Eliciting expert judgements in HTA

1. Introduction

Decisions in Health Technology Assessment (HTA) require judgements about interventions and treatments, particularly as new technologies are assessed progressively closer to their launch on the market. Evidence used to establish cost-effectiveness is typically uncertain, for example the evidence may not be on 'final' outcomes (e.g. cancer products licensed on evidence of progression-free survival), or the evidence base may not be well developed (e.g. in diagnostics, medical devices, early access to medicines scheme).

The judgements required for decision making can be requested of experts, which are typically, but not exclusively, clinicians [1]. Iglesias, et al [2] distinguishes processes intended to obtain qualitative judgements that resolve features of existing evidence (expert opinion) from processes that seek to generate quantitative estimates from experts where evidence does not exist or is not appropriate (expert elicitation). Both expert opinion and expert elicitation can be conducted using a structured process, where a pre-specified protocol or guide is used, or an unstructured process, where methods used are tailored to the specific circumstances and are not necessarily documented in a consistent way.

Expert elicitation and expert opinion (collectively referred to as expert judgement) can be used to describe uncertainties associated with the cost-effectiveness of competing interventions and used to assess the value of further evidence generation. If conducted in an appropriate manner, biases can be minimised. This may requite the use of explicit processes that can be adopted to suit transparent and accountable decision-making processes.

There are a number of generic and topic specific guidelines for generating expert judgement in the literature [3]. Despite expert judgement already being used in HTA, a review of applied studies found significant heterogeneity in the methodology used and a lack of consideration for existing guidance on the topic [1]. It is thought that the observed heterogeneity in methods is due to the lack of guidelines specific for the purposes of HTA. A recent body of work aimed to establish a protocol for structured expert elicitation of quantitative information to inform health-care decision-making, including HTA [3].

This report aims to inform the NICE methods update, and considers the methods available for expert judgement, how expert judgement has been used in health technology assessment to date and how it may be used in particular contexts or areas. Guidance from international organisations on the topic and reporting guidance is also reviewed. The objective of this is to make recommendations on how methods and features of expert judgement can be integrated into the methods update. This work is distinct from the topic of preference elicitation used to inform Multi Criteria Decision Analysis.

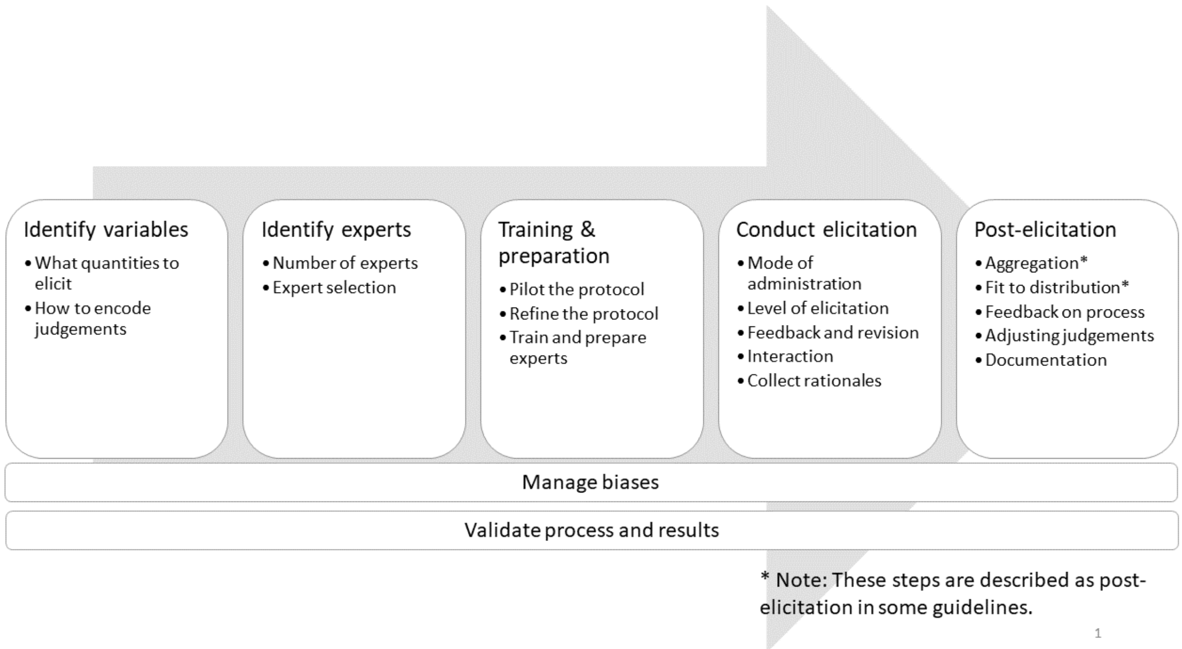
2. Overview of expert elicitation

Expert judgement is used in a number of disciplines, including natural hazards, environmental management, food safety, healthcare, security and counterterrorism, economic and geopolitical forecasting, and risk and reliability analysis [4-8]. How judgements are generated is critical to the quality of the results, and hence ultimate decisions and policies that use these judgements. Whilst some methods research on elicitation should be of value across disciplines, an assessment of the suitability of methods to specific decision making contexts, and to the relevant content experts, is also required.

In HTA, expert opinion has been used at several stages of assessing technologies, for example, in topic generation, in the scoping phase, in generating evidence on cost-effectiveness and as part of decision-making processes themselves. Historically, methods used to seek expert judgement have been ad hoc [9], for example consulting single experts for 'best guesses' of parameters of a decision model for which data were missing or inappropriate. This suggests that these have been unstructured processes, that request a measure of central tendency and no indication of uncertainty. However, there are a few examples of the use of more formal and structured processes in HTA [1].

An overview of the underlying process for elicitation, summarised from a review of guidelines [3], is depicted in **Error! Not a valid bookmark self-reference.**. All the 16 guidelines included in this review recommend documenting the elicitation process and its results. The guidelines typically recommend documentation to be presented, that includes the elicitation questions, experts' responses (both individual , if elicited, and aggregated), experts rationales, and a detailed description of the procedures and design of the elicitation, including the reasoning behind any methodological decision.

Figure 1: The elicitation process



At each step of the elicitation process, analysts are faced with a variety of methodological choices. For example, what quantities to elicit, which experts to include, how the elicitation will be conducted and how results will be analysed.

3. Roles SEE could play in decision making

For this work, a survey was developed to identify key areas and aspects of HTA where expert elicitation has a special role in informing decision making.

3.1. Survey design

The survey was targeted to deision makers and was administered online using Google Forms. Out of the 41 individuals that were invited, 11 completed the survey (i.e. response rate = 26.8%). Invitees included chairs and members of NICE appraisal committees, Public Health (PH) Advisory committees, and decision makers from other jurisdictions.

The survey comprised 12 questions (see Appendix D: for survey questions), and asked participants about their experiences of expert judgement, including any examples of more structured, protocol driven processes they had encountered, as well as its potential future role. Respondents were asked about the ease of incorporating each form of expert opinion in decision-making and how specific factors such as any conflicts of interest or poor reporting of methods could lead to down-weighting of the evidence provided by experts. Finally, respondents were asked how the use of expert elicitation could be improved, for example through better reporting. Open-ended questions were included allowing participants to add further relevant information/areas/considerations that were not covered by the survey questions.

3.2. Summary of survey results

The results of the survey are presented in the supplementary material; here we summarise its results. All participants were members and/or chairs of decision-making committees; the majority were involved in NICE Technology Appraisal committee (n=7/11) and there was one affiliate for each of the following: NICE Diagnostics Appraisal Committee, NICE HST, NOMA, or PH (the latter without HTA involvement). Some participants also had experience in critiquing evidence submitted by companies or as expert advisors.

All participants reported that expert opinion is used in both committee meetings and as part of the evidence generation primarily in the form of quantitative opinions. When quantitative estimates are used in practice, these most often comprise both a central estimate and a measure of uncertainty. Additionally to elicited evidence being used at committee meetings and within the evidence generation process, participants suggested that expert opinion is also being used in price negotiations, for example in patient access schemes, in decision scoping and the `technical engagement' processes.

In terms of different forms of expert opinion, on the whole, participants suggested that qualitative opinions are the hardest to incorporate into decision-making, whilst a quantitative measure of the central estimate and uncertainty is the most straightforward. The vast majority said that they would value expert opinion more if it is elicited quantitatively and accompanied by a measure of uncertainty.

In what concerns perceived limitations of expert opinion, participants indicated that expert opinion is most strongly down-weighted when experts are conflicted, followed by when the reporting of elicitation methods is poor. A degree of down weighting (moderate) is also apparent due to the fact that evidence elicited from experts is generally considered of low quality–because it is not empirical or experimental evidence. Participants added that experts' opinions may also be discounted when there is lack of consensus amongst them.

Across all decision models inputs (e.g. treatment effects, natural history etc), most often, expert opinion was used in a qualitative form, whilst quantitative measures (including both central estimates and uncertainty) were least frequently used. Participants indicated that expert opinion is also used to inform adverse events, establish how surrogate outcomes link to clinical outcomes, identify details of care pathways, among others.

Regarding the various features of decision problems, participants identified that expert opinion is currently being used appropriatly; however, some of the open field responses suggested that decision making processes have tended to select conflicted experts without attempting to minimise biases.

In what concerns the description of the methods involved in the conduct of the elicitation exercise and the reporting of results, participants identified that minimal or no detail is generally reported. Most participants have not encountered examples where a more structured process of eliciting quantitative parameters was used or were unable to determine whether a structured process was utilised given the lack of presented detail.

There is agreement across participants that expert opinion (qualitative and quantitative) could become more useful if it is conducted on key areas of uncertainty, if it ensures that experts are appropriately identified, elicitation methods are described in detail and a structured process is used, and, finally, if expert elicitation is conducted prior to the committee meeting. However, one participant expressed that structuring expert's input any further than currently done could reduce the opportunity to explore and challenge their opinions. Some participants suggested that having a template/protocol for conducting and reporting expert elicitation exercises would be extremely useful.

4. Guidance from HTA organisations

Guidance available from HTA organisations was reviewed to determine recommendations made regarding the use of expert opinion. Fourteen, country-specific, HTA methods guides were included. These were searched using the keywords that were used in the review of applied papers (see Appendix 1:a)i)(1)(a)Appendix B: Appendix 1:a)i)(1)(a)Appendix B:). Table 1 provides a brief summary of the main characteristics of the guidance documents reviewed.

4.1. Summary of findings

Except for a few [10-13], most HTA organisations mention that expert opinion may be used to inform multiple types of parameters, such as efficacy, HRQoL, costs and resource use and adverse events. These guidance consistently suggest that evidence elicited from experts is of low quality and recommend its use only in the absence of other (empirical) evidence [14-21]. Most HTA organisations require that when evidence derived from experts is used, the process of experts' selection and methods for elicitation is clearly described to ensure transparency [14-21] and that sensitivity analyses using other sources of data should be conducted where such data is available [14, 22]; however, only two provide additional details on which processes are considered appropriate and how experts should be selected or experts' beliefs elicited and combined [14, 15]. These will be described in more detail next.

Table 1: Summary of guidance on expert elicitation methods across HTA organisations

Jurisdiction	HTA Agency	Last updated	Elicitation mentioned as a source of evidence?	Level of detail on methods for expert elicitation	Link
England & Wales	NICE	2013	Yes	Minimal detail on methods that should be used and sensitivity analyses that should be conducted	link
Australia	PBAC	2016	Yes	Some detail on the information that need to be provided when elicitation is conducted, the design and conduct of the exercise, the process of expert selection and aggregation of expert opinions and on additional sensitivity analyses	link (See Appendix 1)
Canada	CADTH	2007	Yes	Minimal detail on design and conduct of the elicitation exercise as well as on the process of experts' selection. They refer to[23] and [24] for further information	link
USA	ICER	2018	No. (only for provider mark-up)	No detail	link
Ireland	HIQA	2019	Yes	Minimal detail on the design and conduct of the elicitation exercise as well as on additional sensitivity analyses	link
France	HAS	2012	Yes	Minimal detail on the design and conduct of the elicitation exercise as well as on the process of expert selection	link
Germany	IQWIG	2016	No	No detail	link
Sweden	SBU	2018	Yes. (Only to establish diagnostic tests reference standard)	No detail	link
Netherlands	ZIN	2016	Yes	Some detail on the design and conduct of the elicitation exercise	link (See section 3.4)
Portugal	INFARMED	2019	Yes	Considerable detail on the information that need to be provided when elicitation is conducted, the design of the elicitation exercise, the methods of elicitation, the process of experts' selection, and additional sensitivity analyses	link See section 12
Norway	NOKC	2014	No	No detail	link
Scotland	SMC	2019	No	No detail	link
Poland	AOTMiT	2016	Yes	Minimal detail on the information that need to be provided when expert opinion is used	link
Thailand	HITAP	2014	Yes	Minimal detail on the information that need to be provided when expert opinion is used	link

4.2. Detailed guidance provided by other HTA organisations

Current NICE methods guidance [22] suggests that information may be elicited from experts if there is a lack of data to inform the mean value and associated distribution of a parameter used in cost-effectiveness analysis, and that formal elicitation methods should be used. However, the methods guide does not provide any details regarding the methods that are considered appropriate. Two national HTA bodies that have updated their methods guides more recently (the Australian and the Portuguese) have included more detailed methodological recommendations on selection of experts, how their beliefs should be elicited and how to aggregate their opinions [15, 25]. Their main recommendations are briefly described below.

According to the Australian guidelines for expert elicitation[15], it is recommended that a random set of experts is chosen instead of panels created by sponsors, because the generalisability of the latter is hard to assess. It is also suggested that, if possible, a large number of experts should be consulted and that any conflicts of interest should be declared. Further, an assessment of whether the non-respondents are likely to reduce the representativeness of the sample of experts should take place and any background information, provided to experts prior to the conduct of the exercise, should be included in the submission. PBAC also recommends that the potential for bias in the questions asked should be assessed. Finally, where opinion has been elicited from multiple experts, their concordance should be evaluated and sensitivity analyses should be conducted around the overall estimates derived from experts as well as using other sources of evidence, if available.

Without specifying particular recommendations, PBAC also requires that detail is provided with respect to the way that experts were approached and any incentives offerred. It also states that any tools used should be described, along with details of any piloting of these prior to the conduct of the exercise. In addition, the medium that was used to collect expert judgements should be stated and any iteration methods used (such as the Delphi method) explained in detail. Finally, PBAC requires that when experts are not in agreement, the approach used to derive the final estimates and present the variability across expert' opinions should be clearly presented.

The Portuguese guidelines[25] also recommend that the process of selecting experts should be clearly described and any conflicts of interest declared. They further suggest that all experts should have the necessary skills to elicit the required quantities and the degree of expertise should be justified for each expert. The expert sample should strive to be representative of the range of possible views/settings. Between-expert differences should be presented, explained, and reflected in the uncertainty over the pooled estimate. In what concerns the conduct of the elicitation exercise, INFARMED recommends that experts should be asked to express their opinions quantitatively and only quantities that are observable by experts should be elicited. Importantly, elicited parameters must be fit-for-purpose to ensure mathematical coherence with other parameters of the model, respecting the structure of the cost-effectiveness model and that future collection of empirical evidence is considered for all elicited parameters. In addition, it is suggested that the elicitation process is designed in a transparent way that attempts to minimise biases and heuristics and any protocols used in that process are summarised and reported. Finally, the guidance document recommends that a pooled estimate that considers all experts' opinions is used in the base-case. Ensuring consistency in the methods used across assessments is deemed important and the Portuguese methods guide. Therefore it presents a reference case for methods of elicitation to be used in all submissions (see Table 2). However, the guidance incentiveses companies to also submit other analyses where justified. This can include group opinions derived by consensus using, for example, the Delphi process.

Table 2: Reference methods of elicitation of expert opinion in the Portuguese guidelines

Component	Reference methods
Experts	Substantive expertise or experience, and training in elicitation techniques and methods
Quantities elicited	Observable quantities to the experts, elicited under uncertainty
Approach to elicitation	Individual elicitation (with no, or limited, interaction between experts)
Method	Chips and bins (preferred)
Aggregation	Linear pooling with equal weighting of experts
Delivery	Face-to-face where possible to allow a facilitator to deliver training to the expert

5. How has elicitation been used in HTA?

Expert judgements can be gathered in diffierent ways which can range from informal requests to fomal and structured processes.

5.1. Overview of the use of expert elicitation in HTA

A recent review[26] of 25 company submissions appraised by NICE (single technology appraisal, STA, process) found that expert judgement is ubiquitously used in company submissions. Out of the 25 submissions, only 2 did not report using any form of expert judgement to inform or validate the decision model or its inputs. Most commonly, the submissions referred to gathering expert judgements using informal, unstructured methods, that can be described as either ad-hoc methods (55.1%) and advisory boards (38.8%).

Additionally, a review of the current use of experts within NICE guidance-making programmes, conducted interviews with NICE staff to determine the principal ways in which experts are involved in guidance making in each of the NICE programmes. Again the distinction was made between expert opinion (a qualitative process) and expert elicitation (a quantitative process). The review concluded that, generally, expert opinion is sought in an unstructured way, during committee meetings, by telephone or email. The process is flexible and questions are asked relevant to the particular topic and for anything that committee members and chairs deem appropriate. The use of expert elicitation was undertaken by all NICE programmes, but only in particular circumstances. The use of expert elicitation is thought to be less common than the use of expert opinion and is generally restricted to cases where there are significant gaps in the evidence base.

A previous review [9] and subsequent update[1], looked at the use of formal and structured methods of elicitation in economic evaluation. The review found only a limited number of published examples. To provide a more in depth look at how structured elicitation has been used in HTA, we have here updated the abovementioned reviews of applications of expert elicitation for economic evaluation.

5.2. Review of applied studies of structured expert elicitation

5.2.1. Methods

The review conducted by Soares, et al [1] was updated, using the same search strategy (Appendix B:) and the same data extraction tool. For the update, searches were conducted in Ovid SP (Medline).

Articles were included for full text screening if they described the use of expert opinion with the purpose of informing a cost-effectiveness model/analysis. Articles screened in full-text were excluded if they aimed to elicit only point estimates (not probability distributions), or did not describe any aspect of the elicitation methods used. Both structured (using a pre defined set of methods) and non-structured applied examples were included. Where several papers described aspects of the same elicitation exercise, only the reference containing the most complete account of the elicitation was included.

The data extraction tool aimed to collect information on the methodological choices made in each of the exercises. For each of these fields, information was extracted on the justification for the choices (using an open ended field). An additional open field was used to extract challenges with the design, conduct and analyses of the SEE, reported throughout the text and in the discussion. These were later categorised and grouped for reporting.

5.2.2. Results

As illustrated in the PRISMA flow chart (Figure C-1 in Appendix B:), 95 articles were identified and screened against the inclusion and exclusion listed below. Overall, 4 studies fulfilled the inclusion criteria and were added to the 21 studies identified in [1]. The full list and titles of the included studies is presented in Appendix C: and a brief summary of the main characteristics of the 25 included is provided in Table 3Table 3: A summary of the main characteristics of the identified applications.

The following sections describe the details of the applied studies. Specifically, the parameters elicited using which quantities, the methods of elicitation and methods of analysis

<u>What parameters and quantities are elicited?</u>

Across the identified studies, expert opinion was elicited for the following parameters of decision models: event probabilities [23, 27-44], counts of events [42, 45], time-to-event [27, 29, 30, 45, 46], diagnostic accuracy [28, 31, 47], minimum important clinical difference [31], and relative effectiveness [28, 30, 33, 35-37, 40, 48]. Importantly, model input parameters were often not elicited directly but calculated from one or more elicited quantities. For instance, a time-constant transition probability could be indirectly elicited by asking experts for the mean time at which an event is observed.

The choice of quantities to be elicited was based on three main criteria. The first was appropriateness for experts. For instance, given that decision model parameters may be complex and not directly observable by experts, relative effectiveness parameters may be elicited as probabilities [28, 31, 37, 48] or sensitivities and specificities as probabilities of the true disease status conditional on the test results [31]. Second, elicited quantities need to be fit-for-purpose, allowing elicited evidence to be combined with existing empirical evidence and ensuring statistical coherence between elicited quantities [27, 37]. For example, mutually exclusive events probabilities need to be simultaneously elicited to guarantee that they sum to 1 [42, 43]. Finally, in order to reduce burden to experts, the number of parameters to be elicited should be limited and parameters should be homogenous throughout the exercise (e.g. all probability parameters) [37].

How were quantities elicited?

An important requirement for model-based economic evaluation is the need to elicit uncertainty of experts' judgements in the form of a distribution. This implies that a number of summaries need to be elicited for each quantity to define the shape of a distribution. To do this, applications have typically used one of two approaches: Fixed interval method (FIM) [23, 28, 31, 32, 34, 37-39, 44, 48, 49] or variable interval method (VIM) [27, 33, 35, 36, 41-43, 45-47]. In an FIM, experts are provided with ranges of values and asked to assess the probability that the quantity lies in each. In a VIM, experts are asked to specify values of the quantity of interest for predefined percentiles of the distribution. Although one application chose FIM because the literature suggested that it returns higher variance [28], it was more common for authors to consider both approaches.

The majority of studies individually elicited experts' beliefs and subsequently aggregated them mathematically [23, 27, 28, 31, 32, 35-39, 44, 45, 47, 48, 50], whilst a small number of applications aimed to achieve consensus among experts [29, 33, 41, 46] or did not explicitly report the method of aggregation used [30, 34, 40]. Authors who preferred mathematical aggregation, did so to reflect variation within and between experts [31], and because consensus is known to potentially lead to overconfident results [28], incoherent probability statements [37], and practical difficulties in convening experts [43].

All applications recruited health care professionals based on their knowledge and clinical experience [23, 28, 32, 35, 37, 42-47], research experience [28, 35, 43, 45], relevant jurisdiction recognition by peers [28], lack of involvement in product development [47]. Some applications recruited experts from various specialties [28, 31, 40, 42-44, 46], clinical settings [28, 37, 40, 42, 43] and geographical areas [28, 43, 45] to capture heterogeneity in beliefs. Sample sizes ranged from 2 [48] to 23 [37].. Across applications, constraints to experts sampling related to limited available resources [28] and numbers of relevant experts [35], geographic distance [23, 35].

Across studies, two types of biases were recognised: motivational and cognitive. The former relates to conscious or subconscious distortions of judgements, because of self-interest, and the latter due to the use of heuristics i.e. cognitive shortcuts that individuals use when asked for complex judgements. Two applications made explicit efforts to avoid recruiting experts who may have motivational biases [40, 47] and three studies provided information on cognitive biases in the training session [37, 39, 42].

How were results analysed?

Most applications that used mathematical aggregation applied equal weighting across experts, whilst five studies explored differential weighting based on experts' responses to 'seed' questions (performance-based weighting) [23, 37, 39, 45], or on experts' clinical background (objective weighting) [47].Overall, consensus exercises were typically face-to-face group exercises, whilst studies using mathematical aggregation adopted a mix of formats, ranging from individual interview to remote completion via email.

Generally, inputs to decision models were pooled across experts, except in one study that ran the model with each of the individual elicited distribution and linearly pooled resulting outputs [48]. Of the 18 studies that aggregated experts' beliefs mathematically, 1 did not generate a group estimate and instead used the responses of each expert individually [27]. Thirteen linearly pooled by averaging individual distributions (with or without weighting). Authors justify this choice on the basis of the lack of published evidence that more complex methods outperform linear pooling [28]. Two studies used the predictive distribution from a random-effects meta-analysis of individual elicited distributions [23, 36]; however, given that the random-effects model results in a combined distribution that can be more precise than any of the individual distributions, this pooling method has been deemed inappropriate for use in model based economic evaluation [23].

In some applications, elicited evidence was used directly as inputs to the cost-effectiveness model [27-29, 31, 39, 40, 42-44]. When external evidence existed on elicited parameters, some authors presented both sources separately using scenarios [31, 33], whereas other combined them using Bayesian updating [30, 37, 41]. The latter is consistent only under the assumption that experts did not consider existing evidence when formulating their judgements [37].

Across applications, three types of validity checks have been implemented. One study [35] contrasted qualitative and quantitative responses in order to identify inconsistent responses (internal validity). For instance, the statement "I don't know, this isn't my area of research" was accompanied by extremely certain probability estimates. A second type of validity check compared the elicited beliefs across experts [27, 34, 37, 45, 48, 49], with some studies valuing good agreement [27, 31] and others accepting variation between experts [28, 37]. Finally, when external evidence was available, this was compared with elicited beliefs (external validity) [27, 31, 37, 49].

Table 3: A summary of the main characteristics of the identified applications

Study	Type of parameter elicited	Method (summaries elicited)	Aggregation approach	Mode of administration	Format/ software
Garthwaite et.al [27]	Event probabilities, time to event, dependency	Median and quartiles	Mathematical	Individual F2F and remote (telephone) interview	Interview and specialised software
Leal et. Al [28]	Event probabilities, relative effectiveness (limited reporting), diagnostic accuracy (limited reporting)	Four complementary intervals	Mathematical	Remote (email) and individual F2F interviews	Excel-based
Girling et. Al [29]	Event probabilities, time to event	NR	Consensus	Group F2F interviews	NA
Stevenson et. al [30](surgical)	Event probabilities, time to event, relative effectiveness	NR	NR	NR	NR
Meads et. Al [31]	Event probabilities, diagnostic accuracy, minimum important clinical difference	Chips and bins	Mathematical	Group F2F and individual F2F interviews	Paper
McKenna et. Al [32]	Event probabilities	Chips and bins	Mathematical	NR	Excel-based
Haakma et. Al [47]	Diagnostic accuracy	Mode and 95% CI	Mathematical	Individual F2F interviews	Excel-based
Stevenson et. Al [33] (RCT)	Event probabilities, relative effectiveness	Median and quartiles	Consensus	Group F2F	NR
Speight et. Al [34]	Event probabilities	Chips and bins	NR	NR	Paper
Sperber et. Al [35]	Event probabilities, relative effectiveness	Median and quartiles	Mathematical	Remote (telephone) interview	Excel-based
Brodtkorb et. Al [49]	Performed several elicitation exercises, but detail on each is sparse	NR	NR	NR	NR
Colbourn et. Al [36]	Event probabilities, relative effectiveness	Mean and 95% CI	Mathematical	NR	NR
Soares et. Al [37]	Event probabilities, relative effectiveness	Chips and bins	Mathematical	Group F2F interviews	Excel-based

Bojke et. Al [23]	Relative effectiveness, dependency	Chips and bins	Mathematical	Individual F2F	Excel-based
Cao et. Al [48]	Relative effectiveness	Mode and one percentile	Mathematical	NR	NR
Fischer et. Al [45]	Counts, time-to-event	Median and 80% CI	Mathematical	Group F2F, individual F2F and remote (telephone) interviews	Paper
Poncet et. Al [38]	Event Event probabilities	Chips and bins	Mathematical	NR	NR
Grigore et. Al [39]	Event probabilities	Chips and bins + four complementary intervals	Mathematical	Individual F2F interviews	Excel-based
Wilson et. Al [40]	Event probabilities, relative effectiveness	NR	NR	NR	SHELF
Meeyai et. Al [41]	Event probabilities	Mode and quartiles	Consensus	NR	SHELF
Grimm et. Al [50]	Diffusion-related*	NR	Mathematical	NR	NR
Rossi et. al [42]	Event probabilities	Median and 95% CrI	Mathematical	Individual F2F and remote (telephone) interviews	Excel-based
Wilson et. Al [43]	Event probabilities	Median and 95% CI	Mathematical	Remote (webinar)	Excel-based
Vargas et. Al [44]	Event probabilities	Chips and bins	Mathematical	NR	NR
Cope et. Al [46]	Time-to-event	Most likely value, upper plausible limit and lower plausible limit (operationalised as the 99% CI)	Consensus	Remote (web-based application)	Based on SHELF

NR: Not reported, F2F: Face-to-face

6. Reporting guidelines

Reporting guidelines for expert elicitation may encourage harmonisation of methodologies. In this Section we report a review of the relevant literature

6.1. Relevant literature

A paper published in 2016 outlined suggested criteria for the use of expert judgement in model based economic evaluation [2]. To consider other possible sources of reporting guidance, citation searches were performed using this paper. In addition, Pubmed was searched using the terms "expert" "judgement" "opinion" "elicitation". These searches identified 26 potentially relevant publications which cited Iglesias, et al, 2016 [2]. None of these publications concerned reporting guidance relating to health or healthcare.

6.2. Summary of exiting reporting guidelines

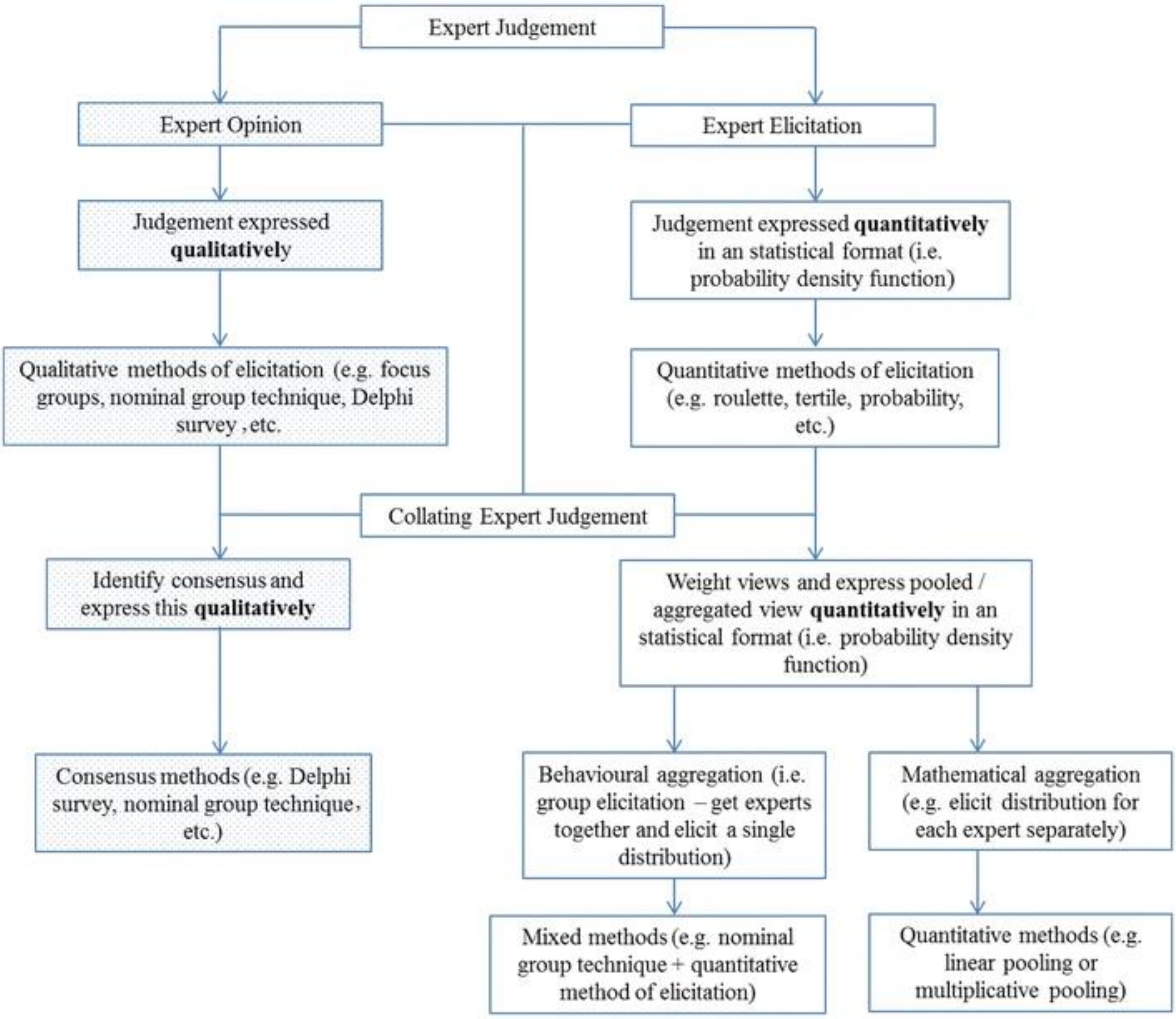
Iglesias, et al, 2016 [2] is the only existing reporting guideline specific to health. This guideline was developed with experts in the field, in particular methodologists with experience of conducting elicitation. They suggest the flow chart shown in Figure 2 as nomenclature for the use of expert judgement. This shows the distinction between expert opinion and expert elicitation, with the latter describing a quantitative process, for example to generate parameter values to populate a decision analytic model. Delphi studies are assumed to concern qualitative opinion, although there is recognition in the criteria that the rounds may involve a quantitative structure.

The guidelines constitute 16 statements for reporting of studies eliciting parameter values or distributions using a structured process. The reporting criteria for an expert elicitation study are presented in Table A-1. These comprise the details of the intended research, the quantities to be elicited, the expert sample, preparation of the exercise including training, data collection, aggregation, calibration, presentation and interpretation of results and ethical issues. The full criteria for expert elicitation is reported in Appendix A: .

For Delphi studies, 11 statements for reporting are suggested. Delphi surveys here are intended to generate a consensus, where the evidence relevant to a particular topic area is either inconclusive or contradictory. The criteria constitute: details of the intended research, details of the evidence base, the expert sample, data collection, number of rounds, data analysis, presentation and interpretation of results and ethical issues. The full criteria for expert elicitation is reported in Appendix A: .

Both reporting criteria do not offer details of what methods should be just to generate expert opinion or judgement, although this may be explicit in the notes that accompany the criterion, for example data aggregation in the expert elicitation reporting criteria refers to the use of mathematical or behavioural aggregation and the use of specific methods such as opinion pooling. For other criterion, details are less defined, such as the methods used to collect data, in particular where distributions are sought from experts.

Figure 2: Suggested nomenclature for expert judgement[2]



7. Existing protocol for the use of structured elicitation in HTA

A recent project funded by the MRC [3] considered the development of a protocol for elicitation in health care decision making. Across the project, a mixed-methods approach was used including a systemic review, targeted searches, experimental work and narrative synthesis. The systematic review identified existing guidelines for structured expert elicitation. This identified approaches used (the 'choices') across different elements of elcitation (across its design, conduct, and reporting) and determined if any dominant approaches exist. Targeted review searches were conducted for selection of experts, level of elicitation, fitting and aggregation, assessing accuracy of judgements and heuristics and biases. The experiments focussed on understanding the accuracy of the method of elicitation, the accuracy of elicitations extrapolating from experts' knowledge base, and understanding how experts revise their estimates, when requested to (e.g. in a Delphi type process).

This work found that there was a lack of consistency across the existing guidelines. In almost all choices, there was a lack of empirical evidence supporting recommendations. The experiments provided some new evidence on the use of FIM and VIM to elicit experts judgement, the impact of between expert heterogeneity and the ability of experts to generalise outside of the setting that they observe.

The experiments showed that FIM and VIM performed similarly although there was some preference for the chips and bins. The second experiment did not show evidence that extrapolation outside of setting, or its level, affects bias, expressions of uncertainty or overall accuracy. The third experiment looked at how and why individuals revise their answers when presented with Delphi-type summaries. Participants were more likely to revise their priors when the group was discordant with their own beliefs and when the group was more certain than they were. Participants were also more likely to revise their priors when the group members were consistent among themselves (although with wider within-participant uncertainty) than when the group elements were inconsistent among themselves (although more precise). In addition, participants were more likely to revise their priors when using chips and bins than bisection. There was no evidence that those who revised have different accuracy to those that who did not revise, which casts some doubts on the benefits of the Delphi iteration process.

To understand the appropriateness of the available choices in the context of health care decision making, a set of principles that underpin the use of structured expert elicitation was defined with particular consideration for constraints in health-care decision-making context. These included, for example, transparency, consistency, the need to reflect uncertainty and the need to recognise and act on biases, amongst others. The principles and the findings from the other components of work were then applied to the set of choices available from existing guidelines to establish a set of reference methods for structured elicitation of quantitative judgements. The reference methods are presented in Table 2. Note that in some circumstances, the principles were unable to provide sufficient justification for discounting particular choices, thus multiple options remain.

8. Discussion

In HTA, expert judgements are commonly used without formal consideration for the methods used to gather this evidence. Structured methods have been sparingly used, even when quantitative descriptions are generated for direct use in models submitted to HTA agencies (Section 3.2).The lack of methodological guidance on elicitation may be responsible for the generalised use of informal, and less onerous, methods. However, it is important to recognise that expert judgement (opinion and elicited quantitative judgements) constitutes the lowest form of evidence being associated with a high likelihood of bias, primarily motivational and cognitive biases. Our survey showed that decision makers downweight the evidence they consider from experts due to perceived sources of biases. However, like in the design of experiments, structured methods are recognised to be more appropriate as they attempt to minimise biases and ensure the tasks requested of experts are clear and accessible. In contrast, unstructured processes are ad hoc, they do not ensure biases are minimised and do not ensure consistency across assessments and decision making processes in NICE.

It is, however, important to recognise that experts are relevant at all stages of HTA, from topic generation, scoping and definition of the decision problem, to the discussion leading to the final decision. Whilst fully protocolled exercises (pre-planned, where for example the quantities to elicit are defined a priori) are desirable, 'responsive' judgements will always be required from experts. Some level of structuring of expert elicitation is possible in 'responsive' situations, for example, by appropriate training of the clinical experts and of chairs of meetings, and by pre-defining (protocolling) some general aspects, such as defining a set of alternative ways to ask about uncertainty, and providing a list of do's and don'ts aimed at minimising biases and heuristics.

What are the strengths and limitations of the elicitation exercises developed to date?

The review of applied studies in HTA identified only examples of structured expert elicitation and highlighted a number of their limitations; namely that applications use a wide range of methods; reporting is generally poor; and that although it is generally agreed that SEE should be designed and conducted in a way that minimises the use of heuristics and other sources of bias, there is little integration of the findings from behavioural research. Applications acknowledge that health care professionals might have only sparse quantitative skills [27, 37, 39, 43] and that, in health care, between-expert variation is expected because of genuine heterogeneity in the populations experts draw upon. Guidance on the use of expert elicitation should have appropriate consideration for how this variation can be reflected in the group estimates obtained. This update of the review conducted here identified that more of the recent applications use a generic elicitation protocol, the SHELF protocol [51].

What uncertainties are associated with the methods for expert elicitation when informing model parameters in HTA?

The MRC project [3] highlighted that there is very limited knowledge regarding the accuracy of alternative methods of elicitation. This means it is difficult to establish where significant uncertainties arise and what methods are more appropriate. However, based on existing methods research and on the principles of elicitation for HTA, this work was able to do a number of considerations on the required choices [3]. We here summarise these considerations, presenting these for the elements of elicitation (Figure 1) in turn.

<u>Identification of variables for the elicitation</u>

The choice of parameters to elicit should be limited to those which have the biggest impact on the decision and/or decision uncertainty. This can be informed by sensitivity analysis to identify which parameters are most influential. Quantities elicited should be observable, such as probabilities (expressed as proportions or frequencies), but not more complex quantities, such as higher moments of a distribution, odds ratios or credible ranges. Instead complex quantities should be decomposed into simpler, observable quantities.

Dependency between parameters being elicited should be considered, either by attempting to ask only about independent variables by expressing dependent variables in terms of independent variables, or by using dependence elicitation methods. The wording of the questions should avoid vagueness; ask questions in a manner consistent with how experts express their knowledge; use neutral wording, avoiding leading questions; and attempt to decompose into simpler quantities when possible.

In terms of capturing judgements about uncertainty, existing guidelines suggest both the fixed interval (FIM) and the variable interval (VIM) can be used to capture judgements. Empirical evidence does not show substantive differences between these methods in terms of accuracy in representing experts beliefs, but individuals have expressed a preference for FIM. A decision-maker should, however, apply one or the other consistently in their setting. Training should always be provided to experts on how to express uncertainty using the method of choice.

Identification of experts

Little is known about how different individuals perform in elicitation. It is, however, generally recognised that a number of skills may be required of them when asked to express their quantitative beliefs, namely: substantive, normative and, in some cases, adaptive skills [52] . There is no evidence on how to identify individuals with normative skills [53]. The focus hence should be on gathering substantive expertise or experience, and developing normative skills during the training session as part of the structured expert elicitation. In recruiting HTA experts, individuals should be selected on the basis of their substantive expertise and willingness to participate. Behavioural research recommends recruiting experts that are free from motivational biases when possible. In all instances, information on personal, financial and conflicts of interest should be collected. A range of viewpoints, for example work setting, should be represented in the sample, with the intention of 'balancing out' or at least diluting the effect of motivational biases. Whilst there is no consensus on how many experts should be included, we here consider that to strive for a range of viewpoints at least five experts should be recruited.

Training and preparation

Piloting of the elicitation exercise should be undertaken prior to the task, and the feedback from this used to revise the questions and format. A proportion of the elicitation exercise should be spent on delivering training, as it is unlikely that experts will have had any previous experience of structured expert elicitation. Training should include a focus on (1) enabling experts to experts their uncertain belief and (2) explaining potential biases in order to minimise them, such as effect of anchoring, adjustment in interval, confirmation bias and overconfidence.

The conduct of the elicitation

Existing guidelines are inconsistent with respect to the level of elicitation – individual or group based. Group discussion can aid less substantive and normative experts; however, face-to-face discussion can be resource and time intensive. Interaction between experts can introduce biases, and in HTA there may be a lack of experienced facilitation to appropriately manage these biases. In these circumstances, an individual-level elicitation may be more appropriate. In the event that consensus approaches are required, an individual elicitation should always be conducted first, followed by group consensus.

Feedback should always be offered to experts with the possibility of revision. Qualitative data from experts on how they formulated their judgements should always be collected (termed *rationales*).

Port-elicitation methods

In order to generate distributions that are fit for purpose, aggregation is preferred over no aggregation. For mathematical aggregation, simple mathematical decision rules, like linear opinion pool with equal weights, are the most commonly applied in HTA. These are straightforward to implement and are thought to allow variability between experts to be reflected in the overall distribution. It is unclear how calibration and adjustment methods [54, 55] fare and further research is needed to advise if and when choices beyond equal weighting are warranted. In terms of documentation, this should be thorough and cover all aspects of the SEE design, conduct and analysis.

Managing bias

In striving to minimise bias, efforts should be made to identify which biases are likely for the sample of experts included, and relevant strategies to minimise these (bias reduction techniques) should be employed. Discussion with experts can help to identify potential biases. The framing of questions can minimise bias and ambiguity. This can include providing experts with relevant background evidence.

Validation

There are differing definitions of validity and discussion of how the concept can be operationalised in an elicitation. Ensuring that the elicited beliefs are fit for purpose could be assessed by its coherence and consistency. The use of internal and external review can help determine if the resulting distributions are valid. There should be an exploration of the implications of between-expert variation and attempt to understand why it is present.

Considerations for the development of NICE guidance

Consistency in methods is important to ensure consistency in decision making. Whlst most HTA guidance documents consulted do not make recommendations for the methods of elicitation permitted, in an attempt to ensure consistency, the Portuguese guidance (Section 4.2) defines a reference set of methods (whilst encouraging the use of alternatives). These recommendations, on the whole, accord with the recommendations from the MRC protocol.

It should be noted that the MRC protocol is the most comprehensive guidance generated to date and is the only one specific to the health care decision making context, focussing primarily on HTA. NICE guidance should be based on this protocol.

9. Proposed recommendations for NICE and their implications

Recommendations

1. Expert judgement is currently used to support decision making in a number of different ways, ranging from qualitative judgements that resolve features of existing evidence (expert opinion) to quantitative estimates where evidence does not exist or is not appropriate (expert elicitation). The context for the use of expert judgement varies, ranging from responsive processes (for example, at scoping or committee meetings) to expert judgement being used to obtain quantitative judgements used as part of the evidence generation.
2. Quantitative judgements requested of experts should attempt to ascertain the level of uncertainty in knowledge so that decision making can recognise uncertainty
3. Experts judgements should always be gathered using a structured process, using pre-specified protocol or guide:
    3.1. For quantitative judgements used as part of evidence generation activities, a pre-defined protocol for structured expert elicitation should be used (more detail below).

3.2. In a responsive processes (for example, where expert judgement is requested at committee discussions), a structured general protocol – a guide – could still be used as a reference point, acknowledging that it will not be tailored to the specific circumstances (more detail below).

4. Expert judgements may be used to inform multiple aspects of the assessment of effectiveness and cost-effectiveness, such as: treatment effects over the short term, for example when existing experimental evidence is limited; extrapolation of treatment effects to longer term (beyond the follow-up of existing experimental evidence); natural history of disease over time; costs and resource use; HRQoL and adverse events; in describing the eligible population or the relevance/generalisability of evidence; establishing care pathways for both the intervention and comparators; describe uncertainty in clinical behaviours, intervention effects and outcomes; ascertain patient issues and real-world experience; help support links between surrogates and clinical outcome; and to comment on the completeness or otherwise of the literature searches.
5. It can also be used to support decisions within the following features of the decision problem. Where the population is significantly heterogeneous, e.g. multiple subgroups or multiple indications, to assess whether the model (and evidence underlying it) fully reflects the expected level of heterogeneity in clinical practice. Where there is uncertainty about the population which may benefit from the technology, expert judgement could be used to assess whether the model (and evidence underlying it) fully reflects the eligible population expected in clinical practice. Where there is uncertainty about the natural history of the condition, expert judgement could be used to assess whether the model (and evidence underlying it) fully reflects the natural history expected in clinical practice. Where there is uncertainty about treatment/diagnostic pathways, expert judgement could be used to validate pathways modelled. Where there is uncertainty about the technology or the mechanism of action of the treatment, expert judgement could be used to support assumptions regarding the effects of the technology/treatments. Where there are multiple alternative positions within treatment/diagnostic pathways, expert judgement could be used to support recommendations regarding different lines of treatment. It could also be used to inform other aspects, such as impact on carers, issues with the real world implementation of the technology/intervention and infrastructure requirements, and on generalising evidence from other countries/jurisdictions to the context of interest.

<u>The use of expert elicitation as part of evidence generation activities</u>

6. To ensure consistency across assessments, a set of reference methods should be used to guide the development of protocols of elicitation. These are summarised in the following table:

Table 4: Reference methods for elicitation

Element	**Reference methods for elicitations conducted as part of evidence generation activities**	**Use in responsive elicitations?**
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.	1-3. yes, used as principles in recruitment

	2. Focus on gathering substantive expertise or experience. Normative skills can be developed during the training session as part of the elicitation. 3. Minimize and record conflicts of interest among the experts. Include experts external to the task, i.e. not those involved in developing the task. 4. At least 5 experts should be included.	4. As many experts as feasible
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.	Recommendations as part of a general guide to elicitation
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 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.	Recommendations as part of a general guide to elicitation
Method	1. Both VIM or FIM work well, however decision makers should aim for consistency across applications.	Pre-defined alternative methods of expressing uncertainty could be included in the guide
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.	Not likely to be applicable in a responsive context
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.	Recommendations as part of a general guide to elicitation

Training & piloting	1. Training is crucial and should focus on avoiding bias and expressing uncertainty. 2. Piloting should be undertaken.	Training provided to clinical experts and chairs of meetings
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.	Recommendations as part of a general guide to elicitation

7. The use of expert elicitation in evidence generation activities should always be accompanied by detailed reporting of methods, processes and results of the elicitation. It should include the following documentation:
   7.1. the pre-specified protocol for the elicitation, which should include:
      7.1.1. an identification of variables for the elicitation: describing the model input parameters elicited, the quantities were used to elicit these and the wording of the questions.
      7.1.2. An identification of experts: definitions of experts used, the sample of experts and training provided to them
      7.1.3. The conduct of the elicitation: methods used to capture uncertain judgements, the approach to elicitation (individual or bahavioural), and the mode of administration
      7.1.4. Post-elicitation methods: such as mathematical pooling and the fitting of distributions, and how the distributions generated are to be used withing the decision model
      7.1.5. Strategies adopted throughout for the management of bias
      7.1.6. Any validation attempts
   7.2. The questionnaire used (or a script, if delivered by interview) should be provided
   7.3. A summary of the conduct of the exercise (e.g. who facilitated the exercise, any deviations from protocol accompanied by justification, etc)
   7.4. A detailed description of the results of the elicitation, which should provide not only the group distribution but also responses from each of individual expert. The pooled estimate should be used in the base case, and differences between experts in the values provided should be presented and explained (it may be useful to ask the individual experts to provide rationales for the elicited values). The implications of between-expert differences to cost-effectiveness should be quantified using sensitivity analyses.

<u>The use of expert judgement in responsive situations</u>

8. A typical example of the use of expert judgement in a responsive situation relates to clinical specialists present to support discussions at meetings. Whilst the presence of clinical experts at such responsive processes is essential, some structuring and standardisation should be implemented in an attempt to minimise biases in the evidence provided by experts.
9. Experts should be provided with documentation about elicitation prior to the meeting, identifying the common biases associated with expert opinion and providing information about uncertainty and how to express it. In addition, they should be briefed on those key issues prior to the meeting.
10. A guide to elicitation should be provided to committee members, which would cover:

10.1. Generic information about elicitation and a number of practical considerations to have in responsive processes, for example highlighting that experts best understand quantities that are observable to them, and identifying a range of bias minimisation strategies. The chairs of meetings, which usually address the clinical experts, could receive more in depth training on elicitation.

10.2. This should also be included in the induction for new committee members. The guide should also describe alternative methods to elicit uncertainty. This would enable some level of standardisation and increase accountability. It would also allow clinical experts to be trained on these specific methods prior to committee/scoping meetings.

11. A set of minimum reporting requirements for expert elicitation for transparent decision making should comprise of reporting of methods used in the elicitation, individuals involved in the elicitation and reporting of its results.

<u>Other considerations</u>

12. A number of other considerations may be of relevance for particular decision problems or decision contexts. Elicitations must appropriately justify relaxing the use of the reference methods identified above. Examples are:

12.1. Researchers may have limited access to sufficient experts, for example in rare diseases, therefore expert recruitment may be more challenging and have to rely on peer nomination.

12.2. Adaptive skills may be required for new technologies since indirect evidence may outweigh directly relevant evidence (e.g. childhood diseases may be informed by adult versions with some extrapolation).

Implications of the recommended changes

It is important to acknowledge that whilst formal and structured processes of expert elicitation are desirable the recommendations may have a number of implications for processes and their timelines.

Structured and formal processes to be used as part of evidence generation activities are timely to conduct. The evaluation of the MRC protocol [3] identified that these could take as much as 5 months for a full time researcher to design and conduct the exercise. However, with structured protocols being increasingly used, standardisation of the protocols will allow significant reductions in the time required to design such exercises. We suggest that companies and assessement groups plan prospective structured expert elicitation to realistic timelines. However, there may still be implications for assessment groups, in processes with short timelines, such as STAs. In these cases, assessment groups may need to use the reference methods for the responsive process. In such constrained processes, elicitations can be restricted to a small number of quantities, those key to decisions.

Another important implication relates to the skill set of analysts preparing evidence, and the needs for training of committee members and chairs. This should be mitigated by better guidance and will quickly reduce over time as experience increases.

Areas for further consideration

The recommendations above present two scenarios for the use of expert elicitation, one where elicitation is used for evidence generation where a comprehensive protocol is recommended, and another where expert judgement is required as part of a responsive situation where a pre-specified detailed protocol cannot be defined. However, expert judgement is required across anumber of different NICE appraisal processes and, within these, at different stages. Careful consideration of how the opinions of experts can be gathered using more formal and structured methods within each programme/stage of the appraisals is required, and can involve reference methods that fall in between the two approaches defined here.

10. Any equity/equality considerations

None

Acknowledgements

The project team would like to thank collagues at the Centre for Health Economics for their help in devising the questions for the survey. In addion, we would like to thank the following committee members and chairs for their responses and valuable feedback: Amanda Adler, Nicholas R Latimer, Gary McVeigh, Brian Shine, Lindsay Smith, Alan Maryon-Davis.

References

1. Soares, M.O., et al., *Experiences of Structured Elicitation for Model-Based Cost-Effectiveness Analyses.* Value in Health, 2018. **21**(6): p. 715-723.
2. Iglesias, C.P., et al., *Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations.* Pharmacoeconomics, 2016. **34**(11): p. 1161-1172.
3. Bojke, L., et al., *Developing a reference protocol for expert elicitation in healthcare decision making.* Health Technology Assessment Reports, 2019. **In Press**.
4. Aspinall, W.P. and R.M. Cooke, *Quantifying scientific uncertainty from expert judgement elicitation*, in *Risk and Uncertainty Assessment for Natural Hazards*, J. Rougier, L.J. Hill, and S. Sparks, Editors. 2013, Cambridge University Press: Cambridge. p. 64-99.
5. Authority, E.F.S., *Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment.* EFSA Journal, 2014. **12**(6): p. 3734.
6. Babuscia, A. and K.-M. Cheung, *An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2014. **177**(2): p. 475-497.
7. Clemen, R.T. and R.L. Winkler, *Combining Probability Distributions From Experts in Risk Analysis.* Risk Analysis, 1999. **19**(2): p. 187-203.
8. Nemet, G.F., L.D. Anadon, and E. Verdolini, *Quantifying the Effects of Expert Selection and Elicitation Design on Experts' Confidence in Their Judgments About Future Energy Technologies.* Risk Anal, 2017. **37**(2): p. 315-330.
9. Grigore, B., et al., *Methods to Elicit Probability Distributions from Experts: A Systematic Review of Reported Practice in Health Technology Assessment.* PharmacoEconomics, 2013. **31**(11): p. 991-1003.
10. Institute for Clinical and Economic Review, *ICER's Reference Case for Economic Evaluations: Principles and Rationale.* 2018.
11. Norwegian Medicines Agency, N.K.C.f.H.S., Norwegian Directorate of Health, *The National system for the introduction of new health technologies within the specialist health service* 2014.
12. Institute for Quality and Efficiency in Health Care, *General Methods (Version 5.0).* 2017.
13. Scottish Medicines Consortium, *Working with SMC – A Guide for Manufacturers.* 2017.
14. Products, I.-N.A.o.M.a.H., *Methodological Guidelines for Economic Evaluation Studies of Health Technologies.* 2019.
15. Australian Government: Department of Health, *Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (Version 5.0).* 2016.
16. CADTH, *Guidelines for the Economic Evaluation of Health Technologies: Canada. 4th Edition.* 2017.
17. Health Information and Quality Authority, *Guidelines for the Economic Evaluation of Health Technologies in Ireland.* 2019.
18. Haute Autorité de santé, *Choices in Methods for Economic Evaluation* 2012.
19. Zorgunstituut Nederland, *Guideline for economic evaluations in healthcare.* 2016.
20. The Agency for Health Technology Assessment and Tariff System, *Health Technology Assessment Guidelines. Version 3.0* 2016.
21. Health Intervention and Technology Assessment Program, *Guidelines for Health Technology Assessment in Thailand (Second Edition).* 2014.
22. National Institute for Health and Care Excellence, *Guide to the methods of technology appraisal 2013.* 2013.
23. Bojke, L., et al., *Eliciting Distributions to Populate Decision Analytic Models.* Value in Health, 2010. **13**(5): p. 557-564.

24. Bojke, L. and M. Soares, *Decision analysis: eliciting experts' beliefs to characterize uncertainties.*, in *Encyclopedia of Health Economics*, A.J. Culyer, Editor. 2014, Elsevier.
25. Perelman, J., et al., *Methodological Guidelines for Economic Evaluation Studies.* 2019.
26. van Hest, N., et al., *Trust the Experts? Acceptance of Expert Elicitation in the National Institute for Health and Care Excellence (NICE) Single Technology Appraisal (STA) Process.* Presented at the ISPOR 22nd Annual European Congress, 2019.
27. Garthwaite, P.H., et al., *Use of expert knowledge in evaluating costs and benefits of alternative service provisions: A case study.* International Journal of Technology Assessment in Health Care, 2008. **24**(3): p. 350-357.
28. Leal, J., et al., *Eliciting Expert Opinion for Economic Models: An Applied Example.* Value in Health, 2007. **10**(3): p. 195-203.
29. Girling, A.J., et al., *Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy.* Int J Technol Assess Health Care, 2007. **23**(2): p. 269-77.
30. Stevenson, M.D., et al., *The cost-effectiveness of surgical instrument management policies to reduce the risk of vCJD transmission to humans.* Journal of the Operational Research Society, 2009. **60**(4): p. 506-518.
31. Meads, C., 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 technology assessment (Winchester, England), 2013. **17**(12): p. 1-323.
32. McKenna, C., et al., *Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis.* Health Technol Assess, 2009. **13**(24): p. iii-iv, ix-xi, 1-90.
33. Stevenson, M.D., et al., *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.* Med Decis Making, 2009. **29**(6): p. 678-89.
34. Speight, P.M., et al., *The cost-effectiveness of screening for oral cancer in primary care.* Health Technol Assess, 2006. **10**(14): p. 1-144, iii-iv.
35. Sperber, D., et al., *An Expert on Every Street Corner? Methods for Eliciting Distributions in Geographically Dispersed Opinion Pools.* Value in Health, 2013. **16**(2): p. 434-437.
36. Colbourn, T., et al., *Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses.* Health Technol Assess, 2007. **11**(29): p. 1-226, iii.
37. Soares, M.O., et al., *Methods to elicit experts' beliefs over uncertain quantities: application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration.* Statistics in Medicine, 2011. **30**(19): p. 2363-2380.
38. Poncet, A., et al., *Electrocardiographic Screening for Prolonged QT Interval to Reduce Sudden Cardiac Death in Psychiatric Patients: A Cost-Effectiveness Analysis.* PloS one, 2015. **10**(6): p. e0127213-e0127213.
39. Grigore, B., et al., *A comparison of two methods for expert elicitation in health technology assessments.* BMC Medical Research Methodology, 2016. **16**(1): p. 85.
40. Wilson, E.C.F., G. Stanley, and Z. Mirza, *The Long-Term Cost to the UK NHS and Social Services of Different Durations of IV Thiamine (Vitamin B1) for Chronic Alcohol Misusers with Symptoms of Wernicke's Encephalopathy Presenting at the Emergency Department.* Applied Health Economics and Health Policy, 2016. **14**(2): p. 205-215.
41. Meeyai, A., et al., *Seasonal influenza vaccination for children in Thailand: a cost-effectiveness analysis.* PLoS Med, 2015. **12**(5): p. e1001829; discussion e1001829.
42. Rossi, S.H., et al., *Expert Elicitation to Inform a Cost-Effectiveness Analysis of Screening for Renal Cancer.* Value in Health, 2019. **22**(9): p. 981-987.

43. Wilson, E.C.F., et al., *Expert Elicitation of Multinomial Probabilities for Decision-Analytic Modeling: An Application to Rates of Disease Progression in Undiagnosed and Untreated Melanoma.* Value in Health, 2018. **21**(6): p. 669-676.
44. Vargas, C., et al., *Costs and consequences of chronic pain due to musculoskeletal disorders from a health system perspective in Chile.* The Pain Report, 2018. **3**(5): p. e656.
45. Fischer, K., D. Lewandowski, and M.P. Janssen, *Estimating unknown parameters in haemophilia using expert judgement elicitation.* Haemophilia, 2013. **19**(5): p. e282-e288.
46. Cope, S., 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): p. 182.
47. Haakma, W., et al., *Belief Elicitation to Populate Health Economic Models of Medical Diagnostic Devices in Development.* Applied Health Economics and Health Policy, 2014. **12**(3): p. 327-334.
48. Cao, Q., et al., *Probability Elicitation to Inform Early Health Economic Evaluations of New Medical Technologies: A Case Study in Heart Failure Disease Management.* Value in Health, 2013. **16**(4): p. 529-535.
49. Brodtkorb, T.-H., *Cost-effectiveness analysis of health technologies when evidence is scarce*, in *Linköping Studies in Science and Technology. Dissertations*. 2010, Linköping University Electronic Press: Linköping. p. 41.
50. Grimm, S.E., S. Dixon, and J.W. Stevens, *Assessing the Expected Value of Research Studies in Reducing Uncertainty and Improving Implementation Dynamics.* Med Decis Making, 2017. **37**(5): p. 523-533.
51. O'Hagan, T. and J.E. Oakley, *SHELF: the Sheffield Elicitation Framework (Version 3.0).* 2016: Sheffield: School of Mathematics and Statistics, University of Sheffield.
52. O' Hagan, A., *Uncertain Judgements: Eliciting Experts' Probabilities* 2006: John Wiley & Sons.
53. Jankovic, D., *Using Experts' Beliefs to Inform Public Policy: Capturing and Using the Views of Many*, in *Centre for Health Economics*. 2018, University of York.
54. Hora, S.C. and E. Kardeş, *Calibration, sharpness and the weighting of experts in a linear opinion pool.* Annals of Operations Research, 2015. **229**(1): p. 429-450.
55. Rufo, M.J., J. Martin, and C.J. Perez, *Log-Linear Pool to Combine Prior Distributions: A Suggestion for a Calibration-Based Approach.* Bayesian Anal., 2012. **7**(2): p. 411-438.

Appendix A: Reporting guidelines from Iglesias, et al 2016 [2]

Table A-1: Criteria for an expert elicitation study

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 ‘pre-selection’ 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 type of uncertain quantities	The rationale for the measure type 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
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 and final number of experts recruited to 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
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 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 a limited number of situations it may be feasible to collect the data using a self-administered online or postal survey but this is unlikely to be successful in most instances. Both face-to-face and telephone data collection is likely to be supported by using a computer
Training	The use of training materials should be reported and made available	This may include background training materials sent to the experts and/or training in the use of probabilities and nature of distributions. This document need to provide explanation of efforts made to prevent influencing experts' knowledge and judgement. 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. ehavioural 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	This should include an explanation of how the reader should interpret the results. It should be recognised 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

	used in the model-based economic analysis	

Table A-2: Criteria for a Delphi study

Criterion	**Description**	**Note**
Research problem	The research problem should be clearly defined and ideally framed explicitly as a research question to be addressed	When clarifying the research problem, remember the Delphi process technique is a group facilitation technique and as such only lends itself to group involvement
Research rationale	The topic and use of the Delphi process method should be justified	The Delphi Process is best used when the research requires anonymity to avoid dominance of one opinion. It should also be remembered that the strength of the Delphi process method lies in the use of iteration in which the process of gaining opinion occurs in rounds to allow individuals to change their opinion
Literature review	The rationale for using the Delphi process method must be informed by a clear description of the evidence base for the topic of the study	The focus of using the Delphi process method should be where unanimity of opinion does not exist owing to a poor evidence base. This section should also describe the process of determining the most important issues to refer to in the design of the initial round of the Delphi process
Data collection	This should include a clear explanation of the Delphi process method employed	This should be sufficiently detailed for a reader to be able to duplicate the process of conducting the Delphi process method. This includes a description of the types of questions used (qualitative or quantitative and ranking, rating or scoring scale used). This section should describe which medium was used to collect the data (electronic or written communication). This section should also describe how results from previous rounds were fed back to the experts and whether feedback is given to the group and/or individual response
The survey	A copy of each round of the survey used in the Delphi	The use of journal supplementary appendices should be exploited to allow the

	process method should be presented	reader access to a full copy of the survey used for each round of the Delphi process
Rounds	This should state the number of rounds planned and used together with the plans for moving from one round to the next	The structure of the initial round (either qualitative or quantitative) should be decided from the protocol stage of the study together with the number of rounds to be used
The sample	The sample or 'expert' panel should be described in terms of the definition of an expert in the context of the study and the selection and composition of the panel, including how it was formed from a sampling frame and response rate achieved	It should be noted that the composition of the panel will affect the results obtained from the Delphi process method. Careful thought should be given to the criteria employed to define an expert, the justification of a participant as an 'expert' and the use of non-probability sampling techniques (such as purposive or criterion methods)
Ethical issues	The ethical issues for the expert sample and research community should be described	The use of the Delphi process method should acknowledge the issues of ethical responsibility, anonymity, reliability and validity in an ongoing manner throughout the data collection and analysis process
Data analysis	The management of opinions, analysis and handling of both qualitative and quantitative data should be described	As with any other survey-based approach, a pre-specified data analysis plan should be prepared. This should include a clear description of the meaning of 'consensus' in relation to the stated aim of the study and how 'agreement' is defined. This should also take account of reliability and validity issues identified
Presentation of results	The results for each round, and final round, should be presented clearly while taking account of the audience of the study findings	The response rate for each round should be stated. Careful consideration should be paid on how to present the interim (between-round) and final results in either graphical and/or statistical representations. In round 1, a summary of the total number of issues generated should be presented. In the final round, the strength of overall consensus should be summarised. Reporting data from quantitative questions should acknowledge the limitations associated with eliciting point estimates (e.g. no indication of uncertainty)

Interpretation of results	The interpretation of consensus (not) gained should be presented together with the meaning of the results and direction of further research needed	This should include an explanation of how the reader should interpret the results, and how to digest the findings in relation to the emphasis being placed upon them. It should be recognised that the composition of the panel will affect the results obtained. The interpretation of results should state whether 'outliers' to the overall consensus were asked for the reasons for their answers

Appendix B: Search strategy for applied studies

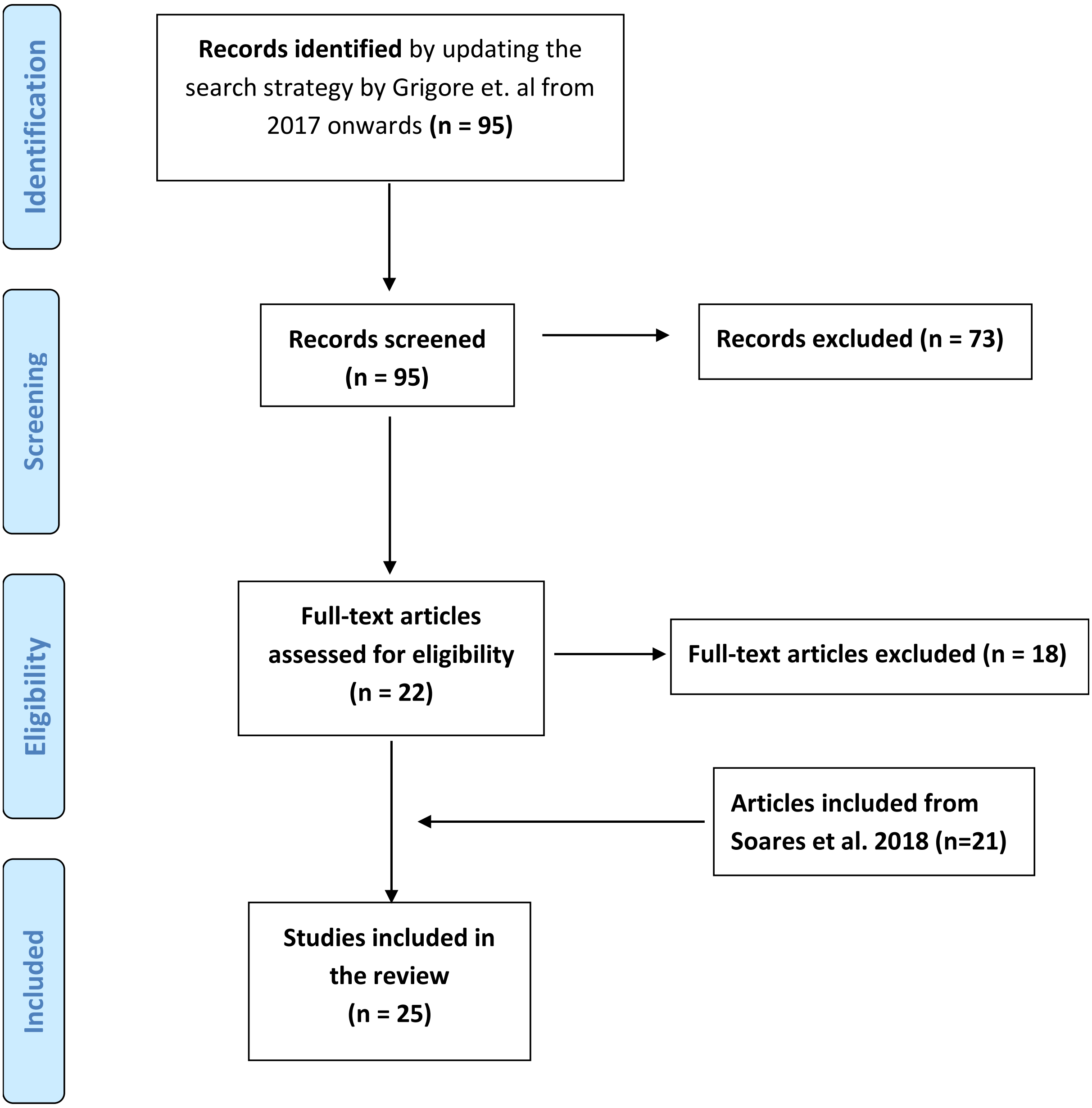
1. elicit*
2. subjective ADJ1 probabilit*
3. bayes* AND prior$
4. probability ADJ1 distribution$
5. value ADJ3 information
6. probabilistic sensitivity analysis
7. 1 OR 2 OR 3 OR 4 OR 5 OR6
8. HTA
9. technology ADJ1 assessment$
10. cost-effectiveness
11. model*
12. comparative ADJ1 effectiveness
13. 8 OR 9 OR 10 OR 11 OR 12
14. expert$ ADJ1 elicitation
15. expert$ ADJ1 opinion
16. expert$ ADJ1 knowledge
17. expert$ ADJ1 judgement
18. expert$ ADJ1 belief$
19. expert$ ADJ1 panel$
20. advisor$ ADJ1 board$

21. 14 OR 15 OR 16 OR 17 OR 18 OR 19 OR 20

22. 7 AND 13 AND 21

Appendix C: PRISMA flow chart and list of papers included directly from Soares et al. 2018 [1], identified in this update, and excluded in this review at the full-text screening step

Figure C-1: PRISMA flow chart of the review of applied studies



List of studies included in Soares et al. 2018 [1]

Leal, J., et al. (2007). "Eliciting Expert Opinion for Economic Models: An Applied Example." VALUE IN HEALTH **10**(3): 195-203.

Cao, Q., et al. (2013). "Probability Elicitation to Inform Early Health Economic Evaluations of New Medical Technologies: A Case Study in Heart Failure Disease Management." VALUE IN HEALTH **16**(4): 529-535.

Meads, C., et al. (2013). "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 Technology Assessment (Winchester, England) **17**(12): 1-323.

Haakma, W., et al. (2014). "Belief Elicitation to Populate Health Economic Models of Medical Diagnostic Devices in Development." Applied Health Economics and Health Policy **12**(3): 327-334.

Garthwaite, P. H., et al. (2008). "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**(3): 350-357.

Girling, A. J., et al. (2007). "Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy." International Journal of Technology Assessment in Health Care **23**(2): 269-277.

Stevenson, M. D., et al. (2009). "The cost-effectiveness of surgical instrument management policies to reduce the risk of vCJD transmission to humans." Journal of the Operational Research Society **60**(4): 506-518.

Bojke, L., et al. (2010). "Eliciting Distributions to Populate Decision Analytic Models." VALUE IN HEALTH **13**(5): 557-564.

McKenna, C., et al. (2009). "Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis." Health Technol Assess **13**(24): iii-iv, ix-xi, 1-90.

Wilson, E. C. F., et al. (2016). "The Long-Term Cost to the UK NHS and Social Services of Different Durations of IV Thiamine (Vitamin B1) for Chronic Alcohol Misusers with Symptoms of Wernicke's Encephalopathy Presenting at the Emergency Department." Applied Health Economics and Health Policy **14**(2): 205-215.

Meeyai, A., et al. (2015). "Seasonal influenza vaccination for children in Thailand: a cost-effectiveness analysis." PLoS Medicine **12**(5): e1001829; discussion e1001829.

Sperber, D., et al. (2013). "An Expert on Every Street Corner? Methods for Eliciting Distributions in Geographically Dispersed Opinion Pools." VALUE IN HEALTH **16**(2): 434-437.

Fischer, K., et al. (2013). "Estimating unknown parameters in haemophilia using expert judgement elicitation." Haemophilia **19**(5): e282-e288.

Grigore, B., et al. (2016). "A comparison of two methods for expert elicitation in health technology assessments." BMC MEDICAL RESEARCH METHODOLOGY **16**(1): 85.

Speight, P. M., et al. (2006). "The cost-effectiveness of screening for oral cancer in primary care." Health Technol Assess **10**(14): 1-144, iii-iv.

Brodtkorb, T.-H. (2010). Cost-effectiveness analysis of health technologies when evidence is scarce. Linköping Studies in Science and Technology. Dissertations. Linköping, Linköping University Electronic Press**:** 41.

Poncet, A., et al. (2015). "Electrocardiographic Screening for Prolonged QT Interval to Reduce Sudden Cardiac Death in Psychiatric Patients: A Cost-Effectiveness Analysis." PLoS ONE **10**(6): e0127213-e0127213.

Colbourn, T., et al. (2007). "Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses." Health Technol Assess **11**(29): 1-226, iii.

Grimm, S. E., et al. (2017). "Assessing the Expected Value of Research Studies in Reducing Uncertainty and Improving Implementation Dynamics." Medical Decision Making **37**(5): 523-533.

Soares, M. O., et al. (2011). "Methods to elicit experts' beliefs over uncertain quantities: application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration." Statistics in Medicine **30**(19): 2363-2380.

Stevenson, M. D., et al. (2009). "The cost-effectiveness of an RCT to establish whether 5 or 10 years of bisphosphonate treatment is the better duration for women with a prior fracture." Medical Decision Making **29**(6): 678-689.

List of additional studies included in the updated review (from 2017 onwards)

Cope, S., et al. (2019). "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 **19**(1): 182.

Rossi, S. H., et al. (2019). "Expert Elicitation to Inform a Cost-Effectiveness Analysis of Screening for Renal Cancer." VALUE IN HEALTH **22**(9): 981-987.

Vargas, C., et al. (2018). "Costs and consequences of chronic pain due to musculoskeletal disorders from a health system perspective in Chile." The Pain Report **3**(5): e656.

Wilson, E. C. F., et al. (2018). "Expert Elicitation of Multinomial Probabilities for Decision-Analytic Modeling: An Application to Rates of Disease Progression in Undiagnosed and Untreated Melanoma." VALUE IN HEALTH **21**(6): 669-676.

List of studies excluded in this review from full-text screening

Bauer, A., et al. (2019). "Access to health care for older people with intellectual disability: a modelling study to explore the cost-effectiveness of health checks." BMC PUBLIC HEALTH **19**(1): 706.

Bullement, A., et al. (2019). "Cost Effectiveness of Avelumab for Metastatic Merkel Cell Carcinoma." PharmacoEconomics Open **3**(3): 377-390.

Cheng, X., et al. (2017). "Health economic evaluation of immunization strategies of hepatitis E vaccine for elderly population." Human vaccines & Immunotherapeutics **13**(8): 1873-1878. sporadic region in China.

Chengat Prakashbabu, B., et al. (2018). "Risk-based inspection as a cost-effective strategy to reduce human exposure to cysticerci of Taenia saginata in low-prevalence settings." Parasites & Vectors [Electronic Resource] **11**(1): 257.

Gouveia, M., et al. (2019). "Cost-effectiveness of the 13-valent pneumococcal conjugate vaccine in adults in Portugal versus "no vaccination" and versus vaccination with the 23-valent pneumococcal polysaccharide vaccine." Human vaccines & Immunotherapeutics **15**(4): 850-858.

Gyftopoulos, S., et al. (2018). "Comparison Between Image-Guided and Landmark-Based Glenohumeral Joint Injections for the Treatment of Adhesive Capsulitis: A Cost-Effectiveness Study." AJR. American Journal of Roentgenology **210**(6): 1279-1287.

Hitimana, R., et al. (2019). "Incremental cost and health gains of the 2016 WHO antenatal care recommendations for Rwanda: results from expert elicitation." Health Research Policy & Systems **17**(1): 36.

Johansson, E., et al. (2018). "Cost effectiveness of ixekizumab versus secukinumab in the treatment of moderate-to-severe plaque psoriasis in Spain." Clinicoeconomics & Outcomes Research **10**: 747-759.

Khan, K. S., et al. (2018). "MRI versus laparoscopy to diagnose the main causes of chronic pelvic pain in women: a test-accuracy study and economic evaluation." Health Technology Assessment (Winchester, England) **22**(40): 1-92.

Meregaglia, M., et al. (2018). "The use of intensive radiological assessments in routine surveillance after treatment for head and neck cancer: An economic evaluation." European Journal of Cancer **93**: 89-98.

Meregaglia, M., et al. (2018). "The short-term economic burden of gestational diabetes mellitus in Italy." BMC Pregnancy & Childbirth **18**(1): 58.

Mistry, H., et al. (2017). "Diabetic retinopathy and the use of laser photocoagulation: is it cost-effective to treat early?" BMJ Open Ophthalmology **2**(1): e000021.

Paleri, V., et al. (2018). "Gastrostomy versus nasogastric tube feeding for chemoradiation patients with head and neck cancer: the TUBE pilot RCT." Health Technology Assessment (Winchester, England) **22**(16): 1-144.

Pinheiro, B., et al. (2020). "Cost-effectiveness of cladribine tablets versus fingolimod in patients with highly active relapsing multiple sclerosis in Portugal." JOURNAL OF MEDICAL ECONOMICS: 1.

Rodriguez-Zepeda, M. D. C., et al. (2018). "Cost-Effectiveness of rFVIIa versus pd-aPCC in the Management of Mild to Moderate Bleeds in Pediatric Patients with Hemophilia A with Inhibitors in Mexico." Value in Health Regional Issues **17**: 164-173.

Si, S., et al. (2018). "Cost-effectiveness evaluation of the 45-49 year old health check versus usual care in Australian general practice: A modelling study." PLoS ONE [Electronic Resource] **13**(11): e0207110.

Wikman-Jorgensen, P. E., et al. (2017). "Microscopic observation drug-susceptibility assay vs. Xpert MTB/RIF for the diagnosis of tuberculosis in a rural African setting: a cost-utility analysis." Tropical Medicine & International Health **22**(6): 734-743.

Zhang, L., et al. (2018). "Clopidogrel Versus Aspirin for the Treatment of Acute Coronary Syndrome After a 12-Month Dual Antiplatelet Therapy: A Cost-effectiveness Analysis From China Payer's Perspective." Clinical Therapeutics **40**(12): 2125-2137.

Appendix D: Online survey questions

Section 1: Introduction

- *What HTA decision making processes do you have experience of?*
- *What was your role in the decision-making process?*

Section 2: In the HTA decision making processes you have experience of, in establishing if a treatment is cost-effective and potentially should be reimbursed, please tell us where and how is expert opinion currently used

- *How is expert opinion used at the committee meeting?*
- *How is expert opinion used as part of evidence generation, i.e. prior to committee meeting as part of evidence submitted or of the critique of this evidence?*
- *Is expert opinion used in contexts other than at the meeting or as part of evidence generation? If so, please describe*.

Section 3: In your experience, how easy is it to incorporate these forms of expert opinion into decision making?

- *Qualitative opinions*
- *Quantitative opinions where experts provide a central estimate*
- *Quantitative opinions where experts provide a central estimate and some measure of uncertainty*

Section 4: The following are some perceived limitations associated with expert opinion. In your experience, is expert opinion taken at face value or is it somehow 'down weighted' due to the following factors:

- *Experts provide subjective opinions and evidence provided should be considered of low quality, i.e. not empirical or experimental evidence*
- *Experts are conflicted, e.g. individual paid to act as expert by the company*
- *Often the reporting of the methods and conduct of elicitation exercises is poor*
- *Are there other factors that lead to 'down weighting' of experts opinions in reaching a decision?*

Section 5: The following are types of evidence required to inform economic models. For each of these, what form of expert opinion is typically sought?

- *Treatment effects for the short term, for example when existing experimental evidence is limited*
- *Extrapolation of treatment effects to longer term (beyond the follow-up of existing experimental evidence) e.g. over a lifetime*
- *Natural history of disease over time*
- *Resource use or costs*
- *Quality of life*
- *Describing the eligible population or the relevance/generalisability of evidence*
- *In your experience, what other types of evidence have expert opinion been used to inform?*

Section 6: Consider the following features of a decision problem. In which circumstances do decision makers seek for expert opinion? And also, when should they seek for experts' opinion more often/consistently?

- *Where the population is significantly heterogeneous, e.g. multiple subgroups or multiple indications. To assess whether the model (and evidence underlying it) fully reflects the expected level of heterogeneity in clinical practice how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there is uncertainty about the population which may benefit from the technology. To assess whether the model (and evidence underlying it) fully reflects the eligible population expected in clinical practice how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there is uncertainty about the natural history of the condition. To assess whether the model (and evidence underlying it) fully reflects the natural history expected in clinical practice how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there is uncertainty about treatment/diagnostic pathways. To validate pathways modelled how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there is uncertainty about the technology or the mechanism of action of the treatment. To support assumptions regarding the effects of the technology/treatments how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there are multiple alternative positions within treatment/diagnostic pathways. To support recommendations regarding different lines of treatment how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *In your experience, what other features of the decision problem is expert opinion CURRENTLY used to inform?*
- *In your experience, what other features of the decision problem SHOULD expert opinion be used to inform?*

Section 7: Consider the following limitations of the evidence used to inform a decision problem. In which circumstances do decision makers seek for expert opinion? And also, when should they seek for experts' opinion more often/consistently?

- *Where clinical/diagnostic accuracy evidence is sparse, e.g. small studies how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where clinical/diagnostic accuracy evidence is compromised for decision making, e.g. from another country how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where there is potential for bias in the evidence, e.g. single arm trial or absence of gold standard in a diagnostic accuracy study how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*
- *Where clinical evidence does not relate to the outcome of interest or is insufficient, e.g. on surrogate outcomes such as response or progression free survival how frequently is expert opinion CURRENTLY used? And how frequently SHOULD expert opinion be used?*

- *Where the observed clinical evidence is insufficient to describe the longer-term effect of treatment or evolution of the disease how frequently is expert opinion CURRENTLY used? and how frequently SHOULD expert opinion be used?*
- *Where evidence on how a diagnostic test changes treatment/diagnostic pathway is insufficient/limited how frequently is expert opinion CURRENTLY used? and how frequently SHOULD expert opinion be used?*
- *Insufficient evidence to describe alternative positioning of treatments/diagnostics, e.g. sequencing of treatments, multiple testing how frequently is expert opinion CURRENTLY used? and how frequently SHOULD expert opinion be used?*
- *In your experience, what other limitations of the evidence base is expert opinion CURRENTLY used to inform?*
- *In your experience, what other limitations of the evidence base SHOULD expert opinion been used to inform?*

Section 8: Where expert opinion is gathered quantitatively as part of evidence generation (i.e. as part of the company's submission or as part of the assessment group's critique), are the methods, conduct and results of elicitation used described in any detail?

- *Is the sample of experts described?*
- *Are the questions asked of experts described?*
- *In eliciting quantitatively, is it clear whether uncertainty was elicited?*
- *In eliciting quantitatively and where uncertainty was elicited, is there a description of how uncertainty was elicited?*
- *Is the conduct of the elicitation described (e.g. whether there was a facilitator, whether experts were trained)*
- *Are the results of the elicitation described?*
- *When more than one expert is included, are the values elicited by each individual expert provided?*

Section 9: Have you encountered any examples of a more structured process of eliciting quantitative parameters, for example using methods of elicitation that were referenced and training experts?

Section 10: What form of expert opinion would you value the most in the decision making processes you are involved with?

Section 11: How could expert opinion be more useful in decision making processes?

- *If it is conducted on key uncertainties needed to reach a decision*
- *If it ensures an appropriately selected sample of experts, e.g. minimises conflicts, sufficient sized sample*
- *If it is well described in terms of methods and conduct*
- *If it is conducted at an appropriate time in the process, e.g. before a committee meeting*
- *If experts are made accountable for their opinions by documenting these*
- *If it uses existing published structured process, e.g. SHELF, IDEAs protocol*
- *Other*

Section 12: Please give any further details you wish to add on the use of expert opinion

Appendix E: Components of elicitation, from Bojke, et al

Table E-1 Summary of the elicitation elements, components, and choices described in SEE guidelines

Element	Component	Choices
Identifying elicitation variables		
What quantities to elicit	Type of parameter	1. Elicit observable quantities 2. Elicit required model parameters directly
	Type of quantity	1. Proportions 2. Frequencies 3. Probabilities 4. Odds ratios
	Selection criteria	1. Define selection criteria (probabilities, consequences, constraints, etc) 2. Minimal assessment of each possible uncertain parameter and sensitivity analysis to see which uncertain parameters have the biggest impact
	Principles for describing quantities	1. Ask clear and well-defined questions 2. Ask questions in a manner consistent with how experts express their knowledge 3. Uncertainty in the elicited variables should impact the model and/or decision 4. Use neutral wording
	Decomposition/disaggregation	1. Decompose variables of interest to aid experts in the elicitation task 2. Don't decompose variables for the experts
	Handling dependence	1. Express dependent variables in terms of independent variables 2. Use conditional probabilities 3. Use other dependence elicitation methods
Encoding judgements	General approach	1. Fixed interval method: • Roulette or chips and bins method • Ask for % that falls within a specific range 2. Variable interval method:

		• Quantiles (Quartiles, Tertiles, 5%, 95% & median, 17%, 83% & median, other) • Bisection • Plausible probabilities (Lowest plausible probability, Highest plausible probability, Best guess for the probability) • Plausible quantities (Upper and lower plausible bounds, best guess, degree of belief) • NUSAP (Numeral, Unit, Spread, Assessment, Pedigree) 3. Hybrid fixed/variable interval methods 4. Summary statistics, moments, measures of central tendency 5. Elicit evidence, not parameter values, and analyst/facilitator defines probability distribution that reflects the body of evidence 6. Other
	Use of visual aids	1. Use to aid elicitation task 2. Do not use
Identifying and selecting experts		
Number of experts	Number of experts	1. Depends on application 1. Options mentioned in different guidelines: about 10; about 5 specialists and 2-3 generalists; 10-20; 6-e12; at least 4; 8 a "rule of thumb"; 5-9
Selecting experts	Roles within SEE	1. Facilitator (assessor, analyst, coordinator): prepare and conduct elicitation 2. Expert (technical expert, specialist, subject-matter expert): provide judgments (and/or evidence) 3. Generalists: may provide judgments, advise on design, or help with the elicitation
	Desired characteristics for those providing judgements	1. Normative expertise 2. Substantive expertise 3. Willingness (interest and availability) to participate

		4. Ability to understand questions 5. Ability to apply skills 6. Notability
	Identification procedure	1. Recommendations by peers, either formally or informally 2. Research outputs 3. Known experience 4. RFP to seek out experts 5. Profile matrix to identify types of expertise required
	Selection procedure	1. Disclosure of personal and financial interests 2. Pursue diversity in opinions, specialisation, area, institution, etc. 3. Pursue diversity in age, gender, culture 4. Formal selection criteria developed and applied 5. Send potential experts a questionnaire 6. Review CVs of possible experts and have a committee select accordingly 7. Match possible experts against profile matrix
	Possible selection criteria	1. Reputation 2. Experience and qualifications 3. Publication history 4. Diversity in background 5. Conflicts of interest 6. Awards 7. Balancing different viewpoints and managing group dynamics 8. Peer assessment (such as GEM) 9. Convenience 10. Balance of internal and external experts (e.g., include at least 2 external experts)
Training and preparation		
Pilot the protocol	Pilot exercise	1. Pilot 2. No mention of pilot
Training and preparation for experts	What to cover in training	1. Probability, including subjective probability, and related concepts 2. Motivation for elicitation 3. Description of what is required from experts

		4. How results will be used 5. Elicitation questions 6. Example and practice questions 7. Review of potential biases 8. Relevant background information, data, and sources 9. Review assumptions and definitions used in the elicitation 10. Description of performance assessment (if relevant) 11. Introduction to dependence (if relevant)
Conducting the elicitation		
Mode of administration	Location	1. Face-to-face • 1-on-1 • Group • Plenary 1. Remote (web, mail, email, phone, video conference, etc)
Level of elicitation	Level of elicitation	2. Individual 3. Group 4. Combination (individual assessment followed by group discussion and assessment)
Feedback and revision	Type of feedback	1. Graphical feedback 2. Fitted distributions 3. Written description of the expert's rationale 4. Rationales from other experts 5. Data collected in the future 6. Discussion of elicited values 7. The expert's performance scores 8. Result of using elicited values in the model 9. Decision resulting from the expert judgment 10. Draft elicitation report
	What to feed back	1. The individual's judgments 2. Aggregated group judgments 3. Judgments from other individual experts
	Opportunity for revision	1. Iterate elicitation/feedback rounds 2. Update after future data is collected

		3. Update for revisions/clarifications after circulating draft elicitation report
Interaction	Opportunity for interaction	1. No interaction 2. Group discussion prior to individual elicitation 3. Group discussion and group elicitation 4. Group discussion following individual elicitation (with opportunity for revision) 5. Remote, anonymized interaction
Rationales	Rationales	1. Collect/record rationales from experts (about how they made their judgments) 2. Collect/record rationales from decision makers (about how they used the expert judgments)
Aggregation	Aggregation	1. Aggregate 2. Don't aggregate • Analyst provides a distribution that captures knowledge from all experts (the Kaplan approach) • Only use individual distributions
	Aggregation approach	1. Mathematical • Opinion pool: equal-weighting, performance-based weighting (with seed questions), analyst-defined weighting (based on rationales, expert qualifications, or other criteria) • Bayesian aggregation 2. Behavioural 3. Combination 4. Other
Fit to distribution	Fit	1. Fit to parametric distribution 2. Use non-parametric approaches 3. Don't fit at all
	Distribution	1. Uniform 2. Triangular 3. Uniform over elicited intervals 4. Normal/beta/other parametric distribution
	Fitting method	1. Minimum least squares 2. Method of moments

		3. Other
Post-elicitation		
Feedback on process	Feedback from experts on process	1. Get feedback on the procedure if future data collection contradicts elicitation results 2. Ask experts to appraise the elicitation exercise after completing it
Adjusting judgements	Methods for adjusting judgments	1. Do not adjust experts' assessments 2. Possible adjustments • Calibrate experts' assessments • Adjust to improve coherence (described by Lindley et al. (1979)) • Small adjustments allowed, if they are fed back to the experts • Drop an expert from the panel
Documentation	What to include	1. Elicitation questions 2. Responses from individual experts (if elicited) 3. Description of process and assumptions for fitting a distribution 4. Discussion of elicitation procedure (and justification for choices made) 5. Rationales 6. Evidence related to elicited quantities 7. Aggregated judgements and/or consensus curves 8. Discussion of use/impact of elicitation results 9. Recording of session(s) 10. List of experts 11. Definitions and assumptions 12. The process for updating judgments
Managing heuristics and biases		
Managing heuristics and biases	Biases relevant for SEE	1. Cognitive biases • Overconfidence • Representativeness • Availability • Anchoring and adjustment • Conservatism • "law of small numbers" • Hindsight bias

<table>
<tr><td></td><td></td><td>
Discrepancy between expert’s beliefs and responses (conscious or unconscious)
Location errors
Tacit assumptions
Inconsistency
2. Motivational biases
Management bias
Expert bias
Social pressure
Group think
Impression management
Wishful thinking
Misinterpretation
Misrepresentation
</td></tr>
<tr><td></td><td>Bias elimination or reduction strategies</td><td>
1. Give experts practice and feedback
2. Identify biases through discussion with experts
3. Provide training on biases
4. Frame questions to minimize biases and ambiguity
5. Provide relevant background evidence
6. Ask for upper/lower bounds first
7. Ask experts to specify the credible interval they have provided
8. Minimize and record conflicts of interest among the experts
9. Require the experts address conflicting information
10. Collect rationales from experts
11. Report anonymous results
12. Anticipate likely biases
13. Ask experts about evidence, not the probability
14. Avoid numbers in questions
</td></tr>
<tr><td colspan="3">Considering the validity of the process and results</td></tr>
<tr><td>Validation</td><td>Characteristics of validity and supporting actions</td><td>
1. Faithfully capturing experts’ beliefs
Provide feedback (graphical feedback often mentioned)
Calibration could be a pragmatic proxy
Test that the question is understood
</td></tr>
</table>

		2. Fitness for purpose 3. Calibration • Ask questions with realizations (i.e., seed questions) that allow calibration to be tested 4. Calibration and informativeness scoring on seed questions (i.e., the Classical Model) • Score experts according to calibration and informativeness • Use scores as basis for performance-based weights (related to Aggregation choices) • Score both individual experts and combinations of experts 5. Coherence • Ask for sets of probabilities that allow coherence to be tested • Overfitting (asking for one more summary than is needed) • Ask for rationales from experts 6. Consistency • Ask for rationales from experts (and check for inconsistencies) • Provide feedback • Derive/give feedback on density function during elicitation • Multiply/integrate decompositions during elicitation • Use different elicitation methods and compare results 7. Internal peer review of process and/or results 8. External peer review of process and/or results

Table E-2 Level of agreement on recommendations and choices in SEE guidelines

Element	Component	Agreement level	Explanation
Identifying elicitation variables			
What quantities to elicit	Type of parameter	Some disagreement	Guidelines agree that observable quantities are preferred, but disagree on whether directly eliciting model parameters is an acceptable choice.
	Type of quantity	Disagreement	Guidelines offer conflicting recommendations on whether eliciting probabilities (compared with other uncertain quantities) is an acceptable choice.
	Selection criteria	Some agreement	Fewer than five guidelines discuss this, but they agree selection criteria should be defined.
	Principles for describing quantities	Some agreement	Some guidelines describe slightly different principles (e.g., asking clear questions, ensuring uncertainty on elicited parameters impacts the final decision or model), but they do not conflict.
	Decomposition	Agreement	The guidelines that discuss decomposing the variables of interest all agree it should be a choice.
	Handling dependence	Some agreement	The guidelines that discuss dependence agree it should be avoided if possible or addressed separately, but they discuss a range of methods for considering dependence.
Encoding judgements	General approach	Disagreement	Guidelines recommend and discuss different, conflicting methods for encoding judgements.
	Use of visual aids	Some agreement	Fewer than five guidelines discuss this, but they agree visual aids can be a useful choice.

Identifying and selecting experts			
Number of experts	Number of experts	Agreement	The experts agree that multiple experts are important, with most guidelines recommending around 5-10 experts.
Selecting experts	Roles within SEE	Agreement	The guidelines are very consistent in their description of the roles involved with elicitation.
	Desired characteristics for those provide judgements	Some agreement	Characteristics discussed in the guidelines are largely consistent, aside from differing views on if normative expertise is a requirement or just desired.
	Identification procedure	Some agreement	Recommendations differ but do not conflict across the guidelines. Agency guidelines tend to offer more detail.
	Selection procedure	Some agreement	Recommendations differ but do not conflict across the guidelines. Agency guidelines tend to offer more detail.
	Possible selection criteria	Some agreement	Recommendations differ but do not conflict across the guidelines.
Training and preparation			
Pilot the protocol	Pilot exercise	Agreement	Almost all guidelines recommend conducting a pilot exercise.
Training and preparation for experts	What to cover in training	Some agreement	The lists of what should be included in training vary across guidelines but do not conflict.
Conducting the elicitation			
Mode of administration	Location	Some agreement	Most guidelines agree that face-to-face administration is preferred, though remote options may be pragmatically useful alternative in some situations.
Level of elicitation	Level of elicitation	Disagreement	Guidelines recommend and discuss conflicting levels of elicitation.

Feedback and revision	Type of feedback	Some agreement	Recommendations differ but do not conflict across the guidelines.
	What to feed back	Some agreement	Recommendations differ but do not conflict across the guidelines.
	Opportunity for revision	Some agreement	Guidelines either recommend revision take place following an elicitation (as part of an iterative process or immediately following the elicitation) or further in the future, following a draft report or additional data collection.
Interaction	Opportunity for interaction	Disagreement	Guidelines offer conflicting recommendations about when and how to facilitation interaction between the experts.
Rationales	Rationales	Agreement	Almost all guidelines recommend collecting expert rationales in some form.
Post-elicitation			
Aggregation	Aggregation	Agreement	All guidelines discuss aggregation as a recommendation or valid choice.
	Aggregation approach	Disagreement	Guidelines offer conflicting recommendations on the approach and method to aggregate judgements.
Fit to distribution	Fit	Some disagreement	The guidelines make few recommendations, but their choices differ.
	Distribution	Some agreement	Fewer than five guidelines discuss this, but they generally agree that many parametric distributions could be chosen.
	Fitting method	Some agreement	Fewer than five guidelines discuss this, but they generally agree that choices include minimum least squares and method of moments.
Feedback on process	Feedback from experts on process	Some agreement	Fewer than five guidelines discuss this, and they recommend complementary approaches.

Adjusting judgements	Methods for adjusting judgements	Some disagreement	Fewer than five guidelines discuss this, but they offer different perspectives.
Documentation	What to include	Some agreement	The lists of what should be included in final documentation vary across guidelines but do not conflict.
Managing heuristics and biases			
Managing heuristics and biases	Biases relevant for SEE	Some agreement	The lists of potential biases vary across guidelines but do not conflict.
	Bias elimination or reduction strategies	Some agreement	The list of possible strategies vary across guidelines but do not conflict.
Considering the validity of the process and results			
Validation	Characteristics/measures	Disagreement	The guidelines differ in their definitions of validity and discussion of how the concept can be operationalised in an elicitation.