Experiences of structured elicitation for model based cost-effectiveness analyses

Running title: Expert elicitation for cost-effectiveness

Marta O Soares (corresponding author), PhD
Centre for Health Economics,
Alcuin 'A' Block, University of York,
Heslington, York
YO10 5DD
UK
marta.soares@york.ac.uk

Linda Sharples, PhD
Medical Statistics Department, London School of Hygiene and Tropical Medicine, London, UK

Alec Morton, PhD
Management Science, University of Strathclyde, Glasgow, UK

Karl Claxton PhD
Centre for Health Economics and Department of Economics, University of York, York, UK

Laura Bojke, PhD
Centre for Health Economics, University of York, York, UK

Funding statement: Work contributing to this manuscript was conducted as part of a Medical Research Council (MRC) grant “HEE: Developing a reference protocol for expert elicitation in health care decision making” (MR/N028511/1). Laura Bojke was also supported in the preparation/submission of this paper by the Health Economics and Outcome Measurement (HEOM) Theme of the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC

YH; www.clahrc-yh.nir.ac.uk, IS-CLA-0113-10020). The views and opinions expressed are those of the authors, and not necessarily those of the MRC, the UK National Health Service (NHS), the NIHR or the Department of Health.

Précis: A review of applications of structured expert elicitation in cost-effectiveness analyses, with the aim of identifying considerations and challenges in the design, conduct and analyses.

Word count: 4000

Number of pages: 32

Number of figures: 0

Number of tables: 3

Appendix:

Pages: 2

Figures: 0

Tables: 0

Supplementary material:

Pages: 3

Figures: 0

Tables: 1

Abstract

Empirical evidence supporting the cost-effectiveness estimates of particular health care technologies may be limited, or it may even be missing entirely. In these situations, additional information, often in the form of expert judgements, is needed to reach a decision. Formal methods to quantify experts' beliefs, termed structured expert elicitation (SEE), but only limited research is available in support of methodological choices. Perhaps as a consequence, the use of SEE in the context of cost-effectiveness modelling is limited. This paper reviews applications of SEE in cost-effectiveness modelling with the aim of summarising the basis for methodological choices made in each application and record the difficulties and challenges reported by the authors in the design, conduct and analyses. This review of experiences of SEE aimed to highlight a number of specificities/constraints that can shape the development of guidance and target future research efforts in this area. The review demonstrates considerable heterogeneity in methods used and authors acknowledge great methodological uncertainty in justifying their choices. Specificities of the context area emerging as potentially important in determining further methodological research in elicitation are: between-expert variation and its interpretation, the fact that substantive experts in the area may not be trained in quantitative subjects, that judgements are often needed on a variety of parameter types, the need for some form of assessment of validity, and the need for more integration with behavioural research to devise relevant debiasing strategies.

1. Context

Reimbursement decisions are often supported by model based economic evaluation (MBEE)[1]. Uncertainty in the evidence used to populate these models can result in uncertain cost-effectiveness estimates.[2] There may be circumstances in which empirical data is limited (for example, a cancer product licensed on the basis of progression-free survival, with limited evidence on survival impacts), or is missing entirely (for example, when assessing the value of a future clinical trial for a medical technology). In these situations, additional information, often in the form of expert judgements, reported as a distribution, is needed to reach a decision. To improve the accountability of the decision making process, the procedure used to derive these judgements should be transparent, with any uncertainty in individual judgements characterised, in addition to between-expert variation[3].

Formal methods to quantify experts' beliefs exist, and are termed structured expert elicitation (SEE)[3,4]. Elicitation has been used in a variety of disciplines including weather forecasting[5] and food and safety risk assessments[6]. However, the existing methodological research on elicitation, both generic and discipline-specific, is inconsistent and non-committal[7]. Methodological uncertainties may be one of the main reasons for the limited use of formal SEE in the context of MBEE. A review of applications in this area, published in 2013[8], identified only a small number(14) of studies reporting the use of SEE. This review did not seek to determine the reasons for heterogeneity of approach, nor did it look at the challenges faced when conducting SEE to support MBEE and inform directions for future research.

In pursuit of further clarity, this paper updates the abovementioned review[8] but instead of reporting the way elicitation is being used in practice, it focusses on summarising the basis for methodological choices made in each application (design, conduct, and analysis) and the difficulties and challenges reported by the authors. In Section 2, the methods for identifying the literature are described and an overview of the contexts in which SEE was used across studies is made. The sections that follow discuss choices, challenges and issues relating to the design (Section 3.2), conduct (Section 3.3) and analyses (Section 3.4) of SEE. In detailing these elements it is necessary to first describe the applications (Section 3.1 and Tables 1-3), and that is where the similarities exist between this review and the 2013[8] review, and also where they end. Finally section 4 sets out specific challenges posed by SEE in MBEE to inform the direction of future research.

2. Methods

To identify applications of SEE, the 2013 review[8] was updated (identifying studies up to 11th April 2017). Further details on the methods of the search are given in the Appendix but, in brief, studies were identified via Ovid SP MEDLINE and, similarly to the 2013 review,[8] were only included if they contained a SEE to elicit uncertain parameters (in the form of a distribution) to inform MBEE. Studies conducting preference elicitation, for example to generate utility estimates for health states, were not included.

The methods used in each application were extracted (the extraction form is reproduced in Tables 1 to 3, that also present results) along with the criteria used to support methodological and practical choices and any issues or challenges discussed in the text. Issues and challenges were extracted using an open field, and then categorised and grouped for reporting.

3. Results

3.1. Summary of applied studies

In total, 21 studies were included. Table 1 and the Appendix provide summary information on each study and highlight that elicitation has been used mainly when data on a particular parameter is limited or absent. Four of the 21 applications were applied in an early modelling context, where there may not be direct clinical experience with the technology of interest, and eight evaluated a diagnostic or screening strategy.

Table 2 summarises the method of recruiting experts, methods of elicitation and methods of aggregation in each of the applied studies. Table 3 reports how the SEE was conducted, including mode of administration and use of any software, and also any analyses that were performed. Each element of the applied studies is considered and choices, challenges and issues discussed in the following sections.

3.2. Aspects related to the design of the SEE

Considerations on the design of SEE were grouped according to: specification of the quantities to elicit , selection of experts , elicitation method , and type of aggregation and weighting of experts' judgements .

Specification of quantities to elicit

In all applications, experts' beliefs were sought for only a few parameters of a decision model, often not elicited directly but calculated from one or more alternative elicited quantities. For example, a time-constant transition probability could be indirectly elicited by asking experts for the mean time at which an event is observed or, alternatively, the proportion of patients that have an event within a particular time period. In the applications, the choice of which quantities to elicit was based on a number of criteria. The first was appropriateness for experts. Parameters in decision models can be complex and may not be directly observable by experts; to account for this some of the studies expressed, for example, relative effectiveness parameters as probabilities[9-12], or sensitivities and specificities into probabilities of the true disease status of the patients conditional on the test results[12]. It may also be more appropriate for different experts to elicit different quantities (for example, in one application[10], geneticists elicited accuracy of a genetic test, and cardiologists elicited parameters related to disease progression) or, in the presence of heterogeneity, for a particular quantity to be elicited separately for population subgroups[12,13].

The second criteria related to statistical concerns. The quantities elicited should be fit-for-purpose, not only in informing decision models (for example, reflecting time dependency) but also in allowing elicited evidence to be combined with any existing empirical evidence[9]. Statistical coherence between quantities elicited should also be ensured.[9,14] For example, where a number of mutually exclusive outcomes is of relevance, eliciting their probabilities independently (with uncertainty) cannot guarantee that they sum to one, but re-expressing parameters as conditional binomial variables does ensure this[9,14]. Additionally, dependencies may exist between the quantities elicited (for example, correlation between relative effectiveness parameters for alternative interventions), between quantities elicited and known covariates or between *a priori* independent quantities that are elicited from the same expert (for example. some experts may be prone to eliciting higher values across the board than others). Of the 7 studies that raised the issue of dependency[9,11,14-18], three did not deal with it at all[15-17]; two re-expressed target parameters as conditionally independent[9,11] and the remaining two studies explicitly elicited dependency[14,18]. In the latter, rather than a correlation parameter being elicited directly, relationships between parameters were captured by asking experts to express how their judgements would change if values for other quantities were known. Methods to elicit dependence directly were, however, generally thought to be complex[9].

The final criterion was burden to experts. Burden can be reduced by: limiting the number of target parameters to elicit; eliciting homogeneous quantities throughout the exercise (e.g. all probability parameters)[9]; using filter questions (e.g. ‘do you think X differs from Y?’) [9,19]; not eliciting dependency or only eliciting it for the covariates identified by the experts as relevant[14].

Selection of experts

All applications recruited health care professionals (but not exclusively[20,21]) based on the following criteria: recognition by peers[10], specialist knowledge or clinical experience[9,10,13,18,19,22,23], based in the relevant jurisdiction[9,10,18,19], research experience[10,22,23] and lack of involvement in product development[13]. In early technology assessment, applications have also looked for other factors such as interaction with colleagues, seen as indicative of the adaptive skills required in this context.

A number of authors[9,14,24] recognised that health care professionals are unlikely to have knowledge of elicitation and may have only sparse quantitative skills. This has been judged by the authors to compromise normative skills; defined as the ability to accurately express judgements in a particular quantitative format, such as probabilities. This has driven the choices made in designing and conducting the SEE, such as training needs, method of elicitation and definition of the quantities to elicit.[9]

Many of the applications have included a varied sample of experts by recruiting them from a range of relevant specialties[10,12,20], clinical settings[9,10,20] and geographical areas/countries[10,23] to capture heterogeneity in beliefs (reflecting underlying heterogeneity in patient populations), and avoid dependency between experts[10].

Across the applications, sampling was purposeful: typically, experts recruited were either collaborators in the research, or were identified by recommendation from clinical colleagues[10,14,18], by contacting professional associations[24], or at specialist conferences[15,23]. Sample sizes ranged from 2[11] to 23[9] (Table 2), generally targeting a small but ‘varied’ sample[10]. One author[22], however, argued that restricting the pool of experts may amplify biases arising, for example, from shared exposure to unrepresentative clinical experience. Many applications mention constraints to sampling due to: resources available to fund the SEE[22], limited number of relevant experts[10], or geographic distance[18,22].

Elicitation method

An important requirement for MBEE 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 methods (FIM)[9-12,18,19,24-27] or variable interval methods (VIM)[13,14,17,21-23,28] (Table 2). In FIMs, experts are provided with ranges of values and asked to assess the probability that the quantity lies in each. In VIMs, experts are asked to specify values of the quantity of interest for pre-defined percentiles of the distribution. Whilst one application[10] chose FIM because the literature suggested that it returns higher variance, it was more common for authors to consider both approaches. Choices were justified on the basis of: pilot exercises designed for the purpose (see below), generic methods research, previous use in MBEE, and claims of lower burden or intuitiveness for experts.

Applications using the VIM elicit either quartiles of the distribution[14,17,21,22] or credible intervals[13,23,28], and in general ask for a very limited number of summaries. Studies using FIM often chose the 'chip and bins' method (histogram technique or probability grid)[9,18,19,25-27]. This method defines a larger number of intervals (typically up to 20) and asks the expert to distribute a fixed number of chips across these intervals. The more chips placed in a particular interval, the stronger the belief that the true value of the quantity of interest lies in that interval. Despite many of the studies arguing for the intuitiveness of the 'chips and bins' method, a pilot study[13] found that two of the three experts included preferred eliciting 95%

probability intervals. Other FIMs, that divide the plausible range of values into 4 or 6 complimentary or overlapping intervals, and ask for quantitative expression of strength of belief for each, have also been used[10]. Pilot testing amongst these found that six complementary intervals resulted in very narrow ranges, and that overlapping intervals were confusing to experts[10]. A separate study[24] comparing the 'chip and bins' method with the 4 complementary intervals method found that the latter required more careful consideration and, because of that, experts perceived it to be more (face-)valid. Also, the resulting pooled distributions were wider. Another FIM application[11] asked experts for a central estimate and elicited for a single interval. This study was noteworthy as it took a more frequentist approach by presenting a hypothetical scenario where 100 different experiments were conducted, and asked experts how many times in those experiments would they expect the observed value to be larger than a particular value.

Consensus vs. mathematical aggregation, weighting of experts

Fourteen studies elicit individually from experts and aggregate mathematically, three aimed to achieve consensus amongst experts[15,17,21] and three others did not explicitly report the method of aggregation used[16,20,25] (Table 3).

None of the three studies using consensus was explicit about the reasons for choosing consensus or the process of achieving it. Therefore, the following focuses on those using a mathematical approach.

Authors justify the choice of mathematical aggregation based on the desirability to reflect variation within and between experts[12], because consensus is known to lead to overconfident results(i.e., narrow distributions)[10] and because it raises practical difficulties of convening experts and providing experienced facilitation. One pilot study[9]additionaly showed consensus produced incoherent probability statements (the median time to healing was greater than the time taken for 70% of patients to heal).

When adopting a mathematical approach there needs to be some consideration on whether to differentially weight the responses of individual experts. Most of the applications reviewed claim insufficient justification for generating differential weights[9,10] and lack of clarity on how to appropriately generate the weights[9,13,26] and hence apply equal weighting. Five studies, however, explored unequal weighting, either based on responses to seed questions[9,18,23,24](performance-based weighting) or using the clinical background of experts[13](objective weighting). Performance-based weighting (commonly called calibration) typically asks experts to respond to one or more 'seed' questions known to the analyst (with certainty) but not the expert. Elicited responses are compared to their known value to generate the weights, with the most commonly used method, the classical method[29], considering both accuracy and informativeness.

Applied studies question the usefulness of calibration and request further methodological research. Uncertainties relate to the number and definition of appropriate seed questions for particular target questions. For example, one study piloted four alternative seed questions[9] and found that, when used separately, these generated divergent weights. In another application, responses to eight seed questions generated zero weights to 17 of the 19 substantive experts – authors expressed discomfort in discarding so much of the information. A third study[18] questioned the relevance of seed questions known with certainty and instead used seeds that were known with uncertainty. Weights were generated based on the overlap of the elicited and true distributions.

3.3. Experiences with the conduct of the exercise

No studies reported major challenges in the conduct of the SEE, despite the complexity of the task.

Consensus exercises were typically face-to-face, in a group; mathematical exercises adopted a mix of formats, ranging from individual interviews to remote completion via email(Table 4). Convening a group facilitated training and a common understanding of the problem[12]; but some studies[10,12,14,26] departed from this format due to time constraints, geographical limitations and availability of experts. One example[9], using a mathematical approach, elicited 18 uncertain parameters in 2 hours following a 2 hour training session.

Administration was via bespoke tools using Excel [9,10,18,19,24,26], paper questionnaires, a generic elicitation package(the Sheffield Elicitation Framework or SHELF)[20,21] and a software package for the elicitation of dependency (Prior Elicitation Graphical Software, PEGS)[14](Table 4). Other studies did not specify mode of administration. The most common tool, bespoke Excel applications, had several perceived advantages, including tailoring the presentation in order to avoid inconsistencies and conditioning questions on previous responses.[9]

Some exercises were explicit about piloting the tool to ensure clear wording of the questions[9,13,19,22], and most offered opportunities for revision and/or graphical feedback (Table 4).

Five applications were explicit about training of experts[9,12,13,18,24] covering: overview of the project and of the role of elicitation[9,12,13,18,24]; quantities required and definitions[9,12,18,24]; explanation and expression of uncertainty[9,13]; consideration of potential biases[9,18,24]; use of the elicitation instrument[9] and delivery of practice exercises[9,12,18,24]. Studies that implemented elicitation remotely generally included some form of instructions, although none reported these in detail.[10]

3.4. Experiences with the analyses and interpretation of elicited evidence

Considerations on the experiences of analyses and interpretation of elicited evidence were grouped according to: validity assessment, syntheses of multiple beliefs for mathematical aggregation, deriving smooth prior distributions and further use of elicited evidence in decision modelling.

Considerations on validity

Aspects related to validity in applied studies were missingness, validity checks and self-reported face-validity. Reporting of missingness was poor: no applications provided recruitment rates, only a few provided the number of recruited experts who did not turn up or did not return the elicitation form[10-12,14,24], and none was explicit about missing responses to individual questions. No studies dealt with missing responses, either formally or informally.

Three types of validity checks have been implemented. One study[22] contrasted qualitative and quantitative responses(internal validity) and found a small number of inconsistent responses; for example, 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 of multiple experts[9,11,14,23,25,26]. Whilst some authors[12,14] valued good agreement others[9,10] accepted variation between individuals (on the basis that individual beliefs are being requested). Finally, when external evidence was available, this was compared with elicited beliefs (external validity).[9,12,14,26] Authors sought agreement, but when differences arose they cautiously justify them based on population differences.[12]

Some applications requested feedback from experts on: the ease of completion of the SEE[9,10,24,26], the basis for experts’ answers (to reveal the sources of evidence considered by the experts and their level of knowledge[10]), or on self-reported face validity[9,10,12,24].

Syntheses of multiple beliefs elicited in mathematical aggregation

Of the 14 studies that used a mathematical approach to aggregation, one did not generate a group estimate and instead used the responses of each expert individually[14]. Nine linearly pooled, by averaging individual distributions (with or without weighting, see section 3.4). Authors justify this choice based on the lack of published evidence that more complex methods outperform linear pooling.[10] Two other studies use the predictive distribution from a random-effects meta-analyses of individual elicited distributions[18,28], a method arising from statistical methodology rather than the wider elicitation literature. Given the random effects model results in a combined distribution that can be more precise that any of the individual distributions, this pooling method has been deemed inappropriate for use in MBEE.[18]

Generally, inputs to decision models were pooled across experts, except in one study that ran the model with each of the individual elicited distributions and linearly pooled resulting outputs.[11]

Deriving smooth prior distribution functions

Some applications were not explicit about how prior distributions were derived from elicited summaries. Those that were explicit used parametric distributions (Table 3), with the choice of distribution either not justified or based on general MBEE literature on distribution choice for probabilistic sensitivity analyses[10]. To fit the distribution (i.e. evaluate the parameters of the distribution that best fit the empirical distributions elicited from experts), some applications cited software[11,13] and others cite the fitting method (e.g. maximum likelihood fitting[10], least squares[17], method of moment[9,18]). Goodness of fit was evaluated either in discussion with the experts[17], or graphically by superimposing the fitted probability density function on the histogram[10]. Bojke et al[18] acknowledged that, whilst the fit of pre-specified parametric distributions was not always ideal in their example, methods that allow fitting of non-parametric distributions are more complex and can complicate further analyses, particularly where Bayesian updating is further required, for example, in value of information calculations.

Further use of elicited evidence in decision modelling

Elicited evidence has been seen as a way of characterising uncertainty for model parameters or assumptions, to inform the decision to acquire further evidence.[9] In some applications elicited evidence was used directly as input to the a cost-effectiveness model[10,12,14,15,20,24]. Where external evidence existed on elicited parameters, some authors present both sources separately using scenarios[12,17], while others combine them using Bayesian updating[9,16,21]. The latter is only consistent under the assumption that the experts did not consider existing evidence when formulating their judgements.[9] Three authors[11,14,19] explored use of individual experts' beliefs and found that results and associated allocation decisions varied between experts.

3.5. Considerations on bias

When seeking to gather experts' opinions, it is important to consider their potential biases, specifically motivational or cognitive.[30] Motivational biases relate to conscious or subconscious distortions of judgments because of self-interest. Cognitive biases are associated with the use of heuristics: cognitive shortcuts that individuals use when asked for complex judgements. When such mental processes are faulty, these may lead to biased judgments.

The potential for bias in expert opinion was recognised in some SEE[9,22] with reported attempts to minimize bias in the design[26]. Two applications make explicit efforts to avoid recruiting experts that may have motivational biases[13,20] Two studies provided information on cognitive biases in the training session[9,24].

4. Conclusions

In the published applied studies, authors generally recognised great potential for using elicitation in MBEE, particularly where evidence is absent (including the early modelling context[13]).

Our critical review demonstrates that reporting is poor (as also identified elsewhere[31]), and there is a lack of consensus on methodology. Given the direct link to healthcare policy decisions, it is important that methodological guidance specific to HTA is generated, with consideration of the constraints inherent to the processes of policy decisions (such as timelines, budget and availability of experts). A number of principles from the elicitation literature are expected to generalise to the MBEE setting, such as the need for piloting and training; however, for many other areas of SEE, it is not clear that methods used in other disciplines translate to HTA. Our review highlights a number of specificities/constraints that can shape the development of guidance and target future research efforts in this area, summarised as follows.

Firstly, there exists important between-expert variation. In other disciplines, variation is generally linked to different levels of bias and hence regarded as undesirable, warranting the use of strategies to reduce or discourage variation, such as consensus methods. The majority of applications in MBEE, however, expect wide variation in the beliefs of multiple experts due to genuine heterogeneity in the populations experts draw upon. Further research efforts should examine the origins of variation, and consider how to appropriately reflect it.

Secondly, substantive experts in HTA are health professionals who may not be trained in quantitative subjects, unlike other areas of science in which elicitation is used such as engineering or meteorology. Further research on SEE should consider the appropriateness of alternative methods of elicitation (e.g. chips and bins, or bisection method) for the potentially less normative experts, or on how to facilitate the elicitation of complex parameters, including dependency. Furthermore, elicitation may have an important role in early modelling where experts' beliefs are required on new technologies. In this case, adaptive skills are required to allow experts' substantive expertise (in the disease area and/or other health care strategies) to be appropriately used. Further research should focus on how to promote the use of adaptive skills, or how to determine better performing individuals in this context.

Cost-effectiveness modelling typically requires judgements on a relatively large number of parameter types (e.g. probabilities, relative treatment effects, costs and Health-Related Quality of Life scores) and the design and conduct of a SEE may well be influenced by what quantities are required. The applications reviewed here elicit a range of different quantities to inform the same parameter; however, they do not draw on evidence or past experiences specific to that quantity. Design of future applications could be aided by a compilation of possible quantities that can be reasonably used to elicit particular parameters types, accompanied by guidance on how to ensure that the multiple quantities elicited in a particular application can be appropriately used within a decision model.

Perhaps given the direct link to decision making, most applied examples seek for assurance on the validity of the particular exercise. It is, however, not clear how such an assessment should proceed. Examples have used self-reported face-validity assessments, sensitivity analyses, and performance weighting (calibration). Particularly for performance weighting, despite a growing (generic) literature discussing the validity of this approach (see for example[32-34]), the applied literature struggles with supporting the methodological choices that need to be made. Whilst some means of correcting for poor performance is welcomed, in the applied literature concerns have been expressed that this should not repress expressions of heterogeneity. If SEE is to be used more systematically in MBEE further guidance is needed on how to demonstrate validity.

Finally, while 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 in the applied literature of the findings from behavioural research. A recent review placing special emphasis on debiasing techniques[30] is a helpful resource to be reflected in future research.

It is worth noting that our review only includes published examples, despite SEE being conducted more widely for MBEE. Moreover, the review is based on analytical reading of the published articles, and hence subject to a certain amount of interpretation. Further research in understanding the landscape of SEE for MBEE could include structured discussions amongst individuals with experience in the area to explore challenges in past exercises and identify those foreseen in future applications.

Table 1: Summary of applications.

	Study	**Type of strategy under investigation**	**Was the aim to inform an early assessment (i.e. R&D) rather than reimbursement?**	**Type of parameter(s) elicited**
1	Garthwaite 2008[14]	Treatment	No	Event probabilities (P), time to event, dependency
2	Leal 2007[10]	Diagnostic/screening	No	P, relative effectiveness (RE), diagnostic accuracy (DA)
3	Girling 2007[15]	Treatment	Yes	P, time to event
4	Stevenson 2009[16]	Prevents transmission	No	P, time to event, RE
5	Meads 2013[12]	Diagnostic/screening	Yes	P, DA, minimum important clinical difference
6	McKenna 2009 [19]	Treatment	No	P
7	Haakma 2014[13]	Diagnostic/screening	Yes	DA
8	Stevenson 2009b[17]	Treatment	No	P, RE
9	Speight 2006[25]	Diagnostic/screening	No	P

10 Sperber 2006[22]	Treatment	No	P, RE
11 Brodtkorb 2010	Several exercises conducted but insufficient detail is reported on each		
12 Colborn 2007[28]	Diagnostic/screening	No	P, RE
13 Soares 2011[9]	Treatment	No	P, RE
14 Bojke 2010 [18]	Treatment	No	RE, dependency
15 Cao 2013[11]	Diagnostic/screening	Yes	RE
16 Fischer 2013[23]	Treatment	No	counts, time to event
17 Poncet 2015[27]	Diagnostic/screening	No	P
18 Grigore 2016[24]	Treatment	No	P
19 Wilson, 2016[20]	Treatment	No	P, RE
20 Meeyai, 2015[21]	Vaccine	No	P
21 Grimm 2017[35]	Diagnostic/screening	No	Diffusion**

* non-dynamic decision model to establish cost-effectiveness of a particular intervention/strategy aimed to inform recommendations on its use for clinical practice

** rate of implementation in clinical practice over time

Table 2: Experts, method of elicitation and method of aggregation

		Experts and recruitment			Approach and method of elicitation		Aggregation		
#	Study	Type of experts	Recruitment	# experts	VIM or FIM+	Method (summaries elicited):	Aggregation approach	Weights used in main exercise?	Nature of weights
1	Garthwaite 2008[14]	NR	NR	4	VIM	Median and quartiles	Mathematical	No	--
2	Leal 2007[10]	Clinicians	purposive	6	FIM	Four complementary intervals	Mathematical	No	--
3	Girling 2007[15]	Clinicians	purposive	5	VIM	NR	Consensus	--	--
4	Stevenson 2009[16]	NR	NR	NR	NR	NR	NR	NR	NR
5	Meads 2013[12]	Clinicians	purposive	21	FIM	Chips and bins **	Mathematical	No	--

6	McKenna 2009[19]	NR	NR	5	FIM	Chips and bins	Mathematical	No	--
7	Haakma 2014[13]	Clinicians	purposive	14	VIM	Mode and 95% CI	Mathematical	Yes	Objective weights
8	Stevenson 2009b[17]	Clinicians	purposive	3	VIM	Median and quartiles	Consensus	--	--
9	Speight 2006[25]	Clinicians	NR	9	FIM	Chips and bins **	NR	NR	
10	Sperber 2006[22]	Clinicians	NR	NR	VIM	Median and quartiles	Mathematical	Yes, but no detail provided	Performance based
12	Colborn 2007[28]	NR	NR	4	VIM	Mean and 95% CI	Mathematical	NR	NR
13	Soares 2011[9]	Clinicians	NR	23	FIM	Chips and bins	Mathematical	No, but explored in a pilot	Performance based weights explored

14	Bojke 2010[18]	Clinicians	purposive	5	FIM	Chips and bins	Mathematical	Yes	performance based weights
15	Cao 2013[11]	Clinicians	NR	2	FIM	Mode and one percentile	Mathematical	No	--
16	Fischer 2013[23]	Clinicians	purposive	19	VIM	Median and 80%CI	Mathematical	No, but explored	Performance based weights explored
17	Poncet 2015[27]	Clinicians	NR	13	FIM	Chips and bins	Mathematical	No	--
18	Grigore 2016[24]	Clinicians	purposive	7	FIM	Chips and bins + four complementary intervals	Mathematical	Yes, alongside equal weighting	Performance based weights explored

19	Wilson, 2016[20]	Clinicians + policy strategist	NR	6	NR	NR	NR	NR	NR
20	Meeyai, 2015[21]	Clinicians + epidemiologists	NR	10	VIM	Mode and quartiles	Consensus	--	--
21	Grimm 2017[35]	NR	NR	3	NR	NR	Mathematical	NR	NR

NR: not reported

+ variable interval method, VIM, or fixed interval method, FIM

*Unclear, but description of results suggests variable interval method has been used

** includes studies that list the following methods: Chips and bins, Frequency chart, and Histogram method

Table 3: summary of applications. Conduct and analyses

	Conduct					**Analyses**		
Study	**Mode of administration**	**opportunities for revision**	**Format/software**	**training**	**piloting**	**Pooling**	**Fitting**	**Pooled distribution used directly within the decision model?**
Garthwaite 2008[14]	Individual face-to-face (IF2F) and remote (telephone) interviews (R)	Unclear	Interview and specialised software	NR	No	No pooling	Independently elicited quantities: NR; dependency elicitation: yes, generalised linear model	No, each experts' distributions used directly

Leal 2007[10]	R (email) + IF2F	Yes	Excel-based	NR	Yes	Linear pooling (LP)	Yes, maximum likelihood	Yes
Girling 2007 [15]	group face to face (GF2F)	Yes	NA	NR	NR	--	Yes, method NR	Yes
Stevenson 2009 [16]	NR	NR	NR	NR	NR	NR	NR	Yes
Meads 2013 [12]	GF2F + IF2F	NR	Paper	Yes	NR	LP	NR	No, Bayesian updating with existing evidence
McKenna 2009[19]	NR	NR	Excel-based	Yes	Yes	LP	Yes, method NR	Yes
Haakma 2014[13]	IF2F	Yes	Excel-based	NR	Yes	LP	Yes, Project Evaluation and Review Technique	Yes

							(PERT software)	
Stevenson 2009b[17]	GF2F	Yes	NR	NR	NR	NA	Yes, least-squares	NR
Speight 2006 [25]	NR	NR	Paper	NR	NR	NR	Yes, method NR	Yes
Sperber 2006 [22]	R	Yes	Excel-based	NR	Yes	LP	Yes, least squares (EasyFit software)	NR
Colborn 2007 [28]	NR	NR	NR	NR	NR	Predictive distribution from random effects meta-	Yes, method NR	Yes

						analysis (REMA)		
Soares 2011 [9]	GF2F	Yes	Excel-based	Yes	Yes	LP	Yes, method of moments	No, Bayesian updating with existing evidence
Bojke 2010 [18]	IF2F	Yes	Excel-based	Yes	NR	LP + REMA	Yes, method of moments.	Yes
Cao 2013 [11]	NR	NR	NR	NR	NR	LP	Yes (BetaBuster software)	Yes
Fischer 2013 [23]	GF2F, IF2F and R	NR	Paper	Yes	Yes	LP	No, empirical distribution used	NR
Poncet 2015. [27]	NR	NR	NR	NR	NR	LP	NR	Yes

Grigore 2016 [24]	IF2F	Yes	Excel-based	Yes	Yes	LP	Yes, method NR	Yes
Wilson, 2016 [20]	NR	NR	SHELF	NR	NR	NR	NR	Yes
Meeyai, 2015 [21]	NR	NR	SHELF	NR	NR	NR	NR	No, Bayesian updating with existing evidence
Grimm 2017 [35]	NR	NR	NR	NR	NR	LP	Yes, least squares	Yes

References

1. Drummond M. Methods for the economic evaluation of health care programmes. Fourth edition. ed. Oxford, United Kingdom ; New York, NY, USA: Oxford University Press; 2015.

2. Claxton K. Exploring uncertainty in cost-effectiveness analysis. Pharmacoeconomics. 2008;26(9):781-798.

3. Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, et al. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley 2006:338.

4. Cooke RM, Goossens LJH. Procedures guide for structured expert judgment. European Commission. 2000;European Atomic Energy Community.

5. Bruno Soares M, Dessai S. Exploring the use of seasonal climate forecasts in Europe through expert elicitation. Climate Risk Management. 2015;10:8-16.

6. European Food Safety A. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal. 2014;12(6):3734-n/a.

7. Ayyub B. Elicitation of Expert Opinions for Uncertainty and Risks. CRC Press. 2001.

8. Grigore B, Peters J, Hyde C, Stein K. Methods to Elicit Probability Distributions from Experts: A Systematic Review of Reported Practice in Health Technology Assessment. PharmacoEconomics. 2013;31:991–1003.

9. Soares MO, Bojke L, Dumville J, Iglesias C, Cullum N, Claxton K. Methods to elicit experts' beliefs over uncertain quantities: application to a cost effectiveness transition model

of negative pressure wound therapy for severe pressure ulceration. Stat Med. 2011;30(19):2363-2380.

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

11. Cao Q, Postmus D, Hillege HL, Buskens E. Probability elicitation to inform early health economic evaluations of new medical technologies: a case study in heart failure disease management. Value Health. 2013;16(4):529-535.

12. Meads C, Auguste P, Davenport 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 Technol Assess. 2013;17(12):1-323.

13. Haakma W, Steuten LMG, Bojke L, IJzerman MJ. Belief Elicitation to Populate Health Economic Models of Medical Diagnostic Devices in Development. Appl Health Econ Health Policy. 2014;12:327-334.

14. Garthwaite PH, Chilcott JB, Jenkinson DJ, Tappenden P. Use of expert knowledge in evaluating costs and benefits of alternative service provisions: a case study. Int J Technol Assess Health Care. 2008;24(3):350-357.

15. Girling AJ, Freeman G, Gordon JP, Poole-Wilson P, Scott DA, Ford RJ. Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy. International Journal of Technology Assessment in Health Care. 2007;23(2):269-277.

16. Stevenson MD, Oakley JE, Chick SE, Chalkidou K. The cost-effectiveness of surgical instrument management policies to reduce the risk of vCJD transmission to humans. J Oper Res Soc. 2009;60(4):506-518.

17. Stevenson MD, Oakley JE, Lloyd Jones M, 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):678-689.

18. Bojke L, Claxton K, Bravo-Vergel Y, Sculpher M, Palmer S, Abrams K. Eliciting distributions to populate decision analytic models. Value Health. 2010;13(5):557-564.

19. McKenna C, McDaid C, Suekarran S, 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):iii-iv, ix-xi, 1-90.

20. Wilson EC, Stanley G, Mirza Z. 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. Appl Health Econ Health Policy. 2016;14(2):205-215.

21. Meeyai A, Praditsitthikorn N, Kotirum S, et al. Seasonal influenza vaccination for children in Thailand: a cost-effectiveness analysis. PLoS Med. 2015;12(5):e1001829; discussion e1001829.

22. Sperber D, Mortimer D, Lorgelly P, Berlowitz D. An Expert on Every Street Corner? Methods for Eliciting Distributions in Geographically Dispersed Opinion Pools. Value in Health. 2013;16:434-437.

23. Fischer K, Lewandowski D, Janssen MP. Estimating unknown parameters in haemophilia using expert judgement elicitation. Haemophilia. 2013;19(5):e282-288.

24. Grigore B, Peters J, Hyde C, Stein K. A comparison of two methods for expert elicitation in health technology assessments. BMC Medical Research Methodology. 2016;16(85).

25. Speight PM, Palmer S, Moles DR, et al. The cost-effectiveness of screening for oral cancer in primary care. Health Technol Assess. 2006;10(14):1-144, iii-iv.

26. Brodtkorb T-H. Cost-effectiveness analysis of health technologies when evidence is scarce Center for Medical Technology Assessment, Department of Medical and Health Sciences, Linköping University, Sweden; 2010.

27. Poncet A, Gencer B, Blondon M, 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):e0127213.

28. Colbourn T, Asseburg C, Bojke L, 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):1-226, iii.

29. Cooke R. Experts in uncertainty : opinion and subjective probability in science. Oxford University Press; 1991.

30. Montibeller G, von Winterfeldt D. Cognitive and Motivational Biases in Decision and Risk Analysis. Risk Anal. 2015;35(7):1230-1251.

31. Iglesias CP, Thompson A, Rogowski WH, Payne K. Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations. Pharmacoeconomics. 2016;34(11):1161-1172.

32. Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. Reliab Eng Syst Safe. 2017;163:109-120.

33. Eggstaff JW, Mazzuchi TA, Sarkani S. The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Safe. 2014;121:72-82.

34. Clemen RT. Comment on Cooke's classical method. Reliab Eng Syst Safe. 2008;93(5):760-765.

35. Grimm SE, Dixon S, Stevens JW. Assessing the Expected Value of Research Studies in Reducing Uncertainty and Improving Implementation Dynamics. Med Decis Making. 2017:272989X16686766.