

Eliciting Expert Opinion for Economic Models: An Applied Example

José Leal, MSc,¹ Sarah Wordsworth, PhD,¹ Rosa Legood, MSc,¹ Edward Blair, MD²

¹University of Oxford, Oxford, UK; ²Churchill Hospital, Oxford, UK

ABSTRACT

Objectives: Expert opinion is considered as a legitimate source of information for decision-analytic modeling where required data are unavailable. Our objective was to develop a practical computer-based tool for eliciting expert opinion about the shape of the uncertainty distribution around individual model parameters.

Methods: We first developed a prepilot survey with departmental colleagues to test a number of alternative approaches to eliciting opinions on the shape of the uncertainty distribution around individual parameters. This information was used to develop a survey instrument for an applied clinical example. This involved eliciting opinions from experts to inform a number of parameters involving Bernoulli processes in an economic model evaluating DNA testing for families with a genetic disease, hypertrophic cardiomyopathy. The experts were cardiologists, clinical geneticists, and laboratory scientists working with cardiomyopathy patient populations and DNA testing.

Results: Our initial prepilot work suggested that the more complex elicitation techniques advocated in the literature were difficult to use in practice. In contrast, our approach achieved a reasonable response rate (50%), provided logical answers, and was generally rated as easy to use by respondents. The computer software user interface permitted graphical feedback throughout the elicitation process. The distributions obtained were incorporated into the model, enabling the use of probabilistic sensitivity analysis.

Conclusion: There is clearly a gap in the literature between theoretical elicitation techniques and tools that can be used in applied decision-analytic models. The results of this methodological study are potentially valuable for other decision analysts deriving expert opinion.

Keywords: decision model, economic evaluation, expert opinion, genetic.

Introduction

As the number of economic evaluations in health care using decision-analytic models increases [1], the issue of obtaining robust data on model parameters becomes important. Frequently, not all the required data can be obtained from observed evidence (randomized controlled trial, cohort studies, etc.) or the literature; therefore, subjective information from experts (usually clinicians) may be required. A potential and relatively simple approach is to ask individual clinicians to provide parameter estimates based on their experience, average the estimates from different clinicians and then use the lower and upper estimates provided by the group to represent the variance around their estimates. Nevertheless, this approach provides limited information about the distribution of uncertainty surrounding parameters.

Understanding the shape of the distribution of uncertainty is particularly important in the context of the increasing use of probabilistic sensitivity analysis (PSA). Essentially PSA allows uncertainty in individual parameters to be propagated across a model simultaneously to explore the overall uncertainty in cost-effectiveness results [2]. PSA requires that distributions be assigned to all model parameters, with these distributions representing the range of values that a given parameter may take. Recent guidelines from the National Institute for Health and Clinical Excellence (NICE) in the UK recommend its routine use in decision models [3].

Recent decision modeling guidelines provide limited advice on methods to employ in order to obtain experts' opinions and quantify the distribution of uncertainty around individual estimates. The guidelines recommend that elicitation methods should be clearly documented and suggest that those methods aiming to reach a consensus (e.g., standard Delphi approaches) are inappropriate, as they may underestimate true parameter uncertainty [4]. While there is some debate in the literature over whether to use consensus or individual elicitation methods, there are

Address correspondence to: José Leal, Health Economics Research Centre, Department of Public Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. E-mail: jose.leal@dphpc.ox.ac.uk
10.1111/j.1524-4733.2007.00169.x

certainly advantages of using the latter, as it is possibly easier and less costly to conduct and, in probability judgements, group assessments have been suggested to lead to overconfident results (i.e., narrow distributions) compared with individual assessments [5].

To date, there is a largely theoretical literature on the most appropriate methods to adopt when eliciting expert opinion (see Garthwaite et al. [6], Kadane and Wolfson [7], and Cooke [8]). In addition, there are a few applied studies where expert opinion has been elicited in other fields such as radioactive wasting/contamination [9,10], clinical trials [11], water industry [12], occupational hygiene [13], and microbiological risk assessment in food processing [14]. In health-care decision modeling, the few practical examples where expert opinions have been used consist of estimating clinical resource use associated with a specific health intervention, or using utility values elicited from clinical experts in the absence of patient or public preferences [15,16]. Nevertheless, most documented examples have not used structured and tested methods to elicit opinions from experts as probabilistic distributions.

Our study aims to address this gap in available approaches, by developing a survey using a practical computer-based tool for eliciting individual expert opinions about the shape of distributions of single model parameters. Initially, we tested alternative approaches for eliciting opinions in prepilot work with departmental colleagues. We then used the approach that was reported to be the easiest to complete and the most accurate representation of their beliefs in an applied example related to the genetic disease hypertrophic cardiomyopathy (HCM). We first provide a brief background to this disease and the decision problem, and then outline the study methods and results.

Methods

Hypertrophic Cardiomyopathy and Decision Problem Background

Hypertrophic cardiomyopathy is a relatively common genetic condition with a disease prevalence of 1/500 [17,18]. It is defined by unexplained asymmetric thickening of the heart and can lead to sudden cardiac death (SCD).

Currently, we are undertaking an economic evaluation to assess the long-term costs and effects of alternative approaches to diagnosing and managing HCM for those at risk of SCD. A Markov model [19] is being used to compare the costs and effects (survival) of a genetic (DNA) with a nongenetic approach (purely clinical tests such as electrocardiogram) of diagnosing HCM within families. Nevertheless, as DNA testing for HCM is new and only recently starting to be moved from research into clinical practice, there is

Box 1 Model parameters where expert opinion was required

Proportion of HCM population at low/medium risk of SCD.
Transition from low/medium to high risk of SCD over a patient's lifetime
Detection of high-risk mutation carriers by the cardiology services
Detection of low-/medium-risk mutation carriers by the cardiology services
Effectiveness of implantable cardioverter defibrillator in the prevention of SCD in high-risk HCM patients
Effectiveness of amiodarone (drug therapy) in the prevention of SCD in high-risk HCM patients
Sensitivity of the genetic diagnostic test
Specificity of the genetic diagnostic test

very limited information on several model parameters, as is common with new technologies.

We identified eight model parameters (Box 1) where observational data were not available relating to: aspects of the natural history of this cardiac disease, treatment effectiveness and accuracy of the DNA test. Data on these parameters were lacking for a number of reasons. For instance, data on the accuracy of the DNA test in a health service setting are not yet available, as few patients have undergone testing. In the absence of observational data, we elicited distributions from experts for several parameters.

In this section of the article, the various steps of our empiric study are described. First, we explain the prepilot where alternative elicitation tools were assessed. Second, we outline the survey design for our applied clinical example. Third, information on the identification and selection of experts for this example is presented. Finally, the techniques used for combining the individual expert results are explained.

Prepiloting of Elicitation Tool

In line with Cooke [8], an important objective of our study was to develop an elicitation tool that was clear, attractive and could be completed reasonably quickly. A further objective was to provide a framework that could provide instant graphical feedback so that participants could observe the shape of distributions that they had specified. Our solution was to build a questionnaire in Microsoft Office Excel, a software package chosen because of its widespread use, its flexibility to perform calculations, and ability to produce graphs.

Prepilot work was undertaken to identify the most appropriate approach to eliciting opinions about the

shape of the distribution. The parameters of interest involved a Bernoulli process, which refers to a probabilistic experiment that can have one of two outcomes, for instance, patient dies or survives, sample is positive or negative for disease. A further example of a Bernoulli process in a model could be the risk of death in the 5 years after disease diagnosis. While it is possible to use a very simple elicitation technique to obtain expert opinion, for example, asking experts the mode and lower and upper estimates, this type of approach provides very limited information about the distribution of uncertainty surrounding parameters. A number of alternative techniques for quantifying a Bernoulli process have been suggested and tested in the literature and overall the quantile method has generally been found to yield a higher relative dispersion of the distribution (i.e., higher variance) compared with the other methods, perceived clarity of use and consistency in a betting situation [6,8,20]. With the quantile method (fixed interval method), the range of values that the parameter can take is elicited from the experts and divided into intervals by the study analyst. For each interval, the expert is then asked to provide the probability that the value will be contained in that interval.

There are different ways in which the quantile method can be applied in practice. Therefore, we used a dummy question on colleagues within the Department of Public Health (University of Oxford) to perform a prepilot and explore three alternative approaches. The group included health economists, other health service researchers, and administrative staff. The question asked was designed to capture indi-

vidual's beliefs about the distribution of uncertainty surrounding the probability that London would host the 2012 Olympic games, before the announcement of the competition results. The unknown parameter of interest was the probability of hosting the Olympic games.

For all approaches, our colleagues were asked to provide the lowest (L) and highest (H) possible value of the probability of London hosting the Olympic games and the most likely value (M) as shown in Figure 1. If an inconsistency occurred, for example, M being higher than H, the software instantly informed the individual. Next, an illustration of these estimates was presented and if the individual felt that it failed to represent their beliefs they were encouraged to alter their initial values.

The participants were then asked to provide probabilities for the quantity lying within certain intervals: again, if inconsistencies occurred (e.g., probabilities not summing to one), Excel alerted the participants. Once all probabilities were imputed, individuals were shown a histogram derived from their estimates and asked whether it represented their beliefs. If the histogram failed to represent their beliefs, participants could easily clear the data and start again.

Three alternative methods for presenting intervals of the distribution were tested in this prepilot (see Fig. 1 and Appendix A for details of formulas used for calculating the intervals):

Six complementary intervals method. Based on the work of Phillips and Wisbey [10], our colleagues were presented with six complementary intervals. These

What is the probability that London will host the 2012 Olympic games?				
Think of a range of values from 0 to 100% to represent this value.				
What is the lowest likely value?	→	20	%	
What is the highest likely value?	→	60	%	
What is the most likely value?	→	50	%	

Overlapping intervals				
What is the probability of your estimated value lying in the following intervals?				
1. Between	20	and	50	→ <input type="text"/> %
2. Between	20	and	35	→ <input type="text"/> %
3. Between	55	and	60	→ <input type="text"/> %
4. Between	20	and	43	→ <input type="text"/> %
5. Between	53	and	60	→ <input type="text"/> %

Six complementary intervals				
What is the probability of your estimated value lying in the following intervals?				
1. Between	20	and	23	→ <input type="text"/> %
2. Between	23	and	40	→ <input type="text"/> %
3. Between	40	and	50	→ <input type="text"/> %
4. Between	50	and	53	→ <input type="text"/> %
5. Between	53	and	57	→ <input type="text"/> %
6. Between	57	and	60	→ <input type="text"/> %

Four complementary intervals				
What is the probability of your estimated value lying in the following intervals?				
1. Between	20	and	35	→ <input type="text"/> %
2. Between	35	and	50	→ <input type="text"/> %
3. Between	50	and	55	→ <input type="text"/> %
4. Between	55	and	60	→ <input type="text"/> %

Figure 1 Approaches tested to capture people's beliefs about the distribution of uncertainty.

were built automatically using a formula to divide the initial distance between each extreme value ([L] and [H]) and the most likely value (M) into three equal parts. Participants were asked to enter the probability that their estimated value was within each interval range.

Overlapping interval method. Based on the work of O'Hagan [21], this method uses intervals that overlap and are wider than the above method. This enables the estimation of six probabilities from the five elicited intervals.

Four complementary intervals method. This is a simplified version of the “six complementary intervals” method whereby only four complementary intervals were used.

Participants reported that with six complementary intervals the ranges were very narrow making the task extremely difficult. Similarly, with the “overlapping interval” method colleagues had considerable difficulty completing the task. Participants experienced confusion when assigning probabilities to intervals that overlapped, and reported that the resulting histogram rarely represented their beliefs. In contrast, participants felt that the “four complementary intervals” method was much easier to complete and more accurately represented their beliefs. This third approach was therefore adopted in our applied survey.

Applied Study Survey Design

We developed two questionnaires to elicit expert opinion on the likely values of eight parameters where

data were lacking in our decision-analytic model. The first questionnaire was for genetics experts (clinical geneticists and molecular genetic scientists) and comprised two questions about the accuracy of the genetic test; the second was for cardiologists and contained the questions on the natural history of disease and treatment effectiveness (see Appendix B for example questions). Each question asked the experts to assess only (hypothetical) observable quantities which they should be familiar with as, for example, the number of HCM patients detected by the cardiology services over a 5-year period. This approach avoided the need for the clinical experts to understand the decision model or to elicit moments of a distribution (e.g., variance, standard deviation, mean, etc.), which individuals have proven poor at [6,7]. All relevant covariates, such as the time frame and/or patient characteristics (e.g., age, HCM status), were explicitly reported in each question to help the experts understand what was being elicited.

The “four complementary interval” elicitation method was then employed to identify the distribution of uncertainty around each parameter. Each questionnaire was piloted with a clinical genetics colleague, whose comments were incorporated into the final questionnaires. Figure 2 shows an example of the user interface for one of the clinical questions.

In addition, in order to capture the basis of their beliefs, the experts were asked what their answer was based on (using a free text box). Such information was considered useful because it revealed the sources of evidence considered and how these were interpreted; it also encouraged experts to state their current level of knowledge.

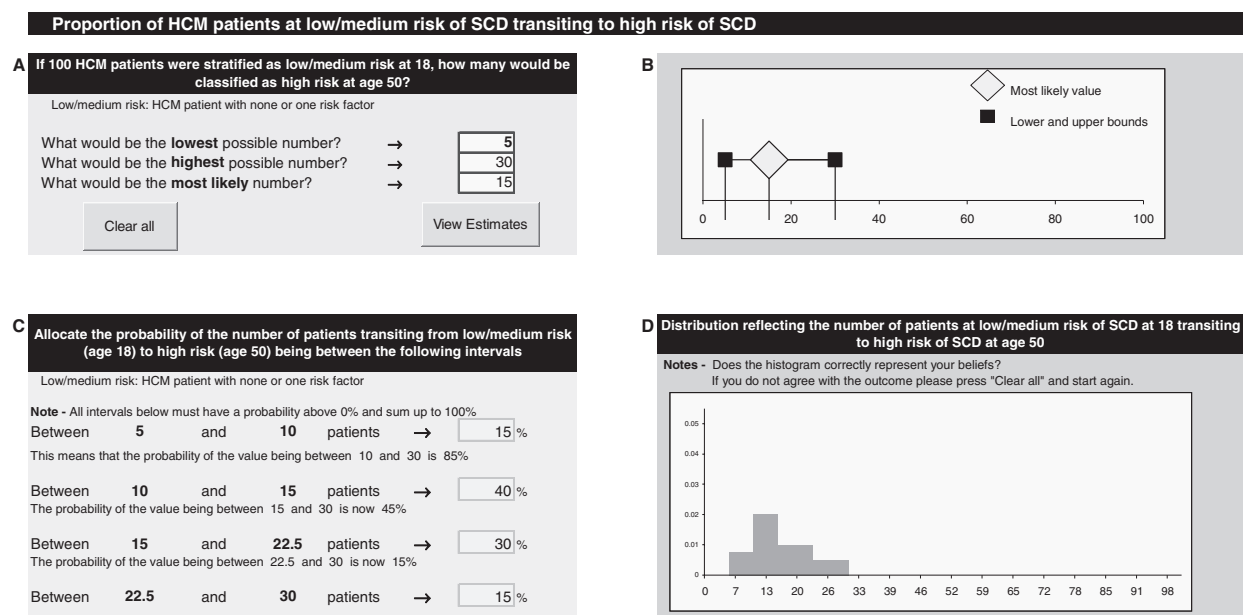


Figure 2 Example of survey question. HCM, hypertrophic cardiomyopathy; SCD, sudden cardiac death.

Identification and Selection of Experts

Because the prevalence of HCM and DNA technologies in the laboratories may differ across countries, only those UK-based experts were considered. We defined “expert” as someone that has specialist knowledge about the subject that we are interested in eliciting opinion about. Twelve leading experts working with HCM populations and DNA technology were identified by advice from our clinical colleagues in Oxford and invited to participate in the survey. The main selection characteristics were recognition by their peers, clinical experience with the subject matter, and being widely published. The experts were based in a mix of teaching and nonteaching institutions, with their specialties being in cardiology, genetics (or both), and molecular science. Six cardiologists and six clinical geneticists formed our expert sample, many of whom had international working experience. These experts were chosen from different areas of the UK to capture different patient populations and avoid eliciting opinions from people with very similar experiences (i.e., same group of patients).

Expert Elicitation

The experts were sent an email explaining the goal of the questionnaire, and how their input would help the construction of the HCM cost-effectiveness model (more information was made available if required). Attached to the email were instructions (in MS Word) for completing the questionnaire and a copy of the actual questionnaire (in Excel).

Respondents were asked to print the instructions before opening the questionnaire to enable them to read the instructions and complete the questionnaire simultaneously. Once the elicitation was completed, the experts were asked to save the file and return it to the authors (via email).

In order to help us gain an insight into how the experts were likely to interpret the questionnaire, we performed individual elicitation sessions with one cardiologist and two geneticists. The experts were briefed on the goals of the study, introduced to the questionnaire software, and we provided assistance by clearing up any misinterpretation of the questions, while being careful not to influence the expert’s results but reinforce our interest in their own opinions.

Once the questionnaires were returned, the experts were sent a feedback form asking about the ease and time necessary for questionnaire completion. The format and content of the questionnaire were classified on a scale of 1 to 5, where 1 was very easy and 5 was very difficult.

Combining Expert Opinions

For this study, it was necessary to combine the individual opinions of experts so that they could be used in

our genetic cost-effectiveness model. Their individual distributions were combined using the linear opinion pool method proposed by Stone [22]:

$$T(p_1, \dots, p_n) = \sum_{i=1}^n w_i p_i$$

where n is the number of experts, p_i is the expert i ’s probability distribution for the parameter of interest, the weights w_i are non-negative and sum to one, and $T(p_1, \dots, p_n)$ represents the summary of the p_i ’s.

For our study, it was perceived that there was little justification for applying different weights to the different experts: as DNA testing in HCM is still a very new clinical area, our experts had similar exposure to information on the testing and we had no further evidence to assign different weights. Therefore, the experts were considered to be equal and the linear opinion pool became a simple arithmetic average.

Fitting Smooth Functions to the Histograms

As the purpose of this exercise was to obtain the opinion of experts concerning the uncertainty in the required model parameters, we could have directly imputed the combined histograms in the model and performed PSA. Nevertheless, the nature of the histogram leads to the same probability being given to all values in a certain interval. Hence, a smooth distribution was considered to be a more realistic way of representing the expert’s opinions, as it allows different probabilities for each possible point estimate and avoids abrupt variations from one point to another [21]. Appropriate parametric distributions were then chosen [2] and assigned using maximum likelihood fitting [23]. Their goodness of fit was evaluated by drawing the probability density function curve and histogram together. Finally, these distributions were introduced in our decision-analytic model, and the uncertainty around the results was quantified.

Results

Response Rate and Feedback

Seven out of 12 experts (58%) returned the survey, with six managing to complete all questions and provide adequate answers (50%). That is, all probabilities for a given parameter added to one, no expert reported disagreeing with the numerical and graphical feedback of how their statements looked and logical responses to questions were provided such as individuals at high risk of SCD being reported as having at least an equal if not higher probability of being detected compared with an equivalent number of low-/medium-risk individuals. Three cardiologists and three geneticists completed the survey. The seventh expert reported being unable to use the software, and did not complete the questionnaire.

Four feedback forms were returned and the format of the elicitation tool was reported to be easy to use

and take less than an hour to complete (cardiology questionnaire required more time than the genetics questionnaire). When asked about how easy it was to complete the questions, most experts reported that they found some of the questions fairly difficult to answer. Nevertheless, they all stated that this was due to the uncertainty in the clinical area, rather than how the questions were asked.

Elicited Parameters

In total, data were obtained on eight parameters using these surveys. As the results were similar across parameters, here we present the elicitation results for one single parameter as an illustration (information on the results from other parameters can be obtained on request from the authors). For this parameter, the proportion of HCM population at low/medium risk of SCD (Fig. 3), the experts provided a variety of different values for the same question. This may reflect some degree of complementary beliefs, suggesting that their opinions arise from different experiences. The experts reported that the basis for their estimates was either data from their own center, literature interpreted in the light of their own clinical experience, or simply clinical experience.

The histograms resulting from each expert elicitation were combined using the simple arithmetic average, and are presented in Figure 3 (under “Combined”). The combined histogram suggested that a beta distribution would represent the data best. As such, a fitted beta distribution together with the original combined histogram is presented in Figure 4.

The elicited parameters were introduced in the decision-analytic model, and the cost-effectiveness of the different model strategies for diagnosing HCM was estimated together with the uncertainty around it.

Discussion

In economic evaluation, decision-analytic models are increasingly used to examine the cost-effectiveness of health-care interventions. A lack of readily available observed evidence has meant that expert opinion is commonly used in these models. Analysts often simply elicit expert opinion on the mean or median values for the parameter of interest and sometimes the minimum and maximum values. Potentially, parametric distributions could be fitted to these values and assumed to represent the expert’s beliefs. Nevertheless, a key limitation of this approach is that it provides insufficient data about the expert’s belief to examine whether the distribution is appropriate. In addition, the increasing use of PSA in decision models requires the correct representation of the uncertainty in model inputs. Hence, the rationale for our work was the need to be explicit and structured about how expert opinion was obtained to inform a probabilistic model.

Proportion of HCM patients at low/medium risk of SCD

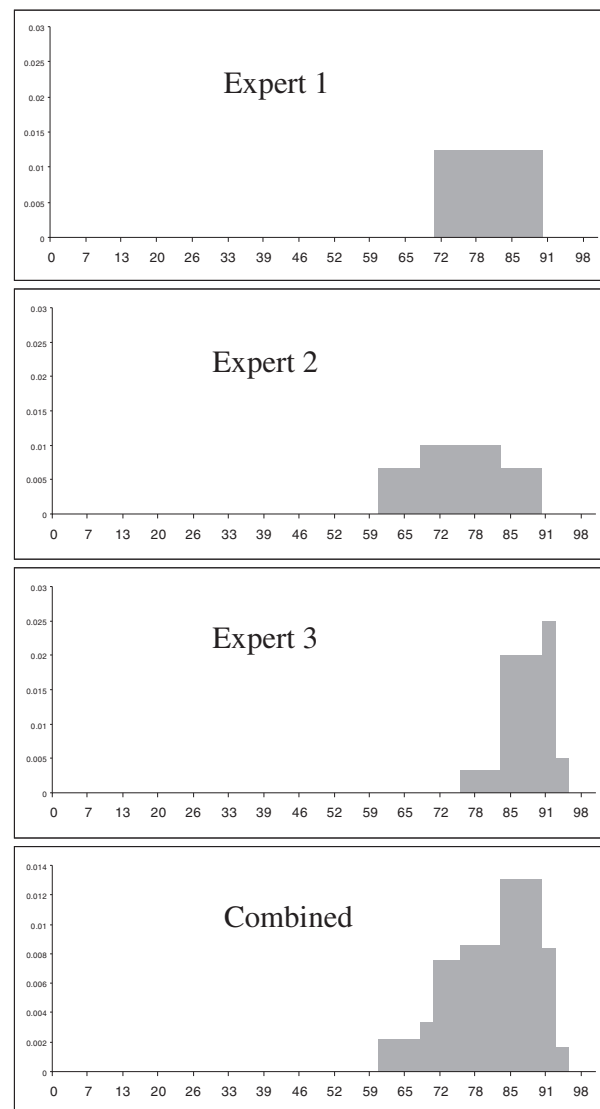


Figure 3 Display of elicited experts’ beliefs about one model parameter. HCM, hypertrophic cardiomyopathy; SCD, sudden cardiac death.

Relevant reviews have been published about the elicitation of experts’ beliefs [6–8] as well as some applied work in other areas than health-care decision modeling [9–14]. Nevertheless, eliciting opinions in health care remains a theoretical and practical challenge that requires further research.

This article has presented the results of a methodological study, which designed a simple questionnaire-based tool using an Excel spreadsheet for eliciting experts’ opinions. The software included graphical feedback so that the experts could immediately see the distributions defined by their estimates and make instant refinements if required.

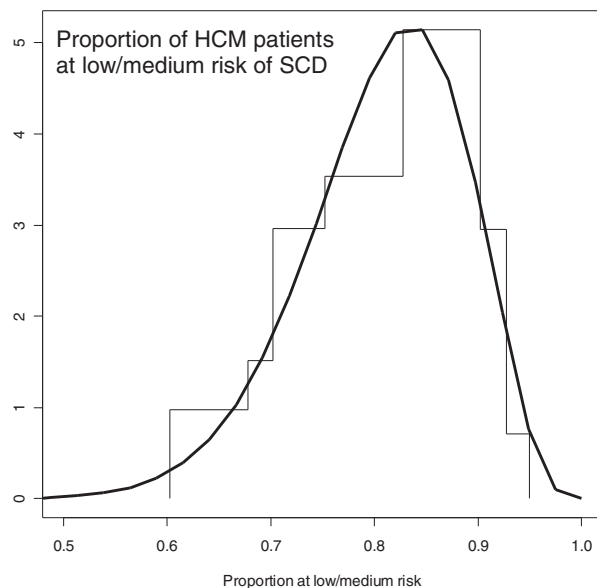


Figure 4 Fitting distributions to combined elicited beliefs. HCM, hypertrophic cardiomyopathy; SCD, sudden cardiac death.

The survey achieved a reasonable response rate, logical answers, and was generally rated as easy to use and all respondents completed the survey in well under an hour. The respondents found some of the questions difficult to answer. Nevertheless, this was attributed to difficulties in answering the clinical question itself as opposed to difficulties with using the survey instrument.

The parameters we were interested in eliciting expert opinion about involved a Bernoulli process (i.e., binary outcomes), which is very useful to model probabilities, risks, prevalence, and odds of health events. Nevertheless, our method could potentially be used to describe the uncertainty about parameters defined by other processes, such as Poisson/gamma processes for resource use/cost data, as we are eliciting the distribution of (hypothetical) observations and placing no parametric constraints on neither the elicitation procedure nor the results.

In terms of the potential limitations of this work, in some ways our approach was crude in respect to the distribution that the experts had to describe only being divided into four intervals. In the literature, other authors using the “quantile method” have elicited extra probabilities or quantiles (up to 12), either using consecutive intervals or overlapping intervals [10,21]. Nevertheless, our prepiloting (Olympic games question for colleagues) highlighted that respondents found both of these methods very time-consuming and difficult to undertake; therefore, we reduced the intervals to four, as we anticipated that the clinical questions we intended to ask would be more complicated than this simple pre-pilot question and hence more difficult to answer.

Undoubtedly, there is a trade-off between the complexity of the elicitation technique and practicality

for obtaining responses that truly reflect the expert’s beliefs. This discrepancy between their “true” beliefs and their elicited distributions is influenced by the way in which experts assess the information provided and how their answers may be affected by likely sources of bias [6,8,24,25]. In this article, we have only considered quantile methods because they are thought to perform better than other methods [6,8,20]. Nevertheless, further research is required to accurately determine which elicitation methods (and variations) may perform better in the context of health. In addition, we do not suggest that the elicited distributions are neither the only correct representation of what the expert actually believes or that their judgment is error free. However, others perceive that if the expert is able to use the survey and believes that the graphical feedback reflects a satisfactory approximation of his beliefs, then this is a useful method [7].

A further potential limitation of our study was that we only report results from three experts per questionnaire. This is partly because the number of individuals with relevant expertise in this genetic area is very limited. Nevertheless, we contacted individuals working with different HCM populations to avoid eliciting highly correlated opinions (i.e., informed experts who do not differ in the way they may think about the question) and to be able to average out any consistent tendency of a particular expert to overestimate or underestimate their opinion (i.e., bias). Also, such small numbers may not be too problematic because it has been suggested in the literature that it is not necessary to have a large number of experts because of diminishing marginal returns associated with large number of experts, and most benefit is suggested to occur with the first three to four experts [5,26]. Hence, in any field the main focus should be on selecting and contacting experts with different opinions/experiences rather than on seeking large number of individuals with the same/similar opinions [27]. Further work is required to test appropriate sample sizes in health care.

We would also have preferred for one of the research team to have been present during the elicitation to provide experts with software training and clarify any issues. Unfortunately, in this case study, this was not possible to do with all the experts because of time and geographic constraints. Instead, some experts were sent the questionnaire via email. While those that did respond to our survey managed to complete the questionnaire, a number of experts failed to respond.

Potentially, we could have used some form of weighting to reflect different knowledge levels between experts. Indeed, techniques for weighting and calibrating the opinions of different experts are available in the literature [8]. Nevertheless, in the absence of any relevant empiric data known to us (but unknown to the expert) and because the DNA test being examined

was so new, we felt that there was no clear justification for weighting our experts differently.

There are a number of more complex mathematical methods than the linear opinion pool for combining opinions [5,8,28]. Nevertheless, it is not clear from the literature whether more complex approaches actually perform better in practice than simpler approaches [5].

In this article, we have used individual elicitation methods in line with recent guidelines for health-care decision modeling [4]. An alternative would have been to have used forced consensus methods (e.g., Delphi, Nominal Group Technique, etc.). Nevertheless, a recent review concluded that during the aggregation of expert opinion there are no clear benefits of interaction (i.e., forced consensus) over no interaction (i.e., individual elicitation methods) despite its intuitive appeal [5]. Hence, there is a lack of strong support for consensus methods, and issues such as the need for extensive facilitation during the elicitation to avoid group polarization [29], the difficulty of convening experts from different parts of the country at a time and place suitable for all, and the reality that a distribution obtained through consensus is still based on individuals' beliefs, may not justify the effort and cost over eliciting experts' opinions individually. Nevertheless, there may be additional benefits from the exchange of information between experts, before the individual elicitation itself, about potential sources of evidence to be considered and the definitions/assumptions of the questions posed to them.

In terms of further research, attention needs to be devoted to formally comparing our elicitation tool with other complex approaches of eliciting experts' beliefs. This is a challenge in itself because to determine the accuracy of different approaches, we need to compare the elicited beliefs with the "true" beliefs of the individual, for which methods are yet to be developed to do. Nonetheless, there is scope to explore different ways of eliciting distributions from individual experts and to evaluate the additional benefit of combining these [26]. Furthermore, our experience in this project has highlighted that more formal qualitative assessment of our survey could have been useful. As such, we are incorporating a debriefing aspect into our ongoing research using this elicitation tool and explicitly asking nonresponders for their reasons for not responding. Finally, we are focusing on the identification and development of methods to further test both the validity and reliability of the elicitation tool, as well as evaluating ways of incorporating a histogram smoothing function into the tool.

Conclusion

Health economists often rely on eliciting expert opinion when populating decision-analytic models. Increasingly, we also need to understand the distri-

bution of all parameters to conduct probabilistic sensitivity analyses. The unstructured and implicit approach to eliciting expert opinion in decision models fails to allow the uncertainty in evidence to be correctly and fully represented. As such, this article provides a practical way forward in eliciting expert opinion and highlights some of the challenges in following approaches suggested in the theoretical literature. Additional research is required to further test and refine this simple elicitation tool. Nevertheless, the tool goes some way toward providing a bridge between the theoretical literature and what is practical for use in health-care decision models and will hopefully be of use to other researchers facing the need to use expert opinion in their models.

We thank our colleagues in the Department of Public Health, University of Oxford for comments on the elicitation tool. We also thank the experts whose contribution was invaluable to this project and the comments from three anonymous reviewers. Funding is acknowledged from the Oxford Genetics Knowledge Park (UK Department of Health). Any views expressed in this article are entirely those of the authors.

References

- 1 Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technol Assess* 1999;3:13–21.
- 2 Briggs AH. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 2000;17:479–500.
- 3 NICE. Guide to the Methods of Technology Appraisal. N0515. London: National Institute for Clinical Excellence, 2004.
- 4 Philips Z, Ginnelly L, Sculpher M, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004;8:19–29.
- 5 Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Anal* 1999;19:187–203.
- 6 Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. *J Am Stat Assoc* 2005;100:680–701.
- 7 Kadane JB, Wolfson LJ. Experiences in elicitation (with discussion). *J R Stat Soc* 1998;47:3–19.
- 8 Cooke RM. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford University Press, 1991.
- 9 Cooke RM, Goossens LHJ. Procedures guide for structured expert judgment in accident consequence modelling. *Radiat Prot Dosimetry* 2000;90:303–9.
- 10 Phillips LD, Wisbey SJ. 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/B101. Didcot, UK: Nirex, 1993.
- 11 Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. *Statistician* 1993;42:341–53.

- 12 Garthwaite PH, O'Hagan A. Quantifying expert opinion in the UK water industry: an experimental study. *Statistician* 2000;49:455–77.
- 13 Ramachandran G, Banerjee S, Vincent JH. Expert judgment and occupational hygiene: application to aerosol speciation in the nickel primary production industry. *Ann Occup Hyg* 2003;47:461–75.
- 14 Van der Fels-Klerx HJ, Cooke RM, Nauta MN, et al. A structured expert judgment study for a model of campylobacter transmission during broiler-chicken processing. *Risk Anal* 2005;25:109–24.
- 15 Insinga RP, Laessig RH, Hoffman GL. Newborn screening with tandem mass spectrometry: examining its cost-effectiveness in the Wisconsin Newborn Screening Panel. *J Pediatr* 2002;141:524–31.
- 16 Cooper N, Coyle D, Abrams K, et al. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *J Health Serv Res Policy* 2005;10:245–50.
- 17 Elliott P, McKenna WJ. Hypertrophic cardiomyopathy. *Lancet* 2004;363:1881–91.
- 18 McKenna WJ, Behr ER. Hypertrophic cardiomyopathy: management, risk stratification, and prevention of sudden death. *Heart* 2002;87:169–76.
- 19 Drummond MF, Sculpher MJ, Torrance GW, et al. *Methods for the Economic Evaluation of Health Care Programmes* (3rd ed.). Oxford: Oxford University Press, 2005.
- 20 Winkler RL. The assessment of prior distributions in Bayesian analysis. *J Am Stat Assoc* 1967;62:776–800.
- 21 O'Hagan A. Eliciting expert beliefs in substantial practical applications. *J R Stat Soc* 1998;47:21–35.
- 22 Stone M. The opinion pool. *Ann Math Stat* 1961;32:1339–42.
- 23 Venables WN, Ripley BD. *Modern Applied Statistics with S: Fourth Edition*. New York: Springer-Verlag, 2002.
- 24 Lindley DV, Tversky A, Brown RV. On the reconciliation of probability assessments (with discussion). *J R Stat Soc* 1979;142:146–80.
- 25 Wallsten TS, Budescu DV. Encoding subjective probabilities: a psychological and psychometric review. *Manag Sci* 1983;29:151–73.
- 26 Winkler RL, Clemen RT. Multiple experts vs. Multiple methods: combining correlation assessments. *Decis Anal* 2004;1:167–76.
- 27 Soll JB. Intuitive theories of information: beliefs about the value of redundancy. *Cognit Psychol* 1999;38:317–46.
- 28 Genest C, Zideck JV. Combining probability distributions. A critique and annotated bibliography. *Stat Sci* 1986;1:114–48.
- 29 Plous S. *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill, 1993.

Appendix A

During the prepiloting of the elicitation tool, our colleagues were first asked to provide the lowest (L) and highest (H) possible value of the probability of London hosting the Olympic games, as well as the most likely

value (M). They were then asked to provide probabilities for the quantity lying within certain intervals. We tested three alternative methods for presenting the intervals.

In the “six complementary interval” method, the individual was asked to provide probabilities for the following six intervals:

1. $[L, (L + M)/3]$
2. $[(L + M)/3, (L + 2M)/3]$
3. $[(L + 2M)/3, M]$
4. $[M, (2M + H)/3]$
5. $[(2M + H)/3, (M + 2H)/3]$
6. $[(M + 2H)/3, H]$

In the “overlapping interval” method, the intervals were constructed according to the following formula:

1. $[L, M]$
2. $[L, (L + M)/2]$
3. $[(M + H)/2, H]$
4. $[L, (L + 3M)/4]$
5. $[(3M + U)/4, U]$

Finally, in the “four complementary interval” method, the individual was asked to provide probabilities for the following intervals:

1. $[L, (L + M)/2]$
2. $[(L + M)/2, M]$
3. $[M, (M + H)/2]$
4. $[(M + H)/2, H]$

Appendix B

We present below the information provided to the experts regarding the elicitation question about the proportion of HCM population at low/medium risk of SCD.

Aim: Establish the proportion of HCM patients at low/medium risk of SCD in the whole HCM population.

Background: The predictive clinical features of high-risk patients for SCD are the following:

- Family history of multiple SCD;
- Unexplained syncope;
- Flat or hypotensive blood pressure response during upright exercise;
- Nonsustained ventricular tachycardia during Holter monitoring;
- Severe hypertrophy (>30 mm).

A HCM patient may be considered at high risk of SCD with two or more of the above risk factors, at medium risk with one risk factor, and at low risk with no risk factor.

Question: Out of 100 HCM patients, how many would be classified as low/medium risk of SCD?