Eliciting expert opinion for economic models: Can we do better?

José Leal, Sarah Wordsworth, Rosa Legood and Edward Blair

1. Health Economics Research Centre, Department of Public Health, Oxford University
2. Oxford Radcliffe Hospitals NHS Trust

Paper presented at the Health Economist's Study Group Meeting, Newcastle, June 2005

Correspondence: Jose Leal, Health Economics Research Centre, Department of Public Health, University of Oxford, Old Road Campus, Oxford, OX3 7LF. E-mail: jose.leal@dphpc.ox.ac.uk. Tel: + 44 (0)1865 226690; fax: + 44 (0)1865 226842.

Work in progress; please do not quote without permission

Abstract:

Expert opinion is considered a legitimate source of information where required data are unavailable. Decision-analytic modelling guidelines recommend that the expert elicitation process should be clearly documented, although provide limited advice on the appropriate elicitation methods to employ. This paper presents a methodological study which aimed to develop a practical computer based tool for eliciting expert opinion and produce information on the distributional forms of model parameters.

In this study our applied clinical example was hypertrophic cardiomyopathy (HCM), a relatively common genetic disease and the main cause of sudden cardiac death in the young. The cost-effectiveness model explores the introduction of DNA testing within HCM families, which is a new clinical area lacking data on some key parameters. A computer software interface was first developed and piloted with departmental colleagues. It was then used to obtain model parameters from individual HCM experts (as opposed to Delphi techniques) relating to: 1) natural history of disease and treatment effectiveness; and 2) accuracy of DNA testing. The *quantile method* was used to quantify subjective opinions to avoid asking individuals directly about parameter variances. The interface provided graphical feedback throughout the elicitation process and internal consistency was tested. A small sample of UK cardiologists and geneticists participated in the study.

A smooth distribution has been fitted to the expert summaries. These distributions are being mathematically aggregated and incorporated into the model, enabling the use of probabilistic sensitivity analysis. It is anticipated that this study, once completed, will be of use to other health economists deriving expert opinion for their economic models.

1. Introduction

As the number of economic evaluations using decision analytic models to synthesise cost and effect data increases, the issue of obtaining robust data on model parameters becomes important. Frequently, not all the data can be gathered from observed evidence (RCT, cohort studies etc); therefore subjective information from experts (usually clinicians) is required. To date most health economists using expert opinion have tended to ask for mid point estimates from local clinicians or clinicians involved directly with the economic evaluation being performed. Such an approach may be appropriate in evaluations where the parameters considered are unlikely to drive the cost-effectiveness results, and/or some reasonable information is already known. However, in new clinical areas, such as genetics this could be insufficient.

Recent decision modelling guidelines recommend that the use of expert opinion should be clearly documented (Philips et al. 2004) and suggest that methods aiming to reach a consensus (e.g. Delphi approaches) are inappropriate, as they can fail to capture the true uncertainty in the parameters. This is particularly important, in the context of recent guidelines from the National Institute of Health and Clinical Excellence (NICE) recommending that uncertainty in the results of modelling studies should be assessed using probabilistic sensitivity analysis (PSA), (NICE 2004) which permits the uncertainty in individual parameters to be propagated across the model simultaneously to explore the overall uncertainty in cost-effectiveness results (Briggs 2000). Performing PSA requires assigning distributions to model parameters, with these distributions representing the range of values that a given parameter may take.

Unfortunately, current guidance provides limited advice on the most appropriate elicitation methods to employ in modelling. This paper presents a methodological study which aimed to develop a practical computer based tool for eliciting expert opinion and produce information on the distributional forms of model parameters applied to an economic evaluation of a new genetic test. Rather than providing an extensive review of the various elicitation methods available (see Garthwaite, Kadane and O'Hagan (2004),

Kadane and Wolfson (1998) and Cooke (1991)), or comparing the results of different methods of elicitation, this paper reports the pragmatic process we undertook to eliciting expert opinion in our applied example.

2. Background

Genetics research, particularly DNA testing, shows great potential for identifying those at increased risk of a cardiovascular event (before illness). However, to translate genetic research into routine clinical practice, information is required on the potential costs and benefits of DNA testing in cardiology. Examples of the potential use of DNA testing in cardiology include inherited causes of sudden cardiac death (SCD), such as hypertrophic cardiomyopathy (HCM). HCM for example, has been shown to be a relatively common genetic condition with a disease prevalence of 1/500 (Elliott & McKenna 2004) defined by unexplained asymmetric thickening of the heart and most commonly inherited in an autosomal dominant manner (requires only one parent to have an abnormal gene in order for the child to inherit the disease) .

In addition, the Oxford Genetics Knowledge Park is performing DNA testing for patients and their families (by cascade testing) who are at risk of SCD on a pilot basis. This pilot programme is underway within an NHS laboratory to examine the potential of DNA testing to improve clinical care by increasing the level of certainty in diagnosis, and through the predictive screening of at-risk family members.

The economic evaluation being performed alongside this pilot programme is assessing the long-term costs and effects of alternative approaches to diagnosing and managing HCM for those at risk of SCD. To date we have built a Markov model where the costs (clinical tests, DNA testing, genetic counselling and treatment) and effects (life years gained) of a genetic and non-genetic (mainly clinical) approach to HCM within families are compared.

As this is a relatively new clinical area, there are a number of parameters that cannot or have not been measured in practice. In particular, information on the sensitivity and

specificity of the new DNA test for HCM was limited, as to date only small numbers of patients have undergone testing. Furthermore, there are no reliable studies published concerning the effectiveness of treatments for the primary prevention of SCD for those with HCM. Hence, in the absence of robust observational data, we were faced with eliciting distributions from experts for several these parameters. The 'expert' here is defined as someone with specialist knowledge of inherited cardiac condition, especially HCM.

2. Empirical Study Methods

The elicitation process is defined as: the development of the elicitation tool, the elicitation of experts' opinions, and the analysis of results. In this section of the paper, the various steps of the process are described. Firstly, we outline the variables where expert opinion was required. Secondly, we describe the design and piloting of the elicitation tool, where the different phases of development are examined. Thirdly, the identification and selection of experts is addressed. Finally, we present the elicitation of experts' opinions and the technique used for combining the individual results.

2.1. Identification of elicitation variables

The nine model parameters in our HCM requiring expert opinion were the:

- Proportion of HCM population at high risk of SCD ;
- 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 ICD in the prevention of SCD;
- Effectiveness of amiodarone in the prevention of SCD;
- Sensitivity of the genetic diagnostic test;
- Specificity of the genetic diagnostic test.

These parameters gave rise to eight specific questions, some of which were tailored to cardiology and others to genetic experts (clinical genetics and molecular genetic scientists). For instance, in order to identify the uncertainty around the proportion of HCM population at high risk of SCD, we posed the question "Out of 100 HCM patients, how many would be classified as low/medium risk of SCD?"

2.2. Design and piloting of elicitation tool

In line with Cooke (1991), an important objective of our methodological study was to develop an elicitation tool that was clear, attractive and could be completed within one hour of the expert's time. A further objective was to provide a framework that could interact with the user in an intelligent way and provide graphical feedback. Therefore, our solution was to build a questionnaire in Microsoft Office Excel. This software package was chosen due to its widespread use and flexibility to perform calculations and produce graphs.

The first phase in developing our elicitation tool was to undertake a pre-pilot where we tested a dummy question on fellow colleagues within the Department of Public Health (University of Oxford). The question was designed to capture individual's beliefs about the distribution of uncertainty surrounding the probability that London would host the 2012 Olympic Games.

Based on previous work by Phillips and Wisbey (1993) and O'Hagan (1998), three approaches using *quantile* methods were tested for eliciting distributions of our Olympic Games question. As shown in Figure 1, for all three approaches, our colleagues were asked to provide the lowest and highest possible value of the probability of London hosting the Games, as well as the most likely value.

More specifically, based on Phillips and Wisbey (1993), our colleagues were presented with *six complementary* intervals and asked about the probability of the desired quantity being between the limits of each interval. This approach resulted in narrow intervals, and our sample found it difficult to estimate such small probabilities. O'Hagan (1998) had

also experienced this problem previously and used wider but *overlapping* intervals in order to obtain the six probabilities.

Secondly, wider *overlapping* intervals were shown to the experts who were asked to provide a probability for each. Unfortunately, respondents still found it difficult and confusing to attribute probabilities to overlapping intervals, and the elicitation outcome rarely represented their beliefs.

Finally, we revised the questionnaire and *four complementary* intervals were now presented to our sample who ranked this approach the highest amongst the three. In particular, they reported that this third approach was the easiest and most practical, while still representing their beliefs to some degree of accuracy. This third approach was therefore adopted for our main study eliciting expert opinion for a new DNA test for HCM.

FIGURE 1 HERE

For the main methodological study an example of our elicitation tool is shown in Figure 2. Akin to the method applied by O'Hagan (1998), in our survey the experts were to be asked to provide a high and low limit, H and L respectively, of the elicited quantity, so that no amount above H or below L would be likely to be found. Then, the expert was asked to provide the most likely value M of the desired parameter (step 'A'). If an inconsistency occurred, e.g. M being higher than H, the software instantly informed the expert.

FIGURE 2 HERE

The next step (step B) was to present an illustration of these estimates. If the expert felt that it did not represent their beliefs they were encouraged to alter the initial values.

Thirdly, in 'step C' the expert was asked to provide probabilities for the quantity lying in the following *four* intervals:

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

Again, if inconsistencies occurred (e.g. probabilities not summing to one), the software alerted the respondent to this. Once all probabilities were imputed, the expert was shown a histogram derived using their estimates (step 'D'), and asked whether it represented their beliefs. In the event that the experts felt that the histogram failed to represent their beliefs, they could easily clear the data and restart the process.

In addition, in order to capture the basis of their beliefs, the experts were also asked what their answer was based upon. Asking for such information was considered useful because it revealed whether answers were based on chance or experience, and also hopefully encouraged experts to state their true opinions.

The questions were piloted with a clinical colleague involved in our economic evaluation, whose comments were incorporated into the final questionnaires. Once we had designed the elicitation process we developed two questionnaires to elicit data for the parameters required for our HCM model. The first was for the genetics experts (clinical geneticists and molecular genetic scientists) and comprised the two questions about the accuracy of the genetic test; the second was for the cardiologists and contained the remaining questions.

2.3. Identification and selection of experts

A list of leading experts working with HCM populations and those at risk of sudden cardiac death generally was generated from the literature, and by advice from our clinical colleagues in Oxford. Since the prevalence of HCM and DNA technologies in the laboratories may differ across countries, only UK based experts were considered. The reputation, area of interest and published work were viewed as proxies for the experts'

scientific expertise. Experts from different areas of the UK were sought in order to capture different patient populations and avoid eliciting opinions from people with very similar experiences (e.g. same families). A total of twelve experts were selected and invited to participate in the DNA testing for HCM survey.

2.4. Expert elicitation

The experts received an e-mail 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 e-mail were instructions (in Word) for completing the questionnaire and a copy of the actual questionnaire (in Excel).

Respondents were asked to print the instructions prior to 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.

In order to help us gain an insight into how the experts were likely to interpret the questionnaire, we performed individual elicitation (think aloud) sessions with one cardiologist and two genetics experts. The experts were briefed on the goals of the study, introduced to the questionnaire software, and provided assistance if needed in their probabilistic assessments. Our role was to clear up potential misinterpretation of the questions, and reinforce the interest in the expert's own opinions.

Once the questionnaires were returned, the experts were sent a feedback form asking about the ease and time necessary for the completion of the questionnaire. 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.

2.5. Combining expert opinions

Although though there may be situations where single experts' probability distributions are sufficient, in our methodological study it was necessary to combine them into a single distribution. Hence, the experts' individual distributions were combined using the *linear opinion pool* method proposed by Stone (1961):

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

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 $T(p_1,...,p_n)$ represents the summary of the p_i's.

In terms of differential weightings for experts, 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, and our experts had similar exposure to information on the testing. Therefore, the experts were considered to be equal and the *linear opinion pool* became a simple arithmetic average.

2.6. Fitting smooth functions to the histograms

Parametric distributions were fitted to the combined elicited histograms because a smooth density function was considered a more realistic way of representing the experts' opinions (O'Hagan 1998), and a continuous function can represent the uncertainty around the model parameters (Briggs 2000). Elicited distributions were chosen that fitted best the elicited frequencies and the range of data.

3. Results

3.1. Response rate and feedback

To date 7/12 experts (58%) have returned the survey, with six managing to complete all questions and provide coherent answers. That is, all probabilities for a given parameter added to one, and logical responses to questions were provided such as, e.g. people at high risk of SCD were reported to have at least an equal if not higher probability of being

detected compared to an equivalent number of low/medium risk individuals. The seventh expert reported being unable to use the software, and hence unable to complete the questionnaire.

Of the four feedback forms we have received so far, the format of the elicitation tool was reported to be easy to use and took less than an hour to complete (cardiology questionnaire required more time than the genetics questionnaire). However, in terms of ease of completion for the questions themselves, most reported finding some of the questions fairly difficult.

3.2. Elicited parameters

As there are several questions in the survey, here we present a selection. Figure 3, presents the elicitation results for two parameters, namely: (Question.A) the proportion of HCM population at low/medium risk of SCD, and (Question.B) the transition from low/medium to high risk of SCD over a patient's lifetime. As can be seen, the experts provided a variety of point estimates and probabilities for the same question, which may reflect some degree of complementary of beliefs. The experts reported that the basis for their estimates of parameter A to be either literature interpreted in the light of their own clinical experience, or simply clinical experience. They then reported that the estimates for parameter B were based on best guess estimates, intuition and own impressions from clinical experience.

FIGURE 3 HERE

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

FIGURE 4 HERE

4. 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. Often analysts only elicit expert opinion on the mean or median values for the parameter of interest and sometimes the minimum and maximum values. Specific distributions are then fitted to these values and are said to represent the expert's beliefs. A limitation of this approach is that it provides insufficient data about the expert's belief to examine whether the elicited distribution is appropriate.

This paper has presented the results of a methodological study, which designed a simple tool using an Excel spreadsheet for eliciting expert's opinion and provided an insight into the 'true' distribution of the required parameters. The software included graphical feedback, so that the experts could immediately see the distributions defined by their estimates and make instant refinements if required.

The survey achieved: a reasonable response rate, coherent answers, and was generally rated as easy to use. The respondents found some of the questions difficult to answer and the cardiology questionnaire required more time than the genetics questionnaire to be completed. However, all respondents completed the survey in less than an hour.

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 six), either using consecutive intervals or overlapping intervals (O'Hagan 1998). However, our pre-piloting work (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 robust responses that truly reflect the expert's beliefs. This robustness is influenced by the way experts assess the information provided and how their answers may be affected by likely sources of bias (Cooke 1991;Garthwaite, Kadane, & O'Hagan 2005). Research seems to show that experts are able to estimate modes and proportions of samples, and are reasonable at quantifying beliefs as intervals of values. Hence, in this paper we have only considered *quantile* methods because they are thought to be more robust than other methods, e.g. hypothetical future sample methods (Garthwaite, Kadane, & O'Hagan 2005). However, appropriate empirical research has still to be conducted in order to accurately determine which methods (and variations) actually work.

A 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 HCM is not high and that we are still waiting for some responses. However, 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 due to diminishing marginal returns (Clemen & Winkler 1999).

A further limitation is that several respondents found the questions themselves difficult to answer, although the feedback from the experts highlighted that this was related to the complexity of the disease and the DNA test itself rather than the questionnaire format.

Ideally, it would have been preferable if an assessor could 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 due to time and geographical constraints. Instead some experts were sent the questionnaire via e-mail. Whilst those that did respond to our survey managed to complete the questionnaire and provided coherent answers, a number of experts failed to respond and one reported that they were unable to use the software.

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 (Cooke 1991). However, in the absence of any relevant empirical 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: hence equal weights were applied for all our experts.

There are a number of more complex mathematical methods than the linear opinion pool for combining opinions (see Cooke (1991), Clemen and Winkler (1999) and Genest and Zidek (1986) for a review). However, it is not clear from the literature whether these more complex approaches actually perform any better in practice than simpler approaches (Clemen and Winkler 1999).

The next stage in our research is to complete the data collection exercise and to integrate the results into our cost effectiveness analysis. At the moment we have fitted a distribution to the combined histogram using the maximum likelihood estimation method. We are also planning to explore alternative methods of quantifying the uncertainty in the elicited distribution such as sampling from the actual histogram data.

In conclusion, health economists often have to rely on expert opinion when populating decision analytic models and they also need to understand the distribution of all parameters in order to conduct fully probabilistic sensitivity analyses. However crude the technique used may seem we are faced with few alternatives. There is clearly a gap in terms of literature between theoretical elicitation techniques and tools that can be used in applied economic evaluation. Our future research is exploring whether with further development and testing, this simple elicitation tool could provide a practical robust solution for other health economists deriving expert opinion for their economic models.

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

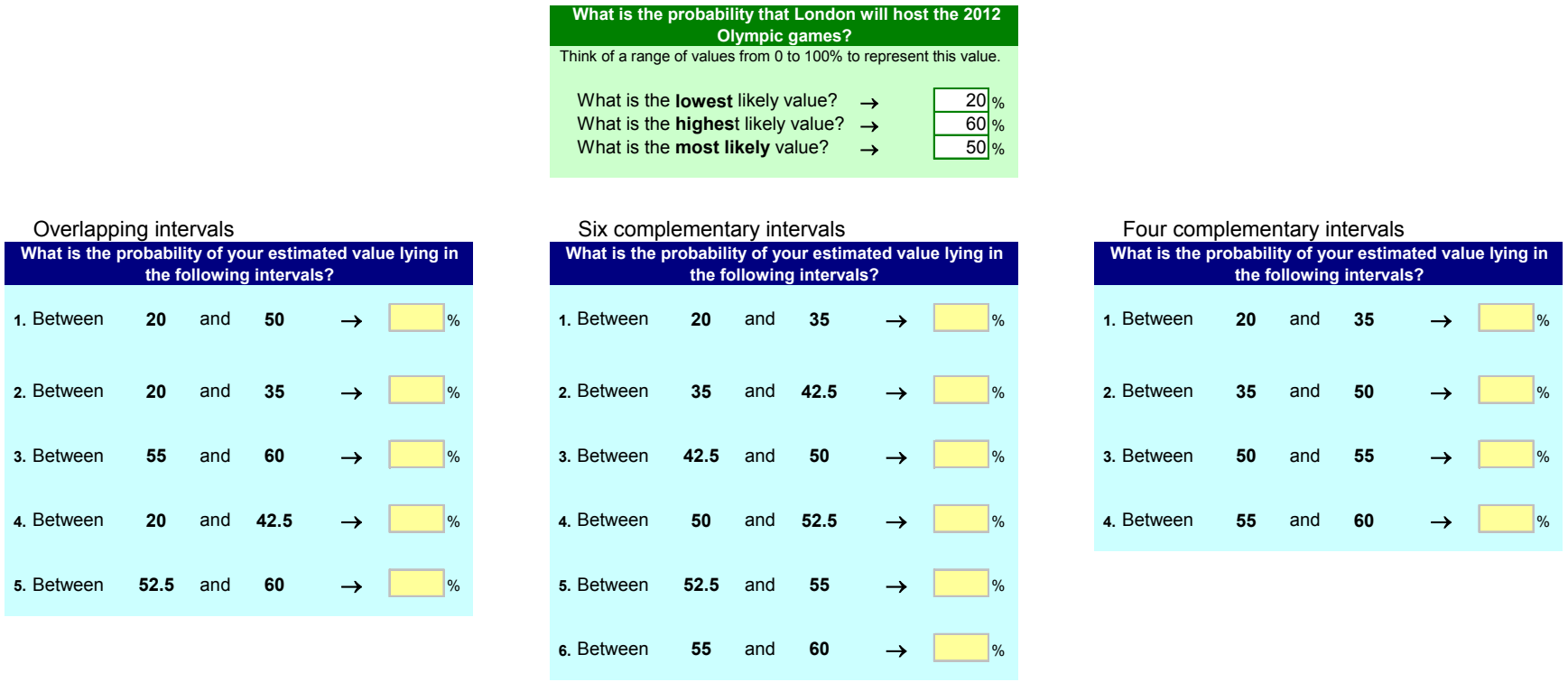


Figure 2. Example of survey question.

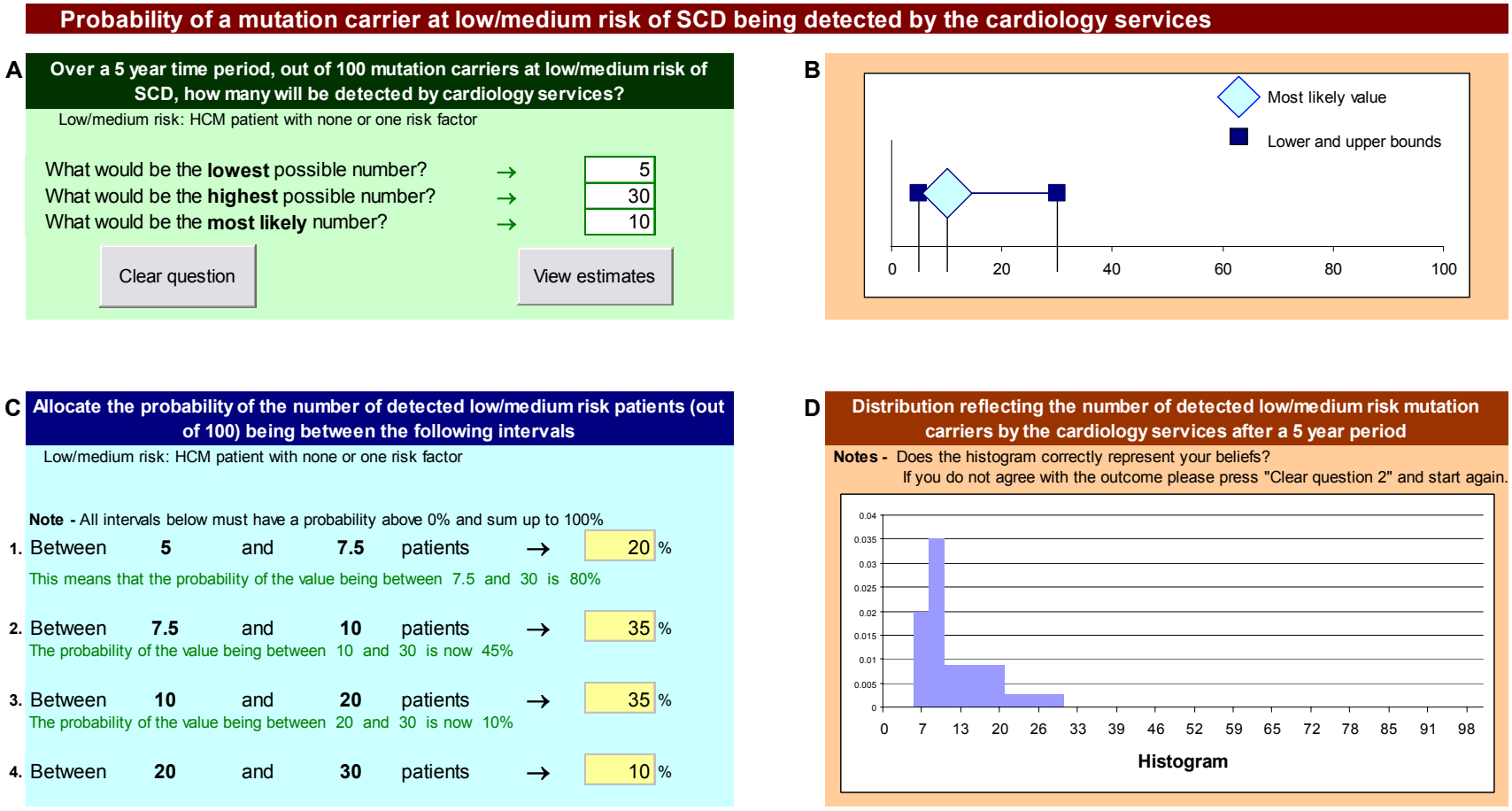


Figure 3. Display of elicited experts' beliefs about two model parameters.

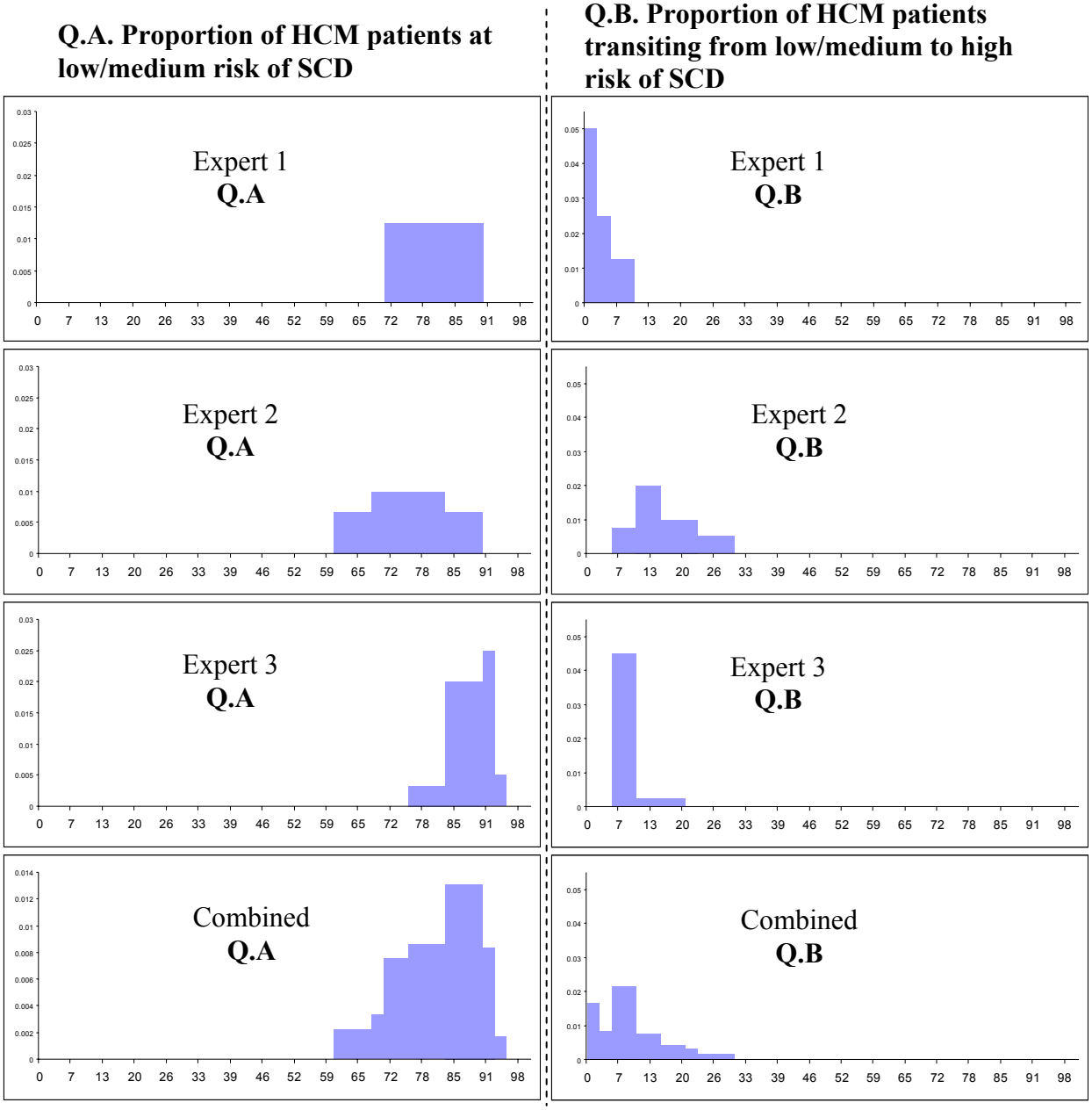
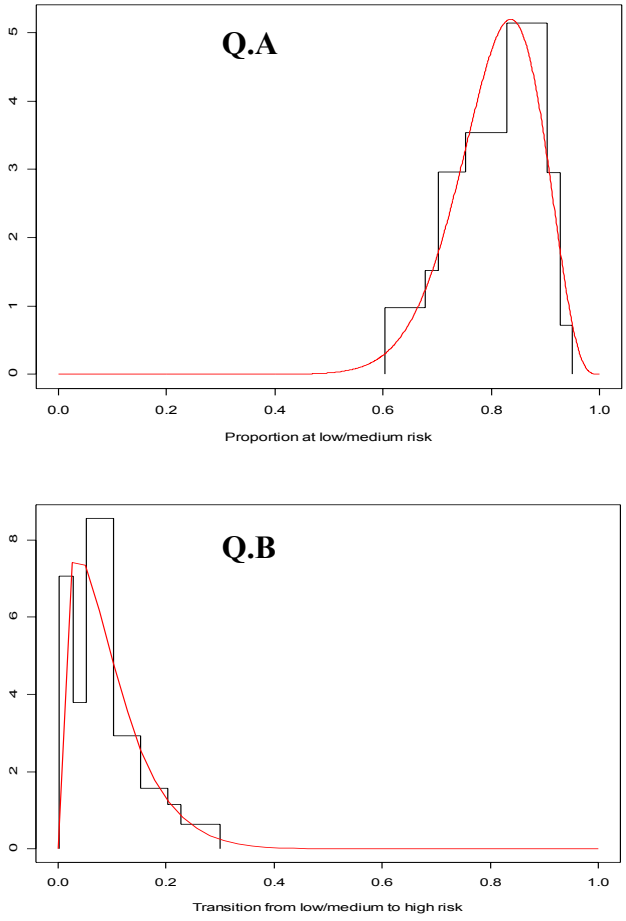


Figure 4. Fitting distributions to combined elicited beliefs.



Appendix

In this appendix we provide an elicitation question together with the background information provided to the expert. The parameter which uncertainty we wish to quantify is the proportion HCM population at low/medium risk of sudden cardiac death.

Aim: Establish the proportion of HCM patients at low/medium risk of sudden cardiac death in the whole HCM population

Background: The predictive clinical features of high risk patients for sudden cardiac death (SCD) are the following:

- Family history of multiple SCDs
- Unexplained syncope
- Flat or hypotensive blood pressure response during upright exercise
- Non-sustained ventricular tachycardia during Holter monitoring
- Severe hypertrophy (>30mm)

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?

References

Briggs, A. H. 2000, "Handling uncertainty in cost-effectiveness models", *Pharmacoeconomics*, vol. 17, no. 5, pp. 479-500.

Clemen, R. T. & Winkler, R. L. 1999, "Combining probability distributions from experts in risk analysis", *Risk Analysis*, vol. 19, no. 2, pp. 187-203.

Cooke, R. M. 1991, *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford University Press.

Drummond, M. F., O'Brien, O., Stoddart, G. L., & Torrance, G. W. 1997, *Methods for the Economic Evaluation of Health Care Programmes: Second Edition.* Oxford University Press.

Elliott, P. & McKenna, W. J. 2004, "Hypertrophic cardiomyopathy", *Lancet*, vol. 363, no. 9424, pp. 1881-1891.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. 2005, "Statistical methods for eliciting prior distributions", *Journal of the American Statistical Association*. (in press).

Genest, C. & Zideck, J. V. 1986, "Combining probability distributions. A critique and annotated bibliography", *Statistical Science*, vol. 1, pp. 114-148.

Kadane, J. B. & Wolfson, L. J. 1998, "Experiences in elicitation", *Journal of the Royal Statistical Society Series D-the Statistician*, vol. 47, no. 1, pp. 3-19.

McKenna, W. J. & Behr, E. R. 2002, "Hypertrophic cardiomyopathy: management, risk stratification, and prevention of sudden death", *Heart*, vol. 87, no. 2, pp. 169-176.

NICE 2004, *Guide to the Methods of Technology Appraisal*, National Institute for Clinical Excellence, London, N0515.

O'Hagan, A. 1998, "Eliciting expert beliefs in substantial practical applications", *Journal of the Royal Statistical Society Series D-the Statistician*, vol. 47, no. 1, pp. 21-35.

Philips, Z., Ginnelly, L., Sculpher, M., Claxton, K., Golder, S., Riemsma, R., Woolacott, N., & Glanville, J. 2004, "Review of guidelines for good practice in decision-analytic modelling in health technology assessment", *Health Technology Assessment*, vol. 8, no. 36.

Phillips, L. D. & Wisbey, S. J. 1993, 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*, Nirex UK, Didcot.

Stone, M. 1961, "The opinion pool", *Annals of Mathematical Statistics*, vol. 32, pp. 1339-1342.