PharmacoEconomics (2013) 31:991–1003
DOI 10.1007/s40273-013-0092-z

SYSTEMATIC REVIEW

Methods to Elicit Probability Distributions from Experts: A Systematic Review of Reported Practice in Health Technology Assessment

Bogdan Grigore · Jaime Peters · Christopher Hyde · Ken Stein

Published online: 9 October 2013

Abstract

Background Elicitation is a technique that can be used to obtain probability distribution from experts about unknown quantities. We conducted a methodology review of reports where probability distributions had been elicited from experts to be used in model-based health technology assessments.

Methods Databases including MEDLINE, EMBASE and the CRD database were searched from inception to April 2013. Reference lists were checked and citation mapping was also used. Studies describing their approach to the elicitation of probability distributions were included. Data was abstracted on pre-defined aspects of the elicitation technique. Reports were critically appraised on their consideration of the validity, reliability and feasibility of the elicitation exercise.

Results Fourteen articles were included. Across these studies, the most marked features were heterogeneity in elicitation approach and failure to report key aspects of the elicitation method. The most frequently used approaches to elicitation were the histogram technique and the bisection method. Only three papers explicitly considered the validity, reliability and feasibility of the elicitation exercises.

Conclusion Judged by the studies identified in the review, reports of expert elicitation are insufficient in detail and this impacts on the perceived usability of expert-elicited probability distributions. In this context, the wider credibility of elicitation will only be improved by better reporting and greater standardisation of approach. Until then, the advantage of eliciting probability distributions from experts may be lost.

Key Points for Decision Makers

- The elicitation of probability distributions from experts to enhance the modelling process is an integral part of health technology assessment
- The majority of reports presenting expert elicitation of probability distribution were incomplete, making critical appraisal of these exercises difficult
- By disseminating reports of such exercises conducted in health technology assessment, research is encouraged towards building a framework for conducting and evaluating the elicitation of probability distributions

1 Background

Model-based economic evaluations carried out as part of health technology assessments (HTA) rely on the synthesis of various types of scientific evidence and economic data to inform policy decisions on the optimal use of healthcare resources. When research data needed for cost-effectiveness decision analytic models are lacking, the opinion of clinical experts is a widely accepted source of evidence, routinely used despite expert opinion being regarded as the least reliable form of scientific evidence [1].

The process of formally capturing expert opinion for use in decision-analytic models in healthcare has been

Electronic supplementary material The online version of this article (doi:10.1007/s40273-013-0092-z) contains supplementary material, which is available to authorized users.

B. Grigore (✉) · J. Peters · C. Hyde · K. Stein
Peninsula Technology Assessment Group (PenTAG), Institute of Health Research, University of Exeter Medical School, University of Exeter, Exeter, UK
e-mail: bogdan.grigore@pcmd.ac.uk

addressed in a number of papers [2, 3], but very little of the literature is concerned with the elicitation of probability distributions from experts on unknown quantities. Recent developments in health research [4] acknowledge the usefulness of eliciting expert opinion in a probabilistic form, especially since probabilistic sensitivity analysis is increasingly becoming the norm in HTA [5], and also because value of information analyses can be conducted to direct further research [6]. However, good practice guidelines for decision-analytic modelling in HTA only recommend that the selection of experts and methods used are documented in a transparent manner [7] and do not provide guidance for eliciting probability distributions in particular. It may not be surprising that exercises eliciting probability distributions from experts are either not used or not reported in economic evaluations of health technologies.

Arguably, eliciting subjective probability distributions from experts is more complex than asking for point estimates; it requires a degree of familiarity with statistics from the expert and usually a formal framework for the elicitation sessions. Many factors influence the elicitation process and researchers have to make methodological choices through all of the elicitation steps in order to achieve a reasonable balance between the accuracy of the elicitation, the resources allocated to the exercise and generally keeping within the time constraints of the HTA project (see Table 1 for examples).

Selection of experts is a key step in a successful elicitation. Ideally, elicitation should be conducted with a number of experts that have the most expertise while not sharing the same perspective [8]. Availability and willingness to participate, as well as potential conflicts of interest also need to be taken into account [4]. There is little literature on the number of experts to elicit from, but research in other fields [9] suggests that between six and twelve experts should be included in most elicitation exercises.

For the elicitation strategy, consideration needs to be given to whether the experts will be asked individually or as a group. Each has advantages and disadvantages and choices need to be made on a case-by-case basis [4].

Elicitation of probability distributions may prove difficult even for highly numeric health professionals, especially when eliciting unknown quantities [10]. To familiarise experts with the task and minimise the impact of heuristics, a preparation step is usually included in the elicitation exercise. Regardless of the elicitation method, usually only a small number of values are elicited from the expert, and an assumption is made about fitting a parametric distribution to these values. The question then is whether to elicit the values that represent the location and spread of these distributions directly or indirectly [11]. O'Hagan et al. [4] provide a detailed overview of various elicitation methods, but only a few have been used by analysts in HTA [12].

One of the most common methods is the histogram technique [13]. This is a fixed interval method that allows quantities of interest to be elicited graphically. It is a discrete form of the probability density function (PDF) [4], where the expert is presented with a frequency chart on which he or she is asked to place a number of crosses (alternatively called chips or tokens). Placing all the crosses in one column would represent complete certainty, while placing all the crosses on the bottom row would

Table 1 Aspects requiring a rational methodological choice by analysts doing elicitation

Aspect	Examples	Description of potential trade-offs
Expert selection	Which experts to select?	If experts are selected from a group of people with very similar experiences, there is a risk of overrepresentation of a certain opinion; however, on an obscure topic, the pool of available experts may be very limited
	The risk of motivational bias	Sometimes, the experts with the best exposure to the relevant topic may want to influence the outcome of the elicitation, irrelevant of their true belief [37]
Method selection	The elicitation method	There is no agreement on which specific elicitation method is most adequate for eliciting expert opinion in a probabilistic form [10]
	Group or individual approach	The combination of opinions from individual experts can be done through behavioural methods (e.g., consensus panels) or mathematical methods (e.g., linear opinion pooling) [16]
	Face-to-face facilitated vs. self-administered elicitation	A face-to-face facilitated session requires a trained analyst to guide the experts through the questions, a process that is resource intensive (considering also time and travel-related expenses), while a self-administered questionnaire is easier to send out, but requires careful preparation of the questionnaire, and response rates are typically low [4]
	Calibration vs. equal weighting for mathematical synthesis	Through calibration, the opinion of more knowledgeable experts can be better represented, but it is difficult to identify relevant weighting criteria (which experts are better informed than others) [23]
Questioning strategy	Framing bias	The way questions are formulated may influence the estimates [38]

represent complete uncertainty about the true value (Fig. 1). The histogram technique has been used in several different forms, including the Trial Roulette [14], where the expert is presented with diagrams representing betting streets, similar to those used on a gaming table.

Another popular method to capture expert opinion based on the PDF is a hybrid method (also known as 'six complementary intervals' [15]). First the lowest (L), highest (H) and most likely value (M) of the probability of a Bernoulli trial are elicited. Intervals are automatically built using a formula to divide the distance between each extreme (L and H) and M into three equal parts. Experts are then asked to enter the probability that their estimated value lies within each interval.

Another common method is the bisection method [10]. This is a variable interval method based on the cumulative distribution function, which entails a sequence of questions to elicit the median and the lower and upper quartiles; the

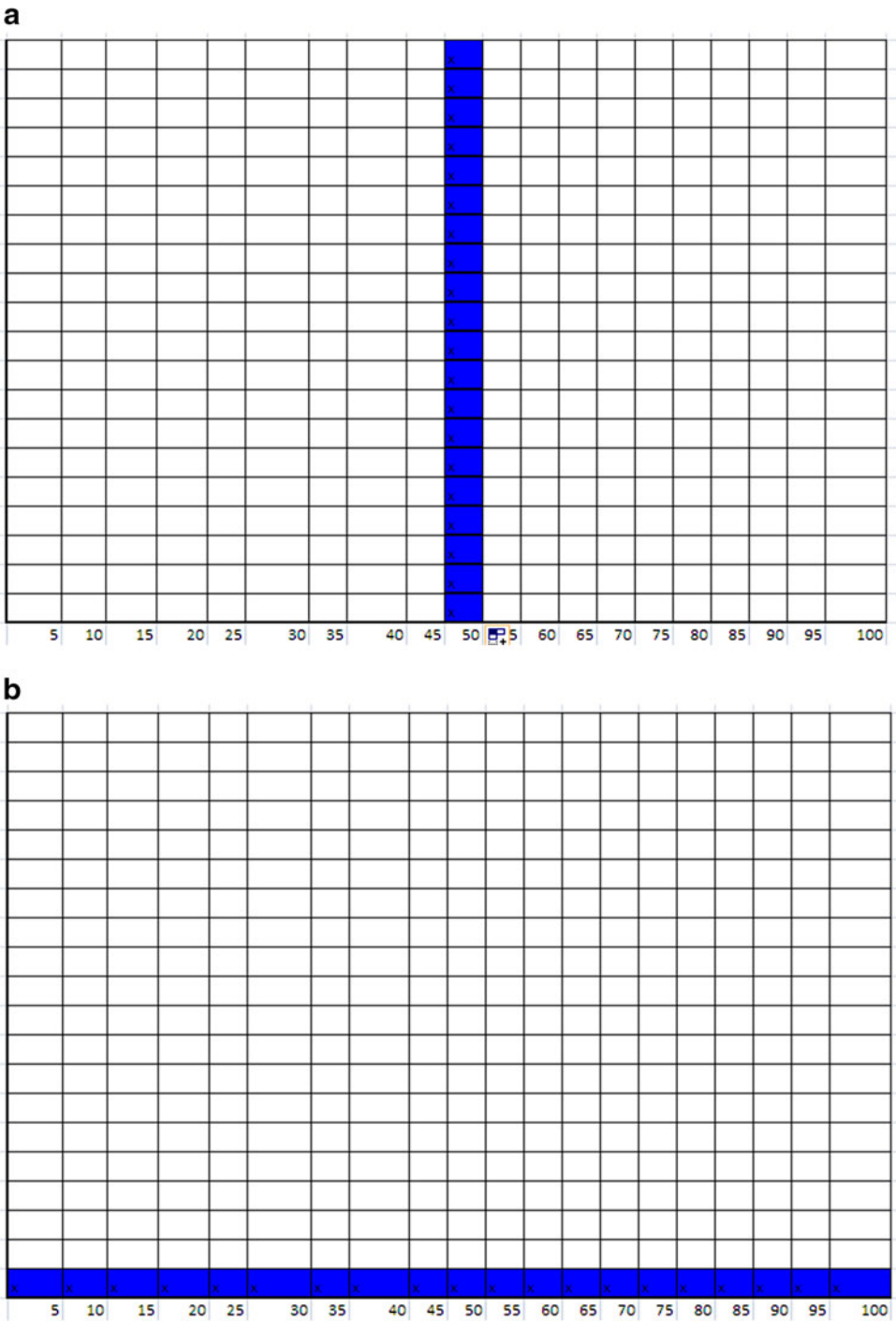

Fig. 1 Complete certainty and complete uncertainty represented with the histogram technique. **a** *Crosses* indicate complete certainty that the value of interest is in the interval 45–50; **b** *Crosses* indicate complete uncertainty, as the value of interest is equally likely to be anywhere between 0 and 100

main characteristic of this technique is that it requires only judgements of equal odds from the expert (e.g., the median would be a point such that the value of interest is equally likely to be less than or greater than it).

When elicited probabilities are obtained from multiple experts, they can be aggregated, either through behavioural or mathematical methods. There is debate on which aggregation method is the most appropriate, with evidence favouring both [4, 16]. In mathematical aggregation, another issue is the calibration of individual opinions (where opinions of experts assessed as 'better' receive more weight in the combined distribution). In practice, experts could be equally weighted (so no expert is considered 'better' than any other expert) [16], or certain criteria could be used to give different weights to the experts; for example, it is possible to validate an expert's opinion against existing data in seed questions (questions for which the answer is known). However, the ability to accurately recall available data is not an accurate predictor of judgement about unknown data [4].

There are currently no published guidelines on good practice elicitation of probability distributions in HTA, nor are there any agreed properties on which to measure elicitation methods. A number of studies [17–19] have, however, tried to define frameworks that would potentially allow the quantification of elicitation methods, by applying the criteria of measurement science [20]. The properties considered for elicitation are validity, reliability, responsiveness and feasibility [18]. With the exception of validity, there has been little discussion of these properties within expert elicitation. In terms of validity, the assessor must be aware that what can be obtained from the expert is not the truth, but the expert's belief about the truth; consequently, a successful elicitation is one that faithfully records the expert's belief [4]. A good measure of face validity of elicited probabilities is obtained when experts are presented with a graphical representation of their expressed belief and are given the opportunity to adjust their response. Validation of the elicited probabilities can be performed by eliciting the same estimate from more than one expert [21], or by comparing them with available empirical data. Validation of an expert's estimates against empirical data is usually not possible for the quantities of interest, but sometimes the expert is asked seed questions [22] in order to have a measure of the reliability of their estimates.

Achieving a good balance among these factors is usually a challenge, with many practical limitations, and so the reporting of such considerations is important when evaluating the elicitation process.

The objective of this study was to systematically summarise the methods used to elicit probability distributions from experts undertaken to inform model-based economic evaluations in healthcare.

2 Methods

2.1 Search Strategy

Identifying studies that reported the formal use of expert opinion to inform decision models was challenging, as pilot searches indicated that not all modelling exercises report in their abstracts whether expert opinion was captured and used as part of the research, much less if probability distributions were elicited.

The search strategy was constructed from a selection of keywords collected from relevant papers uncovered by the pilot search. Keywords were grouped into three categories defining the type of studies (economic evaluations, HTAs), type of input (expert opinion/knowledge/judgement, expert panel, advisory group) and outcome (subjective probabilities, Bayesian priors) (see Electronic Supplementary Material for detailed search strategy). Mapping of keywords to subject headings was not possible, as no topics or Medical Subject Heading (MeSH) terms were identified for expert elicitation, and use of generic headings like 'expert testimony' did not result in any relevant results.

The following databases were searched: EMBASE (1974 to April 2013), MEDLINE (In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to April 2013), Web of Science, CINAHL, the CRD database (including the NHSEED, HTA and DARE databases).

2.2 Inclusion and Exclusion Criteria

After removal of duplicate articles across databases, the titles and abstracts were screened by two reviewers (BG and JP). Articles describing a decision analytic model as part of an HTA, which used expert opinion, were included for full text screening. Any disagreements between the two reviewers were resolved by discussion.

The remaining articles were screened in full text and those that contained only elicited point estimates (not probability distributions) or did not describe the elicitation methods were excluded.

The references of included articles and papers quoting these articles were searched for other potentially eligible studies, through the Quotation Map tool from Web of Science. Where several papers described aspects of the same modelling process, only the reference containing the most complete account of the elicitation was included.

2.3 Data Extraction

Using a data extraction tool constructed specifically for this task, based on the elicitation methods literature [4, 9, 23], the details shown in Table 2 were extracted.

Table 2 Aspects recorded during data extraction

1. Preparation of the elicitation	Purpose of elicitation; type and format of parameters (e.g., epidemiological, clinical, resource use, etc.)
	Reporting of elicitation planning, piloting
	Elicitation strategy:
	Level of elicitation (individual/group elicitation);
	Elicitation method, format of estimates (e.g., intervals, distributions, shape, etc.)
	Mode of administration (e.g., interview, paper, PC, etc.); facilitated or not
	Selection of experts
2. Elicitation session	Reporting of experts preparation
	Opportunity of revision by experts, feedback
3. Post-elicitation	Fit/assignment of smooth function
	Aggregation of estimates from individual experts
4. Reported critical assessment	Consideration of validity
	Consideration of reliability
	Consideration of feasibility

2.4 Critical Appraisal

We have sought to determine whether included articles considered any of the following properties in the reporting of the elicitation exercise.

- Validity: *face validity*, whether the experts were asked if the elicited probability distributions reflected their beliefs, or *criterion validity*, where the elicited distributions are validated against available data [20].
- Reliability: reproducibility of the reported exercise. Whether there is sufficient description of the method and bias management strategies; in particular, the preparation and possible calibration of the experts.
- Feasibility: reporting and discussion of the complexity of the task and the logistics related to the elicitation exercise, any choices and trade-offs made in building the elicitation exercise, and how these may have been reflected in response rates and proportion of valid responses; dealing with invalid responses or failed exercises is also important from a feasibility point of view.

3 Results

Fourteen articles describing elicitation exercises for use in a HTA context were included (Fig. 2). Of these, ten were identified from the database search and four from screening references and citations of included studies. See Table 3 for details on the included papers.

3.1 Summary Statistics

Of the 14 included studies, two were from the grey literature [24, 25], as they were publicly available and met the inclusion criteria. Four of the 12 were exclusively dedicated to the description of the elicitation exercise, independent of the modelling process [12, 21, 26, 27], and one [28] was a methodological paper which also reported an elicitation study. The rest of the studies reported the elicitation exercise with less detail, but explicitly indicated elicitation of probability distributions for a number of model parameters.

3.2 Preparation of the Elicitation

All studies report using health professionals as experts, but three studies do not report the number of experts involved. The median number of experts included in the other ten studies is 5 (mean 9.2, ranging from 3 to 23 experts per study). Details of the strategy used to select experts were reported in seven studies; in six of these [12, 14, 21, 26, 27, 29] substantive expertise is an explicit criterion for expert selection. Two studies [12, 26] report specifically selecting experts with different experiences, while in Stevenson et al. [30], the experts were part of the study team. One study [28] specifically targets geographically dispersed experts, in an effort to explore the widespread collection of quality data. No further aspects of expert selection are reported in the studies.

The elicitation method used is arguably the most important characteristic of the exercise, and all reports contain some information on this. Four studies [12, 25, 26, 29] report constructing pilot exercises to assess various methods to conduct and combine elicitations. Ease of use was an important factor in choosing the appropriate method, as it predicted higher compliance by the experts, especially when a facilitator was not available [25, 26, 28].

Eleven of the 14 studies report providing training in probability elicitation for the participating experts prior to

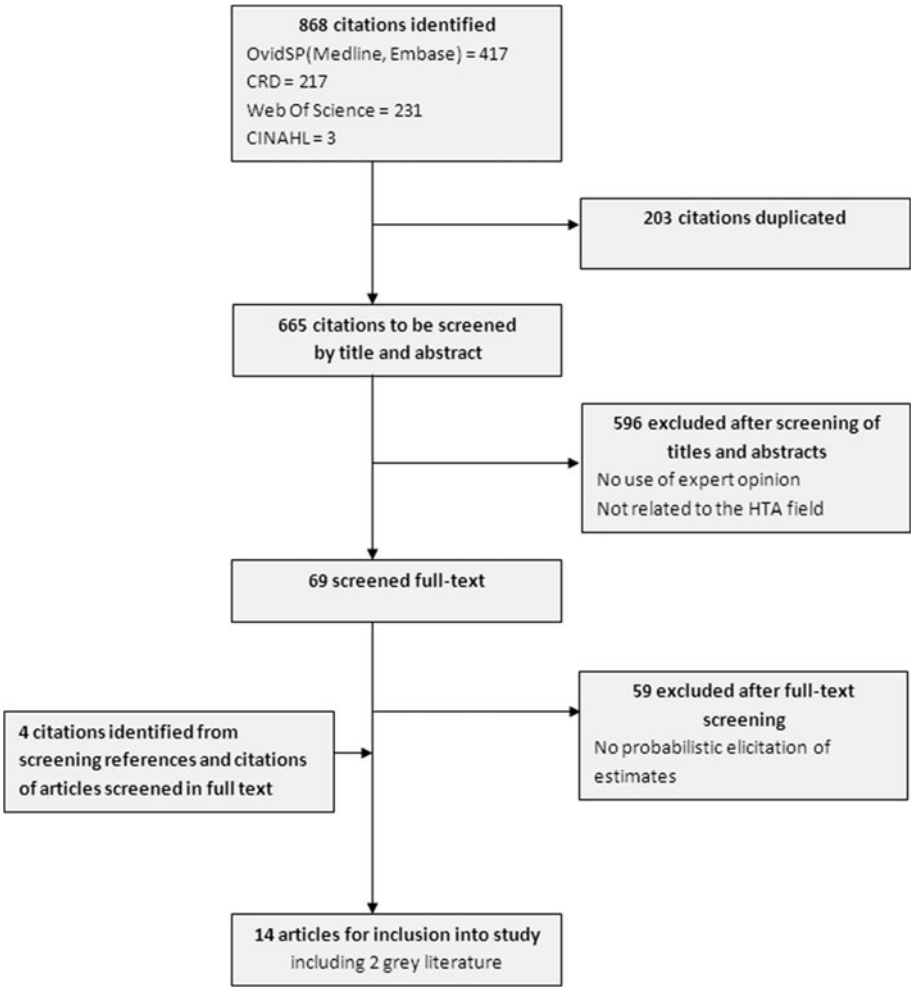

Fig. 2 The flow diagram of search results

undertaking the elicitation session. In seven studies [14, 21, 26–29], experts were given written instructions for completing the questionnaires and in five of these, background information on the modelling exercise was given. Four studies reported the presence of a facilitator during the elicitation [12, 21, 31, 32], and in one study [12], experts attended a 2-hour preparatory training session. Details on the training of experts could not be obtained from the remaining studies. Thirteen studies report conducting the elicitation with experts individually, with only one study [31] using a group consensus method to elicit the probability distributions. See Table 3 for further details.

All included studies used elicitation to inform effectiveness (for treatments), performance (for diagnostic tests) or disease progression parameters, with only one study eliciting resource use [21]. The elicited parameters were in the form of a proportion in six studies, a rate in three studies, time to event in three, and relative risk in two studies.

In three studies, experts were asked directly for their estimates of:

- the mean and the lower and upper 95 % confidence limits [33];
- the mode (described as 'the most likely value') and the lower and upper 95 % confidence limits [25];
- various unspecified quantiles [31].

Four studies [12, 14, 27, 29] used the histogram technique and three studies [21, 28, 30] used the bisection method. Leal et al. [26] report the use of a method based on the 'six complementary intervals' method (the hybrid method), but using instead only four intervals. Meads et al. [32] used the allocation of points technique (similar to the histogram technique in that the experts allocate a total of 100 points to a number of fixed, predefined value ranges). One study [34] does not explicitly report the elicitation method used, but does provide values for the median, 10th and 90th percentiles, while another study [24] gives no indication of the elicitation method used.

Five clinicians participated in a consensus-reaching exercise in Girling et al. [31]. The elicitation entailed the discussion of a small number of quantiles for each parameter. It is not clear from the paper which quantiles

Table 3 Characteristics of the included studies

References	Purpose of elicitation	Types of parameters elicited	Format of parameters	Format of estimates	Administration method	Feedback/ possibility of iteration by expert	Smooth function fit (level, distribution used)	Aggregation
Bojke et al. [27]	Probabilistic decision analytic model to assess the cost-effectiveness of infliximab and etanercept for the treatment of active psoriatic arthritis (PsA), compared with palliative care	Effectiveness parameters Clinical parameters	Rates	Histogram technique	Computer-based; non-facilitated	Y	NR	Linear pooling and random effects meta-analysis (both with and without weighting)
Brodtkorb [24]	Economic evaluation for prosthetic knees for amputees	Effectiveness parameters	Rates Time to event	NR	NR		Y (Weibull)	NR
Colbourn et al. [33]	Cost effectiveness of prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy	Effectiveness parameters	Relative risk	Mean and 95 % CI	Questionnaire; facilitation NR	NR	Beta distribution	Random effects meta-analysis using WinBUGS
Garthwaite et al. [21]	Model for care pathways in colorectal cancer	Epidemiological parameters Clinical parameters Resource use	Time to event Proportions	Median and quartiles (bisection method) or median and min/max	Computer-based; facilitated	Y	Y (only multivariate; as normal distributions)	N
Girling et al. [31]	Early cost-effectiveness assessment of left-ventricular assist devices (LVAD)	Effectiveness parameters	Proportion Time to event	Quantiles	Group elicitation; computer-based; facilitated	Y	Y (in real time)	Behavioural
Haakma [25]	Early evaluation of value of photoacoustic mammography	Performance parameters	Rates	Mode, lower and upper limits (95 % CI)	Face-to-face interview		Y (normal)	Y (linear pooling)
Leal et al. [26]	Economic evaluation of the long-term costs and effects of alternative approaches to diagnosing and managing hypertrophic cardiomyopathy (HCM) for those at risk of sudden cardiac death (SCD)	Epidemiological parameters Performance parameters	Proportions	Mode and upper and lower limit	Email; computer-based; non-facilitated	Y	Y (to aggregated)	Linear pooling
Meads et al. [32]	Economic modelling of PET–CT/CT in detecting and managing recurrent cervical cancer	Epidemiological parameters Performance parameters	Proportions	Allocation of points technique	Pen-and-paper; facilitated	NR	Y (to aggregated, normal)	Y (details are unclear)
McKenna et al. [29]	Model for the cost effectiveness of enhanced external counterpulsation	Effectiveness parameters	Proportions	Frequency chart (histogram)	Computer-based; facilitation NR	Y	Y (to aggregated, beta)	Linear pooling
Soares et al. [12]	Markov decision model for negative pressure wound therapy for severe pressure ulceration	Effectiveness parameters Clinical parameters	Proportions	Histogram technique	Computer-based; facilitated	Y	Y (to aggregated; beta or normal)	Linear pooling
Speight et al. [14]	Model for the cost effectiveness of oral cancer screening	Epidemiological parameters	Proportions	Trial Roulette (histogram)	Questionnaire; facilitation NR	NR	Y (to aggregated, beta)	NR
Sperber et al. [28]	Development of a distance elicitation tool	Clinical parameters	Proportions	Quartile-bisection method	Computer-based; non-facilitated	Y	Y (to individual, beta)	Linear pooling
Stevenson et al. [34]	Model-based assessment of the surgical instrument management policies to reduce the risk of vCJD transmission	Epidemiological parameters Effectiveness parameters	Count Proportions Time to event	Probability distributions	NR; facilitated[a]	NR	Y (beta or normal)	NR

Table 3 continued

References	Purpose of elicitation	Types of parameters elicited	Format of parameters	Format of estimates	Administration method	Feedback/ possibility of iteration by expert	Smooth function fit (level, distribution used)	Aggregation
Stevenson et al. [30]	Cost effectiveness of an RCT to establish whether 5 or 10 years of bisphosphonate treatment is the better duration for women with prior fracture	Effectiveness parameters	Relative risk	Bisection method		Y	Y (to both individual and aggregated; lognormal)	Sampling

NR not reported, *RCT* randomised controlled trial, *vCJD* variant Creutzfeldt-Jakob disease

[a] Not explicit, but implied in the text

were elicited, and although the description fits with the bisection method of elicitation, the actual method used was not reported. Garthwaite et al. [21] also elicited quantities with several covariates using specially developed software.

Beside the facilitated sessions, which favoured interaction with the facilitator [12, 32], other studies report how experts were given the opportunity to adjust their elicitations: either a summary of their answers at the end of the session [12, 26–28] or by seeing the smoothed aggregated distribution of all the experts [34].

In 11 studies, a smooth function was fitted to the individual elicited distributions, and for five studies [12, 14, 26, 29, 32] the smooth function was fitted to the aggregated estimates. In the group elicitation study [31], a computer-generated density function was presented in real time to the clinicians; this was amended as necessary until the whole group was satisfied with the shape.

Analysts fitted beta [12, 14, 25, 28, 29, 33, 34], normal [12, 30, 32, 34] or Weibull [24] distributions to the elicited estimates. Best fit was determined mathematically by the method of moments [12] or the least squares approach [28, 30]. Leal et al. [26] evaluated goodness of fit by drawing the aggregate histogram and PDF curve together. To check the adequacy of the fitted distribution, Stevenson et al. [30] discussed additional quantiles like the 10th and the 90th percentiles. In the case of disagreement, the expert would either modify his or her initial judgements or an alternative parametric distribution would be considered as appropriate. One study reports fitting a normal distribution to multivariate parameters [21]. Although it appears that all elicitation exercises included a fitting of a smooth function, no details are provided in the remaining studies. The issue of uncertainty introduced by smooth function fitting was commented on by only one study [12].

Seven papers include sample questions from the elicitation questionnaire [8, 12, 14, 21, 26, 28, 29], in order to give a clearer image of the elicitation. Sperber et al. [28] also provide the complete elicitation tool used as supplementary, web-only material.

3.3 Post-Elicitation

Mathematical aggregation of the elicited distributions across experts was reported in nine studies. Aggregation of data was conducted by either linearly pooling individual opinions [12, 25–27, 29], using a random effects model [27, 33] or by sampling distributions of each expert and then combining the individual samples [30]. Some studies report using different weights for individual distributions. Calibration of individual experts was done through seeding questions (questions whose answers are known to the investigator, but not to the expert) [27] or a synthetic score based on individual characteristics of the expert (i.e., years of experience) [25]. Bojke et al. [27] compared the results of aggregation using linear pooling and a random effects model, with and without calibration. They found that results were sensitive to aggregation method used and whether calibration was used or not [27]. Soares et al. [12] evaluated the use of calibration in their pilot exercise and reported differences between the combined estimates with and without calibration. However, they discarded calibration for the case study because of difficulties in choosing the most appropriate criteria for weighting the individual elicitations. Sperber et al. [28] also explored several aggregation methods and concluded that unweighted opinion pooling was the optimal method.

In Garthwaite et al. [21], responses for the same question from different experts were empirically compared as a measure of validity and only the estimates of a single expert were used for each parameter. As part of their feasibility assessment, non-facilitated elicitation exercises translated into a mean response rate of 40.63 %, while in elicitation conducted through facilitated sessions, the issue of response rate is not discussed.

Several studies report experts failing to complete the exercise. The main reason for this was the inability to use the software delivering the questions [26, 27]. Other reasons why responses could not be used were invalid answers [12], outlying estimates [25] or the questions being altered

Table 4 Critical appraisal of included studies

Article	Consideration of validity	Consideration of reliability	Consideration of feasibility
Bojke et al. [27]	Experts given a summary of their responses and the opportunity to feedback and/or restart the questionnaire. Authors note heterogeneity of responses and discuss possible causes for this	Training: experts given background information and went through some example questions Elicitation task: method described Calibration: weights were attributed based on a seed question; aggregation of responses conducted both with and without weights	Response rate: 5/16 (31 %) Technical burden is considered as a possible cause for the low response rate Expert selection: authors recognise trade-off between the geographical area for recruitment and the generalisability of the study, and discuss possibility that experts participating might be different from those not participating
Brodtkorb [24]	NR	NR	NR
Colbourn et al. [33]	NR	NR	Expert selection: four experts participated Response rate: NR
Garthwaite et al. [21]	Authors report "good practice in using expert judgement" during the sessions Each expert's responses were validated with those of other experts, but exact procedure is not reported	Elicitation task: method described To prevent the expert from providing unreliable responses, questions were simplified to Bernoulli trials (only two possible outcomes) Multivariate estimates are recorded separately It is unclear whether mathematical aggregation was conducted	Expert selection: experts were "well chosen" Training: experts given an information pack; a facilitator was present in all the sessions Response rate: 4/4 (100 %) Authors noted that some responses provided by one expert could not be used fully
Girling et al. [31]	Authors compare results with unpublished data and conclude they were 'similar'. No other validity considerations are reported	Elicitation task: it is reported that "the elicitation procedure was that described in Garthwaite et al." [10], but method used is unclear, as the referenced paper describes several elicitation methods	Expert selection: five clinical experts Group elicitation
Haakma [25]	Experts were given a visual representation of their estimates and the opportunity to adjust their responses	Training: experts were given background information and explanation of the elicitation purpose and process Elicitation task: method described Calibration: based on the experts' background experience (years of experience, number of cases seen per week)	Alternative elicitation methods were assessed in a pilot study with three experts, and method chosen on basis of ease of use One expert was excluded from the study as responses were considered as an outlier Experts selection: described Response rate: 17/18 (one expert's estimates were not used)
Leal et al. [26]	Feedback was requested from the experts on the distributions fitted to their summaries Experts could also provide feedback on the format and content of the task	Experts selection: chosen to represent the population of patients Training: experts given background information and instructions to complete the questionnaire, but it is unclear if example questions were used Elicitation task: method described Impact of covariates limited by using Bernoulli trials (only two possible outcomes for each question) Calibration: considered but not used, as limited number of experts and the novelty of the topic	Alternative elicitation tools assessed in pilot study and the 'four complementary intervals' method chosen on basis of reported ease of use Response rate: 6/12 (50 %)
Meads et al. [32]	Validity tested by comparison with published data	Training: presentation of the project and the summaries to be elicited given to experts. Example questions were used Elicitation task: method described	Experts selection: 21 experts participated Sessions conducted on the occasions of several specialist meetings Response rate: NR
McKenna et al. [29]	NR	Training: not explicitly reported, but a sample of the questionnaire used is reported with background information and instructions Elicitation task: method described Overconfidence limited by first eliciting extremes	Response rate: 5/7 (71 %) completed the exercise The reasons why two experts failed to complete the questionnaire are not reported

Table 4 continued

Article	Consideration of validity	Consideration of reliability	Consideration of feasibility
Soares et al. [12]	Experts given visual representations of their estimates and could amend, if necessary Elicited opinions were also validated with published data where possible	Expert selection: selection bias minimised by selecting experts from relevant settings (hospital and community-based; both specialists and generalists) Training: experts given a 2-h training session Elicitation task: method described Impact of covariates limited by using binary outcomes Calibration: evaluated in a pilot exercise, but not used in main elicitation exercise, as difficulty in assessing the appropriateness of seed questions	Alternative elicitation methods and calibration assessed in a pilot study Response rate: 23/23 (100 %) Authors report some missing values and one expert providing some inconsistent responses
Speight et al. [14]	NR	Elicitation task: method described	Experts selection: nine clinical experts participated Response rate: NR
Sperber et al. [28]	Experts given visual representations (histogram, fitted distribution) of their responses.	Training: background information given. Purpose and process of elicitation described to experts Elicitation task: method described Calibration: compared results of the mathematical aggregation with response of the clinical collaborator in the study team, and with the distribution of median responses from the experts	Some responses were found invalid or inconsistent, mainly due to lack of specific guidance for the respondents
Stevenson et al. [34]	Elicited responses were shown to other experts for comment	Training: experts given background information and available literature Elicitation task: reported use of 'formal elicitation techniques', but no other details are provided	NR
Stevenson et al. [30]	Fitted distribution was validated by discussing further percentiles from the experts Experts had an opportunity to adjust their summaries, or consider an alternative parametric distribution Combined distribution was also validated by the experts	Expert selection: the participating experts were part of the study team Elicitation task: method described	Experts selection: three clinical experts participated Response rate: NR

NR not reported

by the expert so they could give answers based on data available to them [21]. One paper [28] reports inconsistencies between the expert's stated uncertainty and the provided values, by stating that it wasn't their area of expertise, yet stating 100 % certainty of the value lying between 0 and 0.1.

3.4 Critical Appraisal

In 12 of the included articles, the reports suggest that aspects of validity, reliability and feasibility of the elicitation exercise were considered (see Table 4). However, with very few exceptions [12, 21, 28], based on the study reports, these properties were not explored explicitly.

Nine studies (64 %) mentioned steps to assess the validity of the elicited distributions, usually by asking the expert to reflect on the elicited distributions; ten studies (71 %) consider aspects of reliability of the elicitation exercise (such as providing sufficient description of the method, containing details on elicitation training or pooling distributions elicited from several experts) and eight studies (57 %) report aspects related to the feasibility of the elicitation exercise.

Overall, reporting on the aspects of validity, reliability and feasibility of the elicitation exercises was insufficient across these 14 papers.

4 Discussion

This review uncovered a relatively small number of studies reporting the use of elicitation of probability distributions in HTA, yet these studies were heterogeneous in their reporting of the conduct of the elicitation exercises.

The dominant observed approach to elicitation was getting individual estimates from several experts and combining them mathematically, without calibration. The most frequently used elicitation methods were the histogram technique and the bisection method, with a number of variations. The preferred aggregation methods were linear pooling and random effects meta-analysis. Calibration was considered in some studies but not used, as choosing the

appropriate criteria to calibrate was reported to be a significant deterrent. Other methods of elicitation have been used, including consensus-reaching group elicitation.

Use of strategies for controlling bias was reported across the studies, but varied greatly. Some studies report using empirical measurements for selecting the elicitation method, like piloting of different methods and asking experts for their preferences for one method or the other. Most of the elicited parameters were related to the effectiveness of intervention.

There were a number of reporting limitations in the included studies. Some studies omit reporting relevant data such as the expert selection strategy, the elicitation or aggregation methods. These omissions may be due to limited word counts or because these aspects were not considered relevant. This may negatively affect the perceived quality of the elicitation, even if all reasonable steps were taken to ensure an adequate procedure. Furthermore, there are inconsistencies in terminology, as five different studies report using a very similar method (the histogram technique), but name it in three different ways. This lack of standardised language makes it difficult to categorise and index the articles on the elicitation method used.

The use of expert elicitation of probability distributions is a way to enhance the quality of expert opinion in HTA and is useful in exploring the overall uncertainty that the decision makers are confronted with, as well as in directing further research. However, the perceived complexity of expert elicitation may limit its use. Trade-offs are inherent to the elicitation process, as assessors need to balance limited resources and deadlines with achieving optimal response from the experts [35]. In the included studies, most choices were made with investigators reportedly being aware of potential trade-offs. For instance, there was a significant difference between the rate (and quality) of responses between facilitated and non-facilitated elicitation exercises, with the latter having a low response rate. This was attributed on several occasions to the technical burden on the experts to provide probability distributions in the absence of sufficient statistical knowledge. Some studies indicated that having a facilitator present was not possible because of time constraints and, in at least two studies, investigators anticipated this trade-off and recruited more experts.

Our review is limited by elicitation of probability distributions not being indexed as a subject heading term in the usual search engines and such terms not being present in the abstract or keywords of relevant articles. We tried to deal with this shortcoming by adopting a search strategy that included specific terms such as opinion/knowledge/judgement and subjective probabilities. Although it is possible that this search missed relevant studies if these terms were not used, screening the references of the included papers identified further relevant studies. However, it is still possible that relevant papers were not identified in our search strategy.

No dedicated framework for the critical appraisal of elicitation exercises exists. However, we determined whether authors had considered aspects of validity, reliability and feasibility of the elicitation. The reports suggest that these aspects were considered to a certain extent in the majority of the studies, but not explicitly so. More research is needed to define measurement criteria for the elicitation methods used in HTA. Until these become available, elicitation exercises should be approached alongside strategies that consider the validity and reliability, and should be reported with consideration of methodological trade-offs and the expected impact on the elicitation outcomes. We consider that more complete reporting of the elicitation exercises in the future would help determine whether these aspects had been considered.

Although a checklist for reporting items could be useful to this end, further work is needed in the development of such guidelines [36]. However, it is our opinion that, as a minimum, elements defining the validity and reliability of the elicitation exercise (method description, selection and preparation of experts, calibration and synthesis of individual distributions) should always be reported explicitly in relation to an elicitation exercise. Aspects outside of the analysts' control (like availability of experts) are also important and should be reported transparently as part of the feasibility assessment. Where the description of elicitation is limited by word count, these reports could be provided as web-only appendices.

5 Conclusions

This review summarises the reported current use of expert elicitation of probability distributions in HTAs. Elicitation is a complex process and, as our review shows, it can be difficult to conduct and report. This affects the perceived quality of the exercise and understates the added value of expert opinion obtained as probability distributions.

Based on the findings of this study, we recommend that reporting of elicitation exercises informing HTAs should be more complete, with any consideration of validity, reliability or feasibility explicitly presented. Further efforts should be made towards agreement of definition and content of these measurement properties, in order to improve standardisation of approach. This will facilitate the critical appraisal of reported expert elicitations, greatly enhancing their credibility as genuine tools for HTA, as will the development of reporting guidelines for the elicitation or probability distributions.

Acknowledgments This research was funded by the National Institute of Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Author Contributions All authors contributed to the conception of the study. BG and JP collected and extracted the data. All authors contributed to interpretation of the data and to the final manuscript. Ken Stein is the guarantor for the overall content of this paper.

Conflict of Interest All authors have no conflicts of interest.

References

1. Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. J Clin Nurs. 2003;12(1):77–84.
2. Evans C, Crawford B. Expert judgement in pharmacoeconomic studies: guidance and future use. Pharmacoeconomics. 2000;17(6): 545–53.
3. Sullivan W, Payne K. The appropriate elicitation of expert opinion in economic models: making expert data fit for purpose. Pharmacoeconomics. 2011;29(6):455–9.
4. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: eliciting experts' probabilities. Chichester: Wiley; 2006.
5. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ. 2005;14(4):339–47.
6. Claxton KP, Sculpher MJ. Using value of information analysis to prioritise health research. Pharmacoeconomics. 2006;24(11): 1055–68.
7. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004;8(36):iii–iv, ix–xi, 1–158
8. Winkler RL, Poses RM. Evaluating and combining physicians' probabilities of survival in an intensive care unit. Manag Sci. 1993;39:1526–43.
9. Knol AB, Slottje P, van der Sluijs JP, Lebret E. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environ Health Glob Access Sci Source. 2010;9:19.
10. Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. J Am Stat Assoc. 2005; 100(470):680–701.
11. Winkler RL. The assessment of prior distributions in Bayesian analysis. J Am Stat Assoc. 1967;62(319):776–800.
12. 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):2363–80.
13. Van Noortwijk JM, Dekker A, Cooke RM, Mazzuchi TA. Expert judgment in maintenance optimization. IEEE Trans Reliab. 1992;41(3):427–32.
14. Speight PM, Palmer S, Moles DR, Downer MC, Smith DH, Henriksson M, et al. The cost-effectiveness of screening for oral cancer in primary care. Health Technol Assess 2006;10(14): 1–144, iii–iv
15. Phillips L, Wisbey S. The elicitation of judgemental probability distributions from groups of experts: a description of the methodology and records of seven formal elicitation sessions held in 1991 and 1992. Report nss/r282. Didcot, UK: Nirex UK; 1993
16. Clemen RT, Winkler RL. Aggregating probability distributions. In: Advances in decision analysis: from foundations to applications. USA: Cambridge University Press; 2007. p. 154–76.
17. Kadane J, Wolfson LJ. Experiences in elicitation. J R Stat Soc Ser D (Stat). 1998;47(1):3–19.
18. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. J Clin Epidemiol. 2010;63(4):355–69.
19. Kinnersley N, Day S. Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. Pharm Stat. 2013;12(2):104–13.
20. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010;68(4):319–23.
21. 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):350–7.
22. Cooke RM. Experts in uncertainty: opinion and subjective probability in science. USA: Oxford University Press; 1991.
23. Wolpert RL. Eliciting and combining subjective judgments about uncertainty. Int J Technol Assess Health Care. 1989;5(4):537–57.
24. Brodtkorb TH. Cost-effectiveness analysis of health technologies when evidence is scarce. Linköping: Linkoping University; 2010.
25. Haakma W. Expert elicitation to populate early health economic models of medical diagnostic devices in development. Twente: University of Twente; 2011.
26. Leal J, Wordsworth S, Legood R, Blair E. Eliciting expert opinion for economic models: an applied example. Value Health. 2007;10(3):195–203.
27. 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):557–64.
28. Sperber D, Mortimer D, Lorgelly P, Berlowitz D. An expert on every street corner? Methods for eliciting distributions in geographically dispersed opinion pools. Value Health. 2013;16(2): 434–7.
29. McKenna C, McDaid C, Suekarran S, Hawkins N, Claxton K, Light K, et al. Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic evaluation (structured abstract). Health Technol Assess. 2009;13(24):1–90.
30. Stevenson MD, Oakley JE, Lloyd Jones M, Brennan A, Compston JE, McCloskey EV, 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 Mak. 2009;29(6):678–89.
31. Girling AJ, Freeman G, Gordon JP, Poole-Wilson P, Scott DA, Lilford RJ. Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy (structured abstract). Int J Technol Assess Health Care. 2007;23(2):269–77.
32. Meads C, Auguste P, Davenport C, Małysiak S, Sundar S, Kowalska M, 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.
33. Colbourn T, Asseburg C, Bojke L, Philips Z, Claxton K, Ades A, 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.
34. 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. 2008;60(4):506–18.
35. O'Leary RA, Choy SL, Murray JV, Kynn M, Denham R, Martin TG, et al. Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby *Petrogale penicillata*. Environmetrics. 2009;20(4):379–98.
36. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010;7(2):e1000217.
37. Mazur A II. Disputes between experts. Minerva. 1973;11(2):243–62.
38. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211(4481):453–8.