material should only be provided or made available to those groups of people whose need for or interest in it can reasonably be assumed. Material should be tailored to the audience to whom it is directed
5. Information, claims and comparisons must be accurate, balanced, fair, objective and unambiguous and must be based on an up-to-date evaluation of all the evidence and reflect that evidence clearly. They must not mislead either directly or by implication, by distortion, exaggeration or undue emphasis. Material must be sufficiently complete to enable recipients to form their own opinion of the therapeutic value of the medicine

Data storage

Personal identifiable information should be processed in-line with local regulations. In the EU and UK, the General Data Protection Regulation[13] states that personal data should be:

1. Processed lawfully, fairly and in a transparent manner in relation to individuals
2. Collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes
3. Adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed
4. Accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay
5. Kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organizational measures required by the GDPR in order to safeguard the rights and freedoms of individuals
6. Processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures

References

1. Bojke L, Soares M, Claxton K, et al. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technology Assessment* (Winchester, England). 2021; 25(37):1.

2. Werner C, Bedford T and Quigley J. Sequential refined partitioning for probabilistic dependence assessment. *Risk Analysis*. 2018; 38(12):2683-702.

3. Werner C, Colson A, Morton A and Bedford T. Risk assessment of future antibiotic resistance-eliciting and modelling probabilistic dependencies between multivariate uncertainties of bug-drug combinations. *Frontiers in Applied Mathematics and Statistics: Mathematics of Computation and Data Science*. 2021; 7.

4. Oakley JOHA. The Sheffield Elicitation Framework (SHELF). 2022. Available at: http://www.jeremy-oakley.staff.shef.ac.uk/shelf/. Accessed: 2 June 2022.

5. Colson AR and Cooke RM. Expert elicitation: using the classical model to validate experts' judgments. R*eview of Environmental Economics and Policy*. 2020.

6. National Institute for Health and Care Excellence (NICE). TA592: Cemiplimab for treating metastatic or locally advanced cutaneous squamous cell carcinoma. 2019. Available at: https://www.nice.org.uk/guidance/ta592/history. Accessed: 2 June 2022.

7. Stevens J and Orr M. Using Clinical Experts Beliefs to Compare Survival Models in Health Technology Assessment. *arXiv preprint arXiv*:210906760. 2021.

8. Guyot P, Ades AE, Beasley M, et al. Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making*. 2017; 37(4):353-66.

9. Cope S, Ayers D, Zhang J, et al. Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. *BMC medical research methodology*. 2019; 19(1):1-11.

10. Iglesias CP, Thompson A, Rogowski WH and Payne K. Reporting guidelines for the use of expert judgement in model-based economic evaluations. *Pharmacoeconomics*. 2016; 34(11):1161-72.

11. National Institute for Health and Care Excellence (NICE). NICE health technology evaluations: the manual. 2022. Available at: https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation. Accessed: 2 June 2022.

12. The Association of the British Pharmaceutical Industry (ABPI). ABPI Code of Practice for the Pharmaceutical Industry. 2021. Available at: https://www.abpi.org.uk/. Accessed: 27 May 2022.

13. Intersoft Consulting. General Data Protection Regulation (GDPR),. Available at: https://gdpr-info.eu/. Accessed: 7 June 2022

Lumanity applies incisive thinking and decisive action to cut through complex situations and deliver transformative outcomes to accelerate and optimize access to medical advances. With deep experience in medical, commercial, and regulatory affairs, Lumanity transforms data and information into real-world insights and evidence that powers successful commercialization and empowers patients, providers, payers, and regulators to take timely and decisive action.

Contact us to learn more about how Lumanity can support your unique challenge.

lumanity.com

The Centre for Health Economics (CHE) at the University of York is a world-renowned institute that produces policy relevant research and innovative methods that advance the use of health economics to improve population health.

york.ac.uk/che