International Journal of Technology Assessment in Health Care, 30:4 (2014), 446–453.
© Cambridge University Press 2014
doi:10.1017/S0266462314000403

AN INTERNET-BASED METHOD TO ELICIT EXPERTS' BELIEFS FOR BAYESIAN PRIORS: A CASE STUDY IN INTRACRANIAL STENT EVALUATION

Leslie Pibouleau
Univ Paris Diderot; Sorbonne Paris Cité; INSERM UMR 717; Hop Saint-Louis, Service de Biostatistique et information Médicale
Haute Autorité de Santé, Service Evaluation Economique et Santé Publique

Sylvie Chevret
Univ Paris Diderot; Sorbonne Paris Cité; INSERM UMR 717; Hop Saint-Louis, Service de Biostatistique et information Médicale

Rationale: Bayesian methods provide an interesting approach to assessing an implantable medical device (IMD) that has evolved through successive versions because they allow for explicit incorporation of prior knowledge into the analysis. However, the literature is sparse on the feasibility and reliability of elicitation in cases where expert beliefs are used to form priors.
Objectives: To develop an Internet-based method for eliciting experts' beliefs about the success rate of an intracranial stenting procedure and to assess their impact on the estimated benefit of the latest version.
Study Design and Setting: The elicitation questionnaire was administered to a group of nineteen experts. Elicited experts' beliefs were used to inform the prior distributions of a Bayesian hierarchical meta-analysis model, allowing for the estimation of the success rate of each version. RESULTS: Experts believed that the success rate of the latest version was slightly higher than that of the previous one (median: 80.8 percent versus 75.9 percent). When using noninformative priors in the model, the latest version was found to have a lower success rate (median: 83.1 percent versus 86.0 percent), while no difference between the two versions was detected with informative priors (median: 85.3 percent versus 85.6 percent).
Conclusions: We proposed a practical method to elicit experts' beliefs on the success rates of successive IMD versions and to explicitly combine all available evidence in the evaluation of the latest one. Our results suggest that the experts were overoptimistic about this last version. Nevertheless, the proposed method should be simplified and assessed in larger, representative samples.

Keywords: Bayesian analysis, Prior information, Elicitation method, Internet-based survey, Medical device

Bayesian methods provide an interesting and innovative alternative approach to standard statistical methods, with the specific advantage of explicitly incorporating previous information into current data when evaluating a health technology. Such approaches have been of particular interest in the context of evaluating implantable medical devices (IMDs), and approximately 10 percent of FDA approvals for medical and radiological devices have been based on such analyses (1;2). Indeed, unlike drugs, most IMDs are updated by means of successive introductions of technical changes into new versions, so that Bayes approaches that allow for the accumulation of data on inferences appear appropriate.

This information on previous versions may be useful in the decision process, when data documenting the clinical impact of technical changes are scarce and noncomparative. Indeed, decision makers are often faced with the issue of assessing the latest version of an IMD when only a small amount of clinical data are available on this last version while genuine information on previous versions exists. The example of the NEUROFORM intracranial stents that are used in the treatment of wide-necked aneurysms in conjunction with embolic coil implants is particularly illustrative of such a context. Three successive versions of this IMD had been proposed over 5 years, with accumulated data based on eighteen independent evaluations including 548 patients, when a new, fourth version of the NEUROFORM was proposed (3). In this context, a Bayesian hierarchical meta-analysis model allowed for the assessment of the overall success rate of the IMD by incorporating all of these data, and it also

The authors thank the experts who completed the survey: Dr. Ronit Agid, Dr. Federico Bilotta, Dr. José Cohen, Dr. Harry Cloft, Dr. Stefano Finitisis, Dr. Alejandro Gonzalez, Dr. Ajeet Gordhan, Dr. Andreas Gruber, Dr. David Hasan, Dr. Brad Hindman, Dr. Yutaka Kai, Dr. Dong Joon Kim, Dr. Bernhard Kis, Dr. Angel Lee, Dr. Guo Liang, Dr. Boris Lubicz, Dr. R. Loch MacDonald, Dr. Patrick Mitchell, Dr. J. Mocco, Dr. José Maria Modenesi Freitas, Dr. Andy Molyneux, Dr. Jorge Mura, Dr. Thanh Nguyen, Dr. Shyam Prabhakaran, Dr. M.N. Psychogios, Dr. Andrew Ringer, Dr. Peter Arthur Ringleb, Dr. Christian Schichor, Dr. Romain Sonneville, Dr. Takatoshi Sorimachi, Dr. Christian Taschner, Dr. Walter van den Bergh, Dr. Peter Waldenberger, and Dr. Simon Yu. The authors also thank the experts who participated in the validation of the questionnaire: Dr. Frédéric Clarençon, Dr. Mikaël Mazighi, and Dr. Zoltan Szatmary. Finally, the authors thank Thierry Celeste, who designed the Web site for the survey. This study was supported by a grant from the French Ministry of Health. The funder was not involved in the implementation of the study, the decision to publish, or the preparation of the manuscript.

Table 1. Estimations of the NEUROFORM Success Rate for Each Study

			Independent estimations			
Author, year of publication	Version	y/n	MLE	Bayesian estimation mean estimate	Model estimations Bayesian MA estimation mean estimate	
Dos Santos Souza, 2005	#2	11/18	61.1%	60.0%	68.7%	68.8%
Wanke, 2005	#2	26/26	100%	96.4%	94.3%	94.0%
Lee, 2005	#2	22/23	95.7%	92.0%	91.7%	91.4%
Sani, 2005	#2	9/10	90.0%	83.3%	87.1%	86.9%
Katsaridis, 2006	#2	51/54	94.4%	92.9%	92.5%	92.4%
Biondi, 2007	#2	31/36	86.1%	84.2%	85.9%	85.8%
Wajnberg, 2009	#2	21/26	80.8%	78.6%	82.0%	81.9%
Liang, 2010	#2	65/77	84.4%	83.5%	84.6%	84.5%
Kadkhodayan, 2011	#2	37/59	62.7%	62.3%	65.6%	65.7%
Biondi, 2007	#3	2/2	100%	75.0%	85.8%	85.6%
Liang, 2010	#3	13/16	81.3%	77.8%	82.5%	82.5%
Gordhan, 2011	#3	22/23	95.7%	92.0%	91.5%	91.3%
Kadkhodayan, 2011	#3	42/56	75.0%	74.1%	76.5%	76.5%
Mangubat, 2012	#4	15/21	71.4%	69.6%	74.1%	75.0%

Note. The fourth and fifth columns indicate independent estimations and the last two columns reports the results obtained with the hierarchical Bayesian meta-analysis (MA) model, either with a non-informative prior or with an informative expert-derived prior (results reported on the bottom). Non-informative priors were Norm(0, var $= 10^6$) for the logit of the overall success rate, μ, and LN(0,1) for the inter-version and inter-study standard deviations, ν and σ; informative priors were Norm(1.6, var $= 2.3$) for μ,and Gamma(1,4) for ν; MLE: maximum likelihood estimate y, number of successes; n, number of treated aneurysms;

permitted testing of the heterogeneity of effects across the versions. Based on this model, with noninformative priors for the model parameters, a marked benefit of the version #2 over the version #1 was estimated, while an absence of improvement was found for the version #3 with respect to the version #2 (3). More recently, a single trial on the version #4 of the NEUROFORM stent was published, based on twenty-one patients (4). Decision makers may wish to analyze the overall benefit of the NEUROFORM stent and evaluate the potential benefit achieved by this latest version. One may wonder whether experts' beliefs about the benefits brought by the last versions (version #3 and version #4) of the NEUROFORM stent would provide more helpful insights than those provided by all of the past information that has been modeled. Indeed, in Bayesian inferencing, informative distributions have been recommended for use wherever substantive prior information exists (5).

The integration of experts' beliefs into a Bayesian model requires the formal incorporation of their opinions into a mathematical prior distribution. This process, referred to as the "elicitation" process, presupposes some understanding of how a person assesses the probability of an event, mostly from heuristics (6). To increase acceptance of the use of Bayesian inferencing in clinical research, a sound methodology of eliciting experts' beliefs is required. Numerous elicitation procedures have been reported in the literature, though their measurement properties have rarely been evaluated (7). Computer-based surveys, through interactive and graphical interfaces, may improve elicitation process (8). First, it may allow to achieve elicitation in a transparent, repeatable and robust manner; second, it may give access to a broad range of experts geographically dispersed and at last, the use of interactive tools may help the expert to assess her/his own probability distribution without requiring the presence of someone with statistical and psychological knowledge, namely a "facilitator".

Thus, we developed a computer-based tool to elicit experts' beliefs about the success rate of the NEUROFORM stent. Given the marked lower efficacy of version #1, it was not considered in this study. We showed how experts' beliefs differ from the available data, and we assessed the potential impact of these beliefs on the assessed benefit of the latest version of the device.

MATERIALS AND METHODS

Data

From a total of eleven previously published evaluations of the new versions (#2, #3, and #4) of the NEUROFORM stent, we selected fourteen datasets of complete evaluations, including nine on version #2, four on version #3, and one on version #4. Observed success rates, defined as the placement of the stent in the target artery and an aneurysm occlusion ≥ 95 percent on the immediate postprocedural angiography, are reported in Table 1 for each dataset.

https://doi.org/10.1017/S0266462314000403 Published online by Cambridge University Press

Experts

We elicited experts' beliefs on the success rates of the latest version of the NEUROFORM stent (version #4) and on the success rate of the previous version (version #3) to determine their beliefs about the benefits brought by the latest version.

A group of nineteen experts was selected from the set of corresponding authors who had written articles on intracranial aneurysms that were referenced in PubMed in the last 6 years. The characteristics of the participants are reported in Supplementary Table 1, which can be viewed online at http://dx.doi.org/10.1017/S0266462314000403. The experts were an international group with practices in Europe ($n = 7$), North America ($n = 5$), South America ($n = 3$), and Asia ($n = 3$). They were all clinicians who were equally distributed among three specialities: neuroradiology, neurosurgery, and neurology. The majority of them were highly active in intracranial aneurysm treatment, and 11 of 19 (58 percent) used the NEUROFORM stent. Concerning normative goodness, previous statistical training, considered as a proxy measure for knowledge in statistics (9), was reported by 9 of 19 (47 percent) of the participants.

Elicitation Tool

We developed a Web-based tool elicitation tool that enabled expert elicitation of probabilities, as previously done in many other settings when face-to-face elicitation is not an option, for instance due to time or budget constraints, and to allow to contact experts worldwide on topics about which little or no knowledge is available. A review of existing tools in the field of environment and health has been previously published (10).

The quantity to be elicited was the success rate of a stenting procedure. The elicitation question was formulated as follows: "if 100 patients were to be operated using this stent, what would be in your opinion the number of successes". Indeed, it has been shown that the assessments made using relative frequency, that is, how many successes there would be in a sample of a given size, exhibit less scatter and express complete certainty less often compared with those made with direct probability (11). A brief description of the version of the stent in question was first provided to the experts. The whole questionnaire is available in Supplementary Table 2, which can be viewed online at http://dx.doi.org/10.1017/S0266462314000403.

Various techniques for eliciting a success rate have been described in the literature (12;13), while others focused on a regression coefficient (14;15). Because the elicitation was conducted on a group of nonstatistician experts without the assistance of a facilitator, we used the fixed-interval approach of Leal et al. (16) that appeared more appropriate for nonstatistician experts compared with the variable-interval method (17).

Let X denotes the number of successes among 100 patients, under this fixed-interval method, the expert is asked about the range of plausible values of X and the mode, the most likely value. If any inconsistency occurred, the expert was automatically informed. The distance between each extreme value and the most likely value was then equally divided by 2, and the expert was asked to give a probability (expressed as a "weight of belief") for both resulting intervals (in principle, assessed probabilities must sum to one otherwise they are standardized to one). A histogram was then derived from these estimates (Supplementary Figure 1, which can be viewed online at http://dx.doi.org/10.1017/S0266462314000403), and the expert was asked to confirm whether it represented her/his beliefs or to correct her/his previous estimates if it did not. Once validated, this histogram constituted the expert's distribution of success rate for the considered version of stent.

As recommended (7), we used strategies to minimizing the effect of biases on the elicited beliefs (Table 2). First, to encourage the experts to consider possible sources of uncertainty explicitly (11), they were asked about predictive factors of success. Second, a training exercise was built on the same model as the elicitation exercise. To make them more aware of aleatory uncertainty (18), this exercise focused on a future event which had nothing to do with their domain of expertise. Third, a feedback mechanism with a graphical aid and opportunity for revision was provided.

The questionnaire was implemented in php/html and was initially tested by four departmental colleagues and three experts in neuroradiology and neurology. The definition of success was reformulated to eliminate any ambiguity. Modifications in the design were implemented to make the tool clear, easy and attractive to use.

The methods used to assess the measurement properties of the questionnaire are summarized in Table 3. The belief elicitation procedure demonstrated good agreement for validity and good feasibility (data available upon request).

Construction of Expert Priors

Each expert elicitation resulted in a histogram that constituted the expert's individual distribution of the success rate of the stent version considered. Based on this histogram, the probability of the success rate in each 5 percent interval from 0 to 100 percent was computed.

The experts' individual distributions were then pooled to obtain a collective prior distribution using the linear opinion pool method proposed by Stone (19). Assuming the equivalence of experts, this provided a simple average of the experts' distributions. This method resulted in a so-called "group prior" for each version of the stent. A parametric distribution of the experts' data was then fitted using the parametric family of distributions suggested by the group prior on a logit scale.

Incorporation of the Experts' Beliefs into the Statistical Model

To take into account the heterogeneity of the IMD versions in addition to the inter-study and intra-study heterogeneity, a Bayesian three-level hierarchical model was fitted. A detailed

https://doi.org/10.1017/S0266462314000403 Published online by Cambridge University Press

Table 2. Sources of Bias and Methodological Strategies Used to Control Them

Source of bias (heuristics and other sources)	Interpretation and related bias	Methodological strategy to control bias
Anchoring and adjustment	Tendency to anchor around an initial estimate (such as the mode), and not adjust the final estimate far enough away from this value	Elicit the plausible range of the quantity to be elicited, i.e., the lower and upper bounds, before the mode
Conservatism	Conservatism relates to the process of an expert understating her/his beliefs	Use averaging methods for the group clinical prior
Overconfidence	Overestimating the accuracy of her/his beliefs or alternatively underestimating the uncertainty in a process	- Encourage experts to consider possible sources of uncertainty explicitly - Include experts (sample size greater than 1) - Use averaging methods for the group clinical prior
Representativeness	Providing opinions that are based on situations that are (wrongly or rightly) perceived to be similar	Encourage experts to consider the issue of generalizability of results to other contexts (centres, patient characteristics, health organisations, . . .)
Law of small numbers	Expert bases her/his opinion on small pieces of information and assumes that this extrapolates to the population	Use relative frequency, i.e., how many successes there would be in a sample size of 100, instead of direct probability
Linguistic uncertainty	Misunderstanding the question and / or applying different interpretations to the same term	Define precisely and unequivocally the quantity (success rate) for which a distribution is to be elicited
Normative goodness	Expert's ability to express her/his knowledge in accordance with the calculus of probabilities	Use a training exercise Provision of feedback for verification and opportunity for revision
Substantive goodness	Knowledge of the expert relative to the problem at hand	Selection of experts with a high degree of knowledge in the concerned domain
Conflict of interest		Declaration of any financial or personal interest that the expert might have in the decisions that will depend on the expert's distribution

Note. Based on Johnson et al. (7).

report of this method is published elsewhere (3). Such a model allows to estimating not only the overall success rate of the IMD but also the success rate of each IMD version. This model requires the specification of prior distributions for three parameters (on a logit scale): the overall success rate (μ), the inter-version variability (ν) and the inter-study variability (σ).

We first considered noninformative prior distributions which were: a normal distribution with a mean of 0 and a large variance, N(0, 10^6) for μ, and a log normal LogN(0,1) distribution for σ and ν (3).

Then, to integrate the experts' opinions into the above model, we used their beliefs on the success rate of the version #3, P_3, and that of the version #4, P_4, to inform the priors of μ and ν. The group prior distribution of the success rates of both versions combined on a logit scale was used as the informative prior for m. The prior for ν was obtained through a sampling approach. Details of the model as well as its specification in BUGS code are provided in Supplementary Table 3, which can be viewed online at http://dx.doi.org/10.1017/S0266462314000403.

We generated posterior distributions by combining the elicited prior distribution with the trial data using OpenBUGS software version 3.2.2. A total of 40,000 iterations following 10,000 burn-in iterations were used after checking for convergence. Posterior mean and median estimates of the success rates were computed for each study and for each version of the stent. Last, the posterior probability that the success rate of a new version was higher than that of the previous one was computed.

To detect a version effect, we compared the Deviance Information Criterion (DIC) obtained after fitting the data either with a Bayesian random-effects model or with a hierarchical one. A better fit of the hierarchical model would indicate a version effect, as previously published (3).

All statistical analyses were performed using the R software package, version 2.13.2 (R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org).

RESULTS

Table 4 summarizes all of the information available on the success rates of version #3 and #4 according to the source of information or model.

Elicitation of Experts' Beliefs

The median of the most likely values of the success rates were 75 percent (IQR: 70–88 percent) for version #3 and 90 percent (IQR: 65–93 percent) for version #4, with a pairwise difference

Table 3. Methods of Assessment of Measurement Properties of the Elicitation Questionnaire

Measurement properties	Definition	Method of assessment
Validity	Refers to the degree to which the elicitation questionnaire measures what it purports to measure	Asking experts about the comprehensiveness of the questions Was the question clear? □ Yes □ No Comparing the experts' answers to two questions assessing the effect of the new version of stent Q1: What effect do you believe the new version may have on the success rate compare to the current version of stent? □ improves / □ worsens / □ has no effect Q2: What is the mode (M) for the success rate of each version?
Reliability	Refers to the reproducibility of the measure: - Intra-rater reliability (test–retest) : questionnaire administered to the same participant on two different occasions	Administering the questionnaire to participants on two occasions 4–5 weeks apart
Feasibility	Refers to the ease of use of the tool in terms of ease and time of completion	Asking experts to comment on the ease of each response option and their time to complete the elicitation process: Was the response option easy to use? □ Yes □ No Time to complete the elicitation process □ acceptable □ too long □ < 5 min □ 5 to 10 min □ 10 to 15 min □ 15 to 20 min □ > 20 min

Table 4. Information Available on the Success Rates of the Two Last Versions of the NEUROFORM Stent

	Version #3	Version #4	Pooled
Published data: point estimate	80.8% (79/97)	69.6% (15/21)	79.7% (94/118)
Expert Opinion : median [IQR]			
Most likely rate	77.5% [70.0% - 90.0%]	90.0% [65.0% - 92.5%]	80.0% [70.0% - 90.0%]
Group prior	75.9% [62.2% - 87.8%]	80.8% [61.5% - 91.3%]	78.0% [62.2% - 89.5%]
Posterior estimates : median [IQR]			
Non-informative priors[a]	86.0% [81.9% - 89.4%]	83.1% [76.2% - 87.9%]	85.2% [80.5% - 88.9%]
Informative priors[b]	85.6% [82.9% – 88.2%]	85.3% [82.3% - 88.0%]	85.4% [82.5% - 88.1%]

[a] Non-informative priors were Norm(0, var $= 10^6$) for the logit of the overall success rate, μ, and LogN(0,1) for the inter-version and inter-study standard deviations, ν and σ.

[a]Informative priors were Norm(1.6, var $= 2.3$) for μ, and Gamma(1,4) for ν.

across the two versions of 0.5 percent (IQR: 0–5 percent). Only one expert believed that the success rate of the version #4 was lower than that of the version #3.

Based on the elicited 'group' prior, the median of P_3 was 75.9 percent (IQR: 62.2–87.8 percent) versus 80.8 percent (IQR: 61.5–91.3 percent) for P_4. The resulting "group" priors on a logit scale for each version and for pooled versions are provided in Supplementary Figure 2, which can be viewed online at http://dx.doi.org/10.1017/S0266462314000403. Based on these distributions, a Normal prior N(1.6, 2.3) was chosen for μ and a Gamma prior $\gamma(1, 4)$ was chosen for ν.

Estimation of the Success Rates of the Device Versions

The hierarchical meta-analysis model was first applied to the complete NEUROFORM stent data using noninformative priors

https://doi.org/10.1017/S0266462314000403 Published online by Cambridge University Press

(Table 1). The median posterior success rates of version #2 and version #3 were close, 86.3 percent (IQR: 83.3–88.8 percent) and 86.0 percent (IQR: 81.9–89.4 percent), respectively, whilst that of version #4 was lower, 83.1 percent (IQR: 76.2–87.9 percent) (Table 4).

This result was also illustrated by the 47.9 percent estimated posterior probability that version #3 had a better success rate than the former. This probability dropped to 35.6 percent when considering the superiority in efficacy of version #4 over #3. Actually, based on the model DIC, the random-effect meta-analysis model was the best-fitting model, highlighting the absence of a version effect in efficacy. With this model, the median success rate of the pooled versions was estimated at 85.7 percent (IQR: 83.2–88.2 percent).

The hierarchical meta-analysis model was then applied with the informative 'group' priors (Table 4). The resulting posterior distributions of success rates were close for all versions: the median of P_2 was 85.7 percent (IQR: 83.1–88.1 percent), that of P_3 was 85.6 percent (IQR: 82.6–88.3 percent) and that of P_4 was 85.1 percent (IQR: 81.7–88.0 percent). The posterior probability that P_3 was above P_2 was 48.3 percent, and the posterior probability that P_4 was above P_3 was 44.5 percent.

Compared with the estimate derived from the noninformative priors, the success rate of P_4 had a higher mean estimate (84.5 percent versus 80.5 percent) and a reduced uncertainty (standard deviation of 6.2 percent versus 11.2 percent). Otherwise, the DICs were close regardless of the model.

DISCUSSION

This article describes a case study in the evaluation of an implantable medical device assessing how feasible it is to conduct an elicitation session in a structured manner using a Web questionnaire and to form a probability distribution. An original computer-assisted method of eliciting prior distributions for Bayesian models using expert knowledge was first proposed. Then, we elicited expert opinions using cumulative probabilities to model the imprecision of individual experts as well as the variability between experts. We then presented an estimation method for using this source of elicitation data.

For eliciting the prior probability distribution of the success rate, we used a fixed-interval method though other methods implemented in the various available computer-based tools (10) may have been used. Note that some of these methods were directed toward a regression coefficient as the summary to elicit (14;15). Otherwise, the variable-interval method in which experts are asked about specified percentiles, mostly the median and the quartiles, of their subjective distribution, could appear an alternate candidate. However, the fixed-interval method was reported faster and slightly superior over the variable-interval method along several dimensions such as monotonicity, accuracy, and precision (17). Moreover, a clear-cut preference by most participants for the fixed-interval method was also reported and explained by convenience and ease of use (17). This was confirmed in our survey with 79 percent of the participants who found the questions to be clear and the response options easy to use. Regarding the number of intervals of the fixed-interval approach, we used four intervals whereas such a number varies across studies (11;16;20;21). Nevertheless, a higher number of intervals has been reported to be time-consuming and difficult to undertake by respondents (16).

Among the available computer-based tools, the SHELF framework developed by Oakley and O'Hagan (22) is of particular interest as it proposes five techniques of elicitation with a R package to fit parametric distributions to experts' answers. However, this framework relies on a face-to-face interview, thus could not apply directly in our setting. Nonetheless, exploring the possibilities offered by the SHELF framework is an interesting research perspective for distance elicitation which was recently adopted by Sperber et al. (23).

Once the experts have reported their own answers to the four intervals probabilities, the elicitation task was completed by converting these into a probability distribution, using a mathematical aggregation. Other approaches such as behavioural approaches for reaching consensus could have been used to combine individual expert distributions. Nevertheless, no clear benefit of using interactive approaches over individual elicitation methods has been reported (24). Moreover, guidelines for making healthcare decisions recommend not enforcing consensus in expert opinions (e.g., using standard Delphi methods) to appropriately assess uncertainty in parameters (25). First eliciting the individual beliefs of the experts and then gathering them to gain additional benefits from the exchange of information among them appears promising (22).

When using noninformative priors, success rates were found to be similar between versions #2 and #3 but lower for version #4. By contrast, when using informative priors derived from experts' beliefs, success rates were similar whatever the version. Indeed, as experts were very confident about the success rate of version #4 compared with that of version #3, this translated into a low inter-version variability, bringing the hierarchical model closer to a standard random-effects one. These results suggest that the experts were overly optimistic, so inferences about the IMD based solely on the opinions of experts could be questionable. Moreover, one cannot exclude some conflict of interest with industry, even though only one participant endorsed, that may explain the over-optimistic opinions of the experts regarding the improvement of version #4 over #3. In all cases, respondents are likely a biased sample of the population and this limits the generalizability of our findings. Although the number of participants was greater than the median of the reported sample sizes ($n = 11$) (7), the low response rate was the major limitation of our study. Of the 341 contacted authors, thirty-four (10 percent) connected to the Web site. Our recruitment method, that is, asking by e-mail all authors of articles published on intracranial aneurysms, appears inappropriate to

https://doi.org/10.1017/S0266462314000403 Published online by Cambridge University Press

the complexity of the task requiring a strong implication of the participants. Three reminder e-mails were sent to all authors but difficulty was probably to gain attention from very busy people. Methods allowing to further motivate the experts, for example by contacting members of learned societies, participants in conferences related to the specialty, or clinicians from all of the neurosurgery departments in Paris hospitals, could have increased the response rate.

Of the thirty-four participating authors, fifteen (44 percent) dropped out at the training exercise level and only nineteen fulfilled the elicitation questionnaire. The length of the questionnaire, as well as the training exercise seemed to have discouraged some of the participants, possibly indicating their difficulty of understanding. However, experts of the "dropout group" did not significantly differ from those of the "elicitation group," except with respect to their English skills (first language for 42 percent in the elicitation group versus 27 percent in the dropout group) and statistical training (47 percent versus 20 percent, respectively). The remote process of elicitation may thus have selected experts already at ease with probability calculus, illustrating the need for some help file in the elicitation. Besides, such a normative goodness is actually one expertise required from the participants in addition to substantive goodness, so that, although uncommon, it is likely that the selected participants were actual experts.

Among the nineteen experts of the elicitation group, only eleven of them had already used the NEUROFORM stent. This difference in expertise questioned the assumption of the equivalence of experts we made when building the group prior. Indeed, users had less uncertainty about the success rates of both versions and were also more optimistic than nonusers regarding the benefit brought by the latest version. To take into account the difference in expertise, we derived a new "weighted" group prior with the linear opinion pool method by attributing double weight to users compared with nonusers. The resulting weighted group prior did not markedly differ from the unweighted prior and posterior estimates were not affected by the weighting (data available upon request).

In conclusion, we proposed a practical method of eliciting experts' beliefs about the success rate of the latest version of an IMD with the use of an Internet-based elicitation tool allowing for the estimation of the efficacy of successive versions of the IMD. The proposed elicitation method was easy to implement, with elicitation at a distance that did not require the intervention of a facilitator. A key feature of this work is that all available evidence was included in the evaluation in a transparent way. This method could be applied to many other settings of evaluation, when data are sparse and experts are asked for their opinion.

SUPPLEMENTARY MATERIAL

Supplementary Table 1:
http://dx.doi.org/10.1017/S0266462314000403
Supplementary Table 2:
http://dx.doi.org/10.1017/S0266462314000403
Supplementary Figure 1:
http://dx.doi.org/10.1017/S0266462314000403
Supplementary Table 3:
http://dx.doi.org/10.1017/S0266462314000403
Supplementary Figure 2:
http://dx.doi.org/10.1017/S0266462314000403

CONTACT INFORMATION

Leslie Pibouleau, MSc, PhD, (l.pibouleau@has-sante.fr), Service Evaluation Economique et Santé Publique, Haute Autorité de Santé, 2 avenue du Stade de France, 93 218 Saint-Denis La Plaine, France.
Sylvie Chevret, MD, PhD, Professor of Public Health, Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75 475 Paris Cedex 10, France.

CONFLICTS OF INTEREST

All authors report they have no potential conflicts of interest.

REFERENCES

1. Campbell G. Bayesian statistics in medical devices: Innovation sparked by the FDA. *J Biopharm Stat*. 2011;21:871-887.
2. Pibouleau L, Chevret S. Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices. *J Clin Epidemiol*. 2010;64:270-279.
3. Pibouleau L, Chevret S. Bayesian hierarchical meta-analysis model for medical device evaluation: Application to intracranial stents. *Int J Technol Assess Health Care*. 2013;29:123-130.
4. Mangubat EZ, Johnson AK, Keigher KM, Lopes DK. Initial experience with neuroform EZ in the treatment of wide-neck cerebral aneurysms. *Neurointervention*. 2012;7:34-39.
5. O'Hagan A. *The Bayesian approach to statistics*. http://www.sagepub.com/upm-data/18550_Chapter6.pdf (accessed October 20, 2013).
6. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain judgements - eliciting experts' probabilities*. New York: Wiley; 2006.
7. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: A systematic review. *J Clin Epidemiol*. 2010;63:355-369.
8. Low-Choy S. *Expert elicitation and its interface with technology: A review with a view to designing Elicitator*. unpublished. 2009. http://mssanz.org.au/modsim09/J2/lowchoy.pdf (accessed October 20, 2013).
9. Johnson SR, Tomlinson GA, Hawker GA, et al. A valid and reliable belief elicitation method for Bayesian priors. *J Clin Epidemiol*. 2010;63:370-383.
10. Devilee JLA, Knol AB. Software to support expert elicitation. *RIVM Lett Rep*. 2011.
11. O'Hagan A. Eliciting expert beliefs in substantial practical applications. *Statistician*. 1998;47(Pt 1):21-35.
12. O'Hagan A, Buck CE, Daneshkhah A, et al. Uncertain judgements - eliciting experts' probabilities. New York: Wiley; 2006.
13. Winkler RL. The quantification of judgment: Some methodological suggestions. *J Am Stat Assoc*. 1967;62:1105-1120.
14. Denham R, Mengersen K. Geographically assisted elicitation of expert opinion for regression models. *Bayesian Anal*. 2007;2:99-136.

15. Garthwaite PH, Al-Awadhi SA, Elfadaly FG, Jenkinson DJ. Prior distribution elicitation for generalized linear and picewise-linear models. *J Appl Stat*. 2013;40:59-75.
16. Leal J, Wordsworth S, Legood R, Blair E. Eliciting expert opinion for economic models: An applied example. *Value Health*. 2007;10:195-203.
17. Abbas AE, Budescu DV, Yu HT, Haggerty R. A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Anal*. 2008;5:190-202.
18. Hora SC. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliab Eng Syst Saf*. 1996;54:217-223.
19. Stone M. The opinion pool. *Ann Math Stat*. 1961;32:1339-1342.
20. Phillips LD, Wisbey SJ. *The elicitation of judgemental probability distributions from groupes 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.
21. Soares MO, Bojke L, Dumville J, et al. Methods to elicit experts' beliefs over uncertain quantities: Application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration. *Stat Med*. 2011;30:2363-2380.
22. Oakley JE, O'Hagan A. SHELF: *The Sheffield Elicitation Framework (version 2.0). Sheffield, UK: School of Mathematics and Statistics*, 2010. http://www.tonyohagan.co.uk/shelf/ (accessed October 20, 2013).
23. 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;16:434-437.
24. Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Anal*. 1999;19:187-203.
25. 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:iii-iv, ix-xi, 1-158.

https://doi.org/10.1017/S0266462314000403 Published online by Cambridge University Press