Appl Health Econ Health Policy (2014) 12:327–334
DOI 10.1007/s40258-014-0092-y

ORIGINAL RESEARCH ARTICLE

# Belief Elicitation to Populate Health Economic Models of Medical Diagnostic Devices in Development

**Wieke Haakma · Lotte M. G. Steuten · Laura Bojke · Maarten J. IJzerman**

Published online: 13 March 2014


**Abstract**

*Background and Objective* Bayesian methods can be used to elicit experts' beliefs about the clinical value of healthcare technologies. This study investigates a belief–elicitation method for estimating diagnostic performance in an early stage of development of photoacoustic mammography (PAM) imaging versus magnetic resonance imaging (MRI) for detecting breast cancer.

*Research Design* Eighteen experienced radiologists ranked tumor characteristics regarding their importance to detect malignancies. With reference to MRI, radiologists estimated the true positives and negatives of PAM using the variable interval method. An overall probability density function was determined using linear opinion pooling, weighted for individual experts' experience.

*Result* The most important tumor characteristics are mass margins and mass shape. Respondents considered MRI the better technology to visualize these characteristics. Belief elicitation confirmed this by providing an overall sensitivity of PAM ranging from 58.9 to 85.1 % (mode 75.6 %) and specificity ranging from 52.2 to 77.6 % (mode 66.5 %).

*Conclusion* Belief elicitation allowed estimates to be obtained for the expected diagnostic performance of PAM, although radiologists expressed difficulties in doing so. Heterogeneity within and between experts reflects this uncertainty and the infancy of PAM. Further clinical trials are required to validate the extent to which this belief–elicitation method is predictive for observed test performance.

**Key Points for Decision Makers**

This article presents a new application of belief elicitation to estimate the clinical value of a medical imaging device in an early stage of development.

Belief elicitation in an early stage can identify the potential diagnostic performance of a medical device and can support developers to prioritize between prototypes and features to improve the technology.

This method can characterize uncertainty regarding the potential performance of a medical technology to support development decisions. Yet, several adaptations for the use of belief–elicitation methods in early development stages are identified.

## 1 Introduction

Sound research and development decisions in medical technology development require effective methods and procedures for identifying and assessing the anticipated impact of new technologies. Early health technology

**Electronic supplementary material** The online version of this article (doi:10.1007/s40258-014-0092-y) contains supplementary material, which is available to authorized users.

W. Haakma · L. M. G. Steuten · M. J. IJzerman (✉)
Department of Health Technology and Services Research, MIRA Institute for Biomedical Technology and Technical Medicine, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.j.ijzerman@utwente.nl

W. Haakma
e-mail: wiekehaakma@gmail.com

W. Haakma
Department of Forensic Medicine and Comparative Medicine Lab, Aarhus University, Aarhus, Denmark

L. Bojke
Centre for Health Economics, University of York, York, UK

assessment has been proposed as an approach to inform product development and market access strategies for new medical products [1, 2]. Specifically, several authors have explored early (Bayesian) health economic modeling, which allows for prior beliefs or existing evidence to be updated by new information that becomes available at later stages of development [2–5]. As such, it may support both (1) developers in prioritizing between several competing possible concepts, prototypes, or features of the technology, and (2) identifying parameters that have a large impact on the diagnostic, clinical, and economic value [2].

Because early development stages are typically characterized by scarcity of empirical data, to determine the cost effectiveness of a new technology, early health economic models need to be populated with alternative sources of information, such as elicited priors or beliefs. To incorporate elicited beliefs in health economic models, these have to be expressed in a statistical form, presenting distributions that reflect the uncertainty surrounding them. Several studies have reported methods for constructing such priors, such as Hiance et al. [6] who reported a rational approach to construct an expert-based prior to estimate the 3-year event-free survival of two treatments in chronic lymphocytic leukemia, and Johnson et al. [7] who investigated the feasibility, reliability, and validity of a elicitation method called 'bins-and-chips' to estimate the prior probability of 3-year survival with and without warfarin. In addition, a few applications in health economic modeling have been published, including by Leal et al. [8] who used belief elicitation to estimate the parameters of an economic model for evaluating new DNA testing technologies, Bojke et al. [9] who assessed the cost effectiveness of treatments for active psoriatic arthritis by asking experts to predict unknown parameters, and Soares et al. [10] who used elicitation to inform decision models to estimate the cost effectiveness of negative pressure wound therapy. Although belief elicitation is used in several studies, it has not been applied in early stages of technology development. Obviously, the clinical performance of technologies in these early stages is hard to quantify.

The main objective of this study was to determine the added value of a belief–elicitation method for determining uncertain priors for an early health economic model. The belief–elicitation method used here aims to minimize some of the biases previously identified in belief elicitation [11], in particular clarity bias, while in the analysis specific attention is paid to intra- and inter-respondent heterogeneity [7].

The case study used for the elicitation exercise is a comparison of the expected performance of a prototype photoacoustic imaging technology of the breast with the known performance of standard magnetic resonance (MR) imaging (MRI) for the detection of breast cancer after an inconclusive X-rays or mammogram [12, 13].

## 2 Methods

### 2.1 Photoacoustic Imaging and Magnetic Resonance Imaging as Comparator

A new technology, the photoacoustic mammoscope (PAM), was developed to identify vascularization in tissue. As tumor growth is often associated with increased blood vessel supply, PAM combines light and ultrasound to detect the hemoglobin in vessels. An important application of this technology includes breast cancer visualization. At the time of performing this research, two small clinical trials of 13 patients have been performed in a diagnostic setting using the first prototype of PAM [12, 13]. Therefore, limited information regarding the clinical performance of PAM was available.

In this study, MRI was chosen as a comparator of PAM as an alternative to resolve uncertain findings after X-ray mammography and ultrasound in the diagnostic trajectory of breast cancer. MRI makes use of magnetic fields to visualize the tissue. Gadolinium is often used as a contrast agent to identify angiogenesis (growth of new blood vessels, essential for cancer progression). The main difference between these technologies is the ability to visualize different types of tumor characteristics in breast images. Because MRI requires the use of a contrast agent, there is a small risk of chemical exposure which is not relevant in PAM. Furthermore, MRI has a low specificity in the detection of breast cancer and is more expensive than other techniques [14]. Therefore, there may be a priori reason to believe that PAM could be clinically superior to MRI. As MRI is familiar to radiologists, it was expected that they would perform better in the elicitation task when asked to express beliefs relative to a known imaging technique such as standardized MRI. However, this may come at the cost of potential bias due to anchoring [11].

### 2.2 Study Overview and Selected Experts

In this study, experts (radiologists specialized in examining MR images of breasts in The Netherlands) were asked to estimate the expected diagnostic performance of PAM. As the performance of a diagnostic test is a trade-off between the number of true positives (TP) and the number of false positives and conditional on the prior probability of disease, it is not possible to directly estimate the sensitivity and specificity [5]. Instead, experts were asked to estimate the TP, which is the number of images correctly identified by radiologists as positive (cancer is present), and the true negatives (TN), which is the number of images correctly identified as negative (cancer is not present). Both estimates were derived and then compared to the pooled MRI data. In addition, specific tumor characteristics relevant to

detecting breast cancer were provided and explicitly ranked in the elicitation process to prepare the experts with clinical concepts that resonate well within their usual clinical decision-making framework to avoid a simple decision heuristic affecting comprehensiveness of the elicited beliefs and to reduce bias. By explicitly asking them to rank tumor characteristics, the decision problem is better framed and it avoids experts thinking in simple decision heuristics, neglecting important parts of the information [15]. Finally, a calibration procedure was applied to account for heterogeneity in expertise [16].

Twenty radiologists were invited to participate in this study. Radiologists were chosen as the experts as they are the professionals who will eventually assess images provided by PAM. Experts were selected using purposive sampling based on predefined characteristics such as the expected level of knowledge and their experience and expertise in the detection of breast cancer using MRI. Experts from both academic and non-academic hospitals within The Netherlands were selected to ensure experience with both a specialized academic and general patient population. None of the experts had any involvement in the development of PAM. The number of experts was limited to 20 as this is generally assumed to be sufficient [11, 17].

### 2.3 Rating of Tumor Characteristics

The elicitation commenced by identifying the expected performance of PAM and MRI to detect tumor characteristics as usually judged in the examination of images of breasts. These tumor characteristics are identified from the BI-RADS (Breast Imaging–Reporting and Data System) classification system to grade breast lesions [18] and include (1) mass margins; (2) mass shape; (3) mass size; (4) vascularization; (5) localization; (6) oxygen saturation; and (7) mechanical properties. First, the experts are asked how important the tumor characteristics are in the examination of images using point allocation (100 points). This was calculated using Eq. 1:

$$I_{\mathrm{tc}_j} = \sum_{i=1}^{n} \mathrm{SI}_{\mathrm{tc}_j} * w_i \tag{1}$$

where $\mathrm{tc}_j$ is the individual tumor characteristic, SI is the scored importance for $\mathrm{tc}_j$, and $w_i$ is the weight of that individual expert. Following this, they are asked how well MRI and PAM can visualize these characteristics by grading each characteristic with a value ranging from 0 to 100, where 0 indicates a low performance and 100 indicates a high performance. Subsequently, the expected performance of MRI and PAM was determined by using Eqs. 2 and 3:

$$\mathrm{MRI}_{p(\mathrm{tc}_j)} = \sum_{i=1}^{n} w_i * (\mathrm{MRI}_{p_i}(\mathrm{tc}_j)) \tag{2}$$

$$\mathrm{PAM}_{p(\mathrm{tc}_j)} = \sum_{i=1}^{n} w_i * (\mathrm{PAM}_{p_i}(\mathrm{tc}_j)). \tag{3}$$

where the performance ($p$) of each tumor characteristic ($\mathrm{tc}_j$) was included and $w_i$ accounted for the weight ($w$) of each individual expert ($i$) (see Sect. 2.5).

### 2.4 Direct Elicitation of the Probability Distribution Function of the True Positives and True Negatives

A questionnaire and spreadsheet-based (Microsoft® Excel®) exercise was designed for the elicitation experiment (see Electronic Supplementary Material 1, 2, and 3). Background information regarding PAM was provided and the first results of the technology were shown. This was similar for each individual radiologist. The questionnaire was administered on a face-to-face basis, requiring 30–45 min to complete. During the elicitation process a standardized script was used. The elicitation process and purpose was explained, including how uncertainty should be expressed. Questions were carefully formulated with the help of a clinical collaborator and feedback was provided to check whether the questions were understood. After the each elicitation step (i.e., background questions, the elicitation of tumor characteristic importance, and the elicitation of TP and TN), experts had the opportunity to revise their answers as suggested by Johnson et al. [7]. In the Excel® sheet, radiologists provided their judgments of TP and TN numbers for PAM relative to pooled data for MRI. The pooled TP and TN were based on four studies [19–22] on the basis of clinical relevance (i.e., MRI in a diagnostic setting). We used general approaches for data pooling as used in systematic reviews [3, 4] of diagnostic studies. A $2 \times 2$ table was constructed in which the experts had to estimate the TP (cell 'a') and TN (cell 'd'), which was sufficient to estimate TP, TN, false positives, and false negatives, because the row totals were given and kept constant.

Based on previous research [23] and pilot testing of our Excel®-based experiment, the mode (most likely value) was expected to be the most intuitive parameter for experts to elicit using the variable interval method [3, 5]. In the final Excel® format experts were asked to indicate the mode and the lower and the upper boundaries within a 95 % credible interval. A graphical display was used to represent the expert estimates in a probability density function (PDF) and the Project Evaluation and Review Technique (PERT) approach was applied to calculate the mean (μ) (Eq. 4), standard deviation (σ; given as Stdev in

the equations) (Eq. 5), alpha (α) (Eq. 6), and beta (β) (Eq. 7) [24]. The PERT approach provides a triangular distribution, and is based on the mean of the most optimal estimation and the most pessimistic estimation [24]. A beta distribution was used, since this is a flexible and mathematically convenient class to distribute the PDF [4].

$$\text{Mean} = \frac{\min + 4 * \text{mode} + \max}{6} \tag{4}$$

$$\text{Stdev} = \frac{\max - \min}{6} \tag{5}$$

$$\alpha = \left(\frac{\text{mean} - \min}{\max - \min}\right) * \left(\frac{\text{mean} - \min * (\max - \text{mean})}{\text{stdev}^2}\right) \tag{6}$$

$$\beta = \left(\frac{\max - \text{mean}}{\text{mean} - \min}\right) * \alpha. \tag{7}$$

The radiologists' estimations and weights (see Sect. 2.5) were synthesized using the linear pooling method [9]. The radiologists' weights are aggregated and used to obtain an overall weighted distribution using Eq. 8, where $p(\theta)$ is the probability distribution for the unknown parameter $\theta$ and where $w_i$ is the radiologist $i$'s weight summing up to 1.

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta). \tag{8}$$

### 2.5 Calibration Procedure

As heterogeneity between experts was expected, a calibration method was applied based on clinical background [5] to try to explain some of this heterogeneity. Years of experience was indicated as an important aspect to determine the weight of an expert [25], where experts received a score of 2 points when having more than 3 years of experience, otherwise they received a score of 1 point. As repetition of procedures is suggested to result in higher success rate [26], this factor is included as the average number of MR images examined per week in the calibration procedure. The last aspect involves the examination of MR images in other areas (Table 1). For each aspect a weight was calculated using the score and the importance of this aspect for each individual expert. The individual weights obtained by this calibration procedure of experts are included in the synthesis process.

## 3 Results

Of 20 invited radiologists, two were unable to attend (see Table 2, which shows the predefined characteristics per radiologist and their calibration rates). There was no difference in background between attending and non-attending radiologists. During face-to-face interviews, some radiologists expressed difficulties while formulating their judgments. In one case, the radiologist was resistant to the method and produced inconsistent data leading to a decision to exclude them from further analysis.

Due to incomplete responses, some of the weighted averages were determined using smaller sample sizes. The most important characteristics in the assessment of images of breasts are the mass margins and shape. The importance of the characteristics with their 95 % confidence interval is indicated in Fig. 1. Further information regarding the performance of PAM and MRI can be found in the Electronic Supplementary Material 4.

Characteristics such as mechanical properties and oxygen saturation are ranked less important. MRI was estimated to perform better at visualizing mass margins and mass shape, where PAM was estimated to perform better at visualizing vascularization and mechanical properties by the experts (Fig. 2).

### 3.1 Sensitivity and Specificity

Fourteen radiologists were willing to provide an estimation about the potential performance of PAM.

There is a considerable heterogeneity between and within the estimations of radiologists (Fig. 3). The estimations on the TP ranged from 50 to 292 and the TN ranged from 50 to 308. The overall mean of the TP is 217.3 ($\sigma = 12.8$) and TN is 203.2 ($\sigma = 13$). The combined distribution of all radiologists of the TP ranged from 172 to 248.6, with a mode of 220.8, and the TN ranged from 160.9 to 238.9, with a mode of 204.8. Sensitivity and specificity were derived from the TN and TP and ranged from 58.9 to 85.1 % (sensitivity), with the mode being 75.6 %, and 52.2 to 77.6 % (specificity), with the mode being 66.5 %. In Electronic Supplementary Material 5, a figure showing the distribution of TN is given.

**Table 1** Calibration factors

| Years of experience (weight 0.45) | | Average number of MR images examined per week (weight 0.45) | | Examining MR images in other areas (weight 0.1) | |
|---|---|---|---|---|---|
| $X < 3$ | 1 | $X < 5$ | 1 | $X = 0$ | 1 |
| $X \geq 3$ | 2 | $5 \leq X < 10$ | 2 | $X > 0$ | 2 |
| | | $10 \leq X$ | 3 | | |

*MR* magnetic resonance

**Table 2** Information and calibration weights of radiologists

| Expert | Academic hospital | Years of experience | Average number of MR images examined per week | Examining MR images in other areas | Calibration weight of expert for tumor characteristics | Calibration weight of expert for sensitivity and specificity | Number of publications |
|---|---|---|---|---|---|---|---|
| 1 | Yes | 5 | 6 | 4 | 0.06522 | 0.07824 | NA |
| 2 | Yes | 2 | 5 | 4 | 0.04970 | 0.05949 | 114 |
| 3 | Yes | 10 | 3 | 1 | 0.05116 | 0.06157 | 15 |
| 4 | No | 15 | 3 | 4 | 0.05116 | 0.06157 | NA |
| 5 | No | 10 | 15 | 1 | 0.07928 | 0.09491 | 8 |
| 6 | Yes | 10 | 6 | 2 | 0.06522 | 0.07824 | NA |
| 7 | Yes | 1.5 | 6 | 2 | 0.04970 | 0.05949 | 1 |
| 8 | No | 0.2 | 4 | 4 | 0.03564 | 0.04282 | NA |
| 9 | Yes | 24 | 15 | 1 | 0.07928 | 0.09491 | 7 |
| 10 | No | 15 | 5 | 0 | 0.06219 | 0.07454 | 7 |
| 11 | No | 8 | 15 | 4 | 0.07928 | 0.09491 | 11 |
| 12 | No | 1 | 5 | 4 | 0.04970 | 0.05949 | 1 |
| 13 | No | 5 | 5 | 3 | 0.06522 | 0.07824 | NA |
| 14 | No | 20 | 2 | 3 | 0.05116 | 0.06157 | NA |
| 15 | Yes | 7 | 3 | 1 | 0.05116 | NA | 1 |
| 16 | No | 18 | 10 | 4 | 0.06522 | NA | 1 |
| 17 | No | 2 | 7 | 2 | 0.04970 | NA | 7 |
| 18 | Yes | 17 | 3 | 4 | Excluded from study | Excluded from study | 1 |

*MR* magnetic resonance, *NA* not applicable

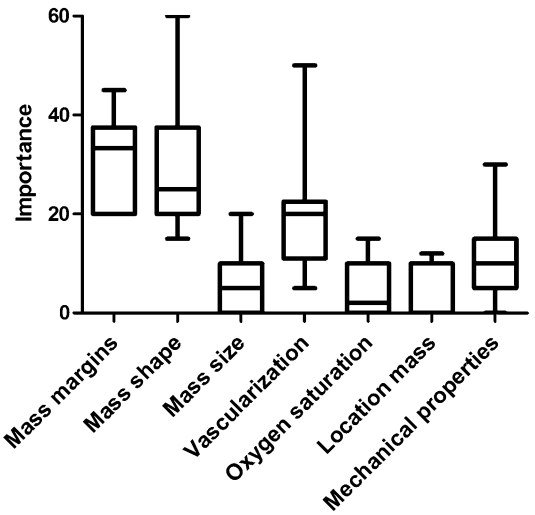


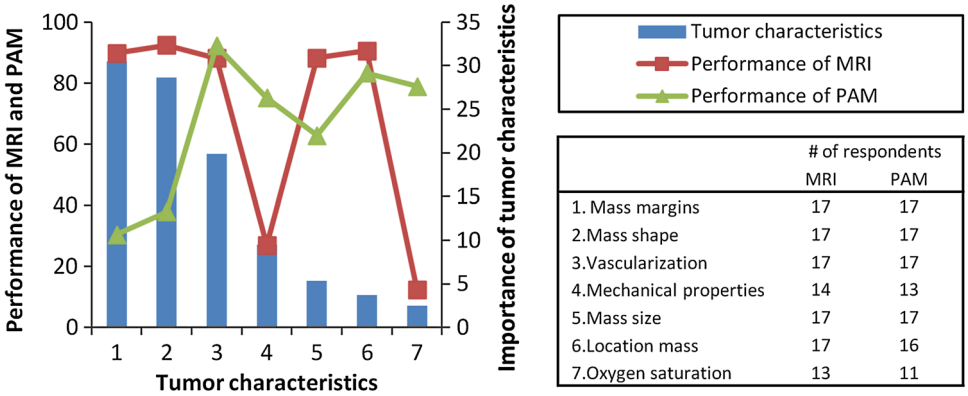

**Fig. 1** Distribution of importance of tumor characteristics indicating the mean within a 95 % confidence interval ranging from 0 (not important) to 100 (very important)

## 4 Discussion/Conclusion

Bayesian methods are broadly accepted as a method to estimate uncertain priors [6, 7, 9]. This study describes a method where belief elicitation is used to construct Bayesian priors regarding the expected diagnostic value of PAM (i.e., sensitivity and specificity) in an early stage of development, which can be used in future health economic models. This can be important to clarify the potential cost effectiveness of the new technology and its potential value in clinical practice.

According to the radiologists, the most important breast tumor characteristics to visualize during the diagnostic process are mass margins and mass shape. This is in accordance with the BI-RADS classification. The experts considered MRI (sensitivity of 90.1 % and specificity of 69.5 %) the better technology to visualize these characteristics. This was also confirmed by the elicitation of the TP and TN, with overall sensitivity of PAM ranging from 58.9 to 85.1 % with a mode of 75.6 % and specificity ranging from 52.2 to 77.6 % with a mode of 66.5 %.

During the elicitation process specific attention was paid to the reduction of clarity and ordering bias by first performing the importance ranking for specific tumor characteristics, before proceeding to the overall elicitation process regarding TP and TN. Still, radiologists indicated that they perceived the elicitation exercise to be difficult. Radiologists indicated that this had to do with the fact that PAM is an early stage technology for which only small-scale, experimental experience was available. However, it remains uncertain if this is a more general problem with providing estimations about technology performance or if it is a specific problem to the present study.




| | # of respondents | |
|---|---|---|
| | MRI | PAM |
| 1. Mass margins | 17 | 17 |
| 2.Mass shape | 17 | 17 |
| 3.Vascularization | 17 | 17 |
| 4.Mechanical properties | 14 | 13 |
| 5.Mass size | 17 | 17 |
| 6.Location mass | 17 | 16 |
| 7.Oxygen saturation | 13 | 11 |

**Fig. 2** Performance of magnetic resonance imaging (MRI) and photoacoustic mammography (PAM), and importance of tumor characteristics

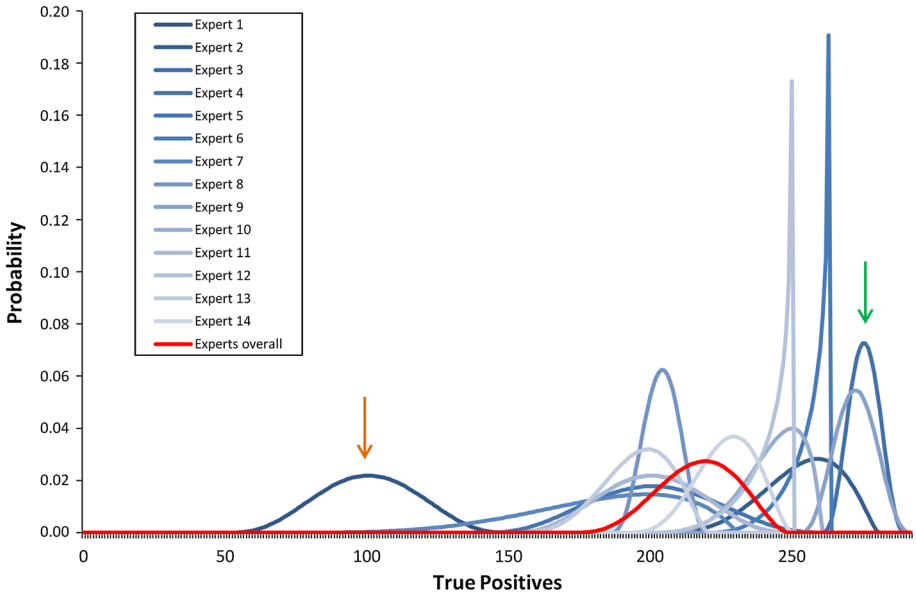



**Fig. 3** Probability distribution of estimations of true positives of 14 radiologists, where the probability ranges from 0 (it definitely will not occur) to 1 (it definitely will occur). The *green arrow* indicates an optimistic probability density function, and the *orange arrow* indicates a skeptical probability density function

It is difficult to identify which sample size is sufficient. An effective sample size of 14 could be an issue in allowing for inferences to be made from the dataset. However, in this study a Bayesian approach was used. Furthermore, it has been argued that beyond 12 experts, the marginal benefit from including more experts begins to decrease [17]. In our elicitation process a substantial heterogeneity within as well as between the estimations provided by the radiologists is revealed, which suggests that the sample size of 14 is sufficient. Within-expert estimation heterogeneity represents the uncertainty of an individual expert, and may be caused by experts' overconfidence when providing beliefs about new technologies. On the other hand, uncertainty by an individual expert can also be caused by a lack of confidence in their estimation. The observed between-expert estimation heterogeneity may be due to diversity in attitudes towards new technology. Radiologists may, for example, be too enthusiastic or skeptic about PAM [27], which leads to a lower or higher expected performance of PAM (see Fig. 3).

There are a number of issues that are worth noting. Firstly, notwithstanding our efforts to reduce bias, the degree of heterogeneity within experts' estimations in this study was quite substantial. The question is to what extent this heterogeneity is due to bias or is simply a true representation of uncertainty regarding the expected performance of such an early-stage technology. When the observed heterogeneity is mainly induced by bias in the elicitation method, this would likely mean that the true prior beliefs are more narrowly distributed, either to the advantage or disadvantage of PAM. When the heterogeneity is mainly a reflection of true uncertainty, then this is not necessarily a negative finding for PAM. As shown by Girling et al. [28], uncertainty in early stages of development may, paradoxically, enhance the future value of a technology. This is because of the flexibility offered by subsequent decision gates [28] in the development process in which (part of) the developmental uncertainty currently surrounding PAM may be solved.

Secondly, in employing methods to reduce clarity bias, the ranking and elicitation of the technologies'

performance for specific tumor characteristics was initiated. To estimate the importance, point allocation was used. In the elicitation of the performance, experts could indicate 0–100 for the performance. It is debatable which method can be used to best elicit these characteristics and whether the outcome will differ when using a different approach. However, the elicitation process of tumor characteristics is expected to prepare the radiologists with clinical concepts that resonate well within their usual clinical decision and to improve the elicitation of the TP and TN numbers and reduce bias. Through this, information that will help developers prioritize between competing technology features in subsequent development stages was also generated. The upper performance bounds provided by radiologists indicate that PAM could obtain a relatively high diagnostic performance compared with MRI, while the lower bound indicates that focusing on the wrong features may lead to a product that is less valuable to clinicians than MRI. For example, PAM was expected to outperform MRI on visualization of mechanical properties and oxygen saturation. Yet, radiologists ranked these characteristics as relatively less important so the added value to be gained from these technology features is questionable. Only if subsequent clinical studies were to show that mechanical properties and oxygen saturation are key indicators for breast cancer diagnosis would PAM be the preferred diagnostic option.

Finally the calibration method used to provide the individual experts' weights may have been flawed. Years of experience and the number of MRIs examined per week (both weight 0.45) are expected to be more important than the examination of MR images in other areas (weight 0.1). Whether this calibration method is necessary to account for the individual performance of each expert and whether the weights of the calibration aspects should be different is debatable. That said, equal weighting of the experts led in this study to similar results regarding the performance of MRI in comparison to PAM (data not shown).

Several recommendations can be made for subsequent elicitation studies that would possibly improve the validity and reproducibility of the results. These recommendations can be divided in general methodological improvements (e.g., (1) use of behavioral approaches) and improvements related to the specific case of breast cancer imaging (e.g., (2) define case mix of tumors, and (3) the use of clinical images from both imaging modalities for improved performance judgment). Regarding methodological improvements, the first recommendation is to elicit the same priors using the alternative behavioral approach, where the focus is to achieve consensus [29]. Comparing both results could provide additional information regarding the diagnostic performance of PAM and it is possible that experts may feel more confident expressing their beliefs as a consensus group rather than individually. A second improvement could be the elicitation of priors for specific tumor types taking into account case mix differences. In this study, priors were elicited for a case mix of all tumor types present in the study population, yet radiologists based their estimation often on their expectations of how PAM will visualize different specific tumor types. Thirdly, in designing further research utilizing belief elicitation in early development stages, it may be worthwhile to modify the approach used here. When better information becomes available of the images produced by PAM of different breast lesions and the related pathological findings, a detailed and exemplary clinical vignette can be developed. Different radiologists can then be asked to examine these images and to indicate the grade of the lesion. Subsequently, when more data becomes available, images obtained with PAM of different lesions could be estimated by the same radiologist.

In conclusion, we have shown that in an early stage of the development of PAM, belief elicitation provides a method to obtain estimates for the expected diagnostic performance of PAM where no other experimental evidence exists. The strength of this research is associated with combining the elicitation of the diagnostic performance of PAM, indicating the potential benefit, with the elicitation of tumor characteristics to help frame the elicitation task for the experts while supporting the developer to prioritize between competing features. The expression of uncertainty surrounding experts' beliefs reflects the infancy of the diagnostic device; however, further clinical trials should be performed to indicate to what extent this method is predictive for observed test performance. Other methods to improve validity of the elicited priors should be tested, to firmly establish if this method is indeed valuable in early stages. Until then, it must be noted that the use of the elicited priors in health economic models requires careful consideration, but that it can provide useful information to inform development decisions.

**Acknowledgments** We would like to thank the radiologists Frank van den Engh, Roland Bezooijen, and Magreet van der Schaaf for providing their input and comments. Furthermore, we would like to thank all radiologists that participated in this study. We would like to thank Srirang Manohar for providing the information about PAM.

There were no sponsors involved in this research and there is no conflict of interest. The submitted manuscript has not been published elsewhere and no funding was received.

**Authors' contribution** WH: design of study, and responsible for data collection and analysis, writing.

LB: design of study, review of expert consultation approach, review of paper.

LS: review of design and data collection, interpretation of findings, review of paper.

MIJ: initiated the study, design and study approach, review of results and paper, responsible for overall content.

## References


1. IJzerman MJ, Steuten LMG. Early assessment of medical technologies to inform product development and market access: a review of methods and applications. Appl Health Econ Health Policy. 2011;9(5):331–47.
2. Vallejo-Torres L, Steuten LMG, Buxton MJ, et al. Integrating health economics modeling in the product development cycle of medical devices: a Bayesian approach. Int J Technol Assess Health Care. 2008;24(4):459–64.
3. Garthwaite AJ, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. J Am Stat Assoc. 2005;100(470):680–701.
4. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. New York: Wiley; 2004.
5. O'Hagan A, Buck CE, Daneshkhah A, et al. Uncertain judgements: eliciting experts' probabilities (statistics in practice). Chichester: Wiley; 2006.
6. Hiance A, Chevret S, Lévy V. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. J Clin Epidemiol. 2009;62(4):431–7.
7. Johnson SR, Tomlinson GA, Hawker GA, et al. A valid and reliable belief elicitation method for Bayesian priors. J Clin Epidemiol. 2010;63(4):370–83.
8. Leal J, Wordsworth S, Legood R, et al. Eliciting expert opinion for economic models: an applied example. Val Health. 2007;10(3):195–203.
9. Bojke L, Claxton K, Bravo-Vergel Y, et al. Eliciting distributions to populate decision analytic models. Val Health. 2010;13(5):557–64.
10. 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(19):2363–80.
11. Johnson SR, Tomlinson GA, Hawker GA, et al. Methods to elicit beliefs for Bayesian priors: a systematic review. J Clin Epidemiol. 2010;63(4):355–69.
12. Jose J, Manohar S, Kolkman RG, et al. Imaging of tumor vasculature using Twente photoacoustic systems. J Biophotonics. 2009;2(12):701–17.
13. Piras D, Wenfeng X, Steenbergen W, et al. Photoacoustic imaging of the breast using the Twente Photoacoustic Mammoscope: present status and future perspectives. IEEE J Sel Top Quant. 2010;16(4):730–9.
14. Mammacarcinoom: Landelijke richtlijn met regionale toevoegingen. Versie 1.1 Nationaal Borstkankeroverleg Nederland (NABON), Vereniging voor Integrale Kankercentra (VIKC), Kwaliteitsinstituut voor de Gezondheidszorg CBO, Amsterdam, The Netherlands; 2008. http://www.oncoline.nl/uploaded/FILES/mammacarcinoom/Richtlijn%20Behandeling%20van%20het%20Mammacarcinoom%20oktober%202005.pdf
15. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol. 2011;62(1):451–82.
16. Cooke R. Experts in uncertainty: opinion and subjective probability in science. Oxford: Oxford University Press; 1991.
17. Knol A, Slottje P, van der Sluijs J, et al. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environ Health. 2010;9(1):19.
18. ACR BI-RADS- MRI Lexicon Classification Form. Reston: American College of Radiology; 2003.
19. Gibbs P, Liney GP, Lowry M, et al. Differentiation of benign and malignant sub-1 cm breast lesions using dynamic contrast enhanced MRI. Breast. 2004;13(2):115–21.
20. Nunes LW, Schnall MD, Orel SG. Update of breast MR imaging architectural interpretation model. Radiology. 2001;219(2):484–94.
21. Bluemke DA, Gatsonis CA, Chen MH, et al. Magnetic resonance imaging of the breast prior to biopsy. JAMA. 2004;292(22):2735–42.
22. Bone B, Aspelin P, Bronge L, et al. Sensitivity and specificity of MR mammography with histopathological correlation in 250 breasts. Acta Radiol. 1996;37(2):208–13.
23. Peterson C, Miller A. Mode, median, and mean as optimal strategies. J Exp Psychol. 1964;68(4):363–7.
24. van Dorp RJ, Kotz S. A novel extension of the triangular distribution and its parameter estimation. J R Stat Soc Ser D-Sta. 2002;51(1):63–79.
25. Miglioretti DL, Gard CC, Carney PA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. Radiology. 2009;253(3):632–40.
26. Liberman L, Benton CL, Dershaw DD, et al. Learning curve for stereotactic breast biopsy: how many cases are enough? Am J Roentgenol. 2001;176(3):721–7.
27. Spiegelhalter DJ, Myles JP, Jones DR, et al. An introduction to bayesian methods in health technology assessment. BMJ. 1999;319(7208):508–12.
28. Girling A, Young T, Brown C, et al. Early-stage valuation of medical devices: the role of developmental uncertainty. Val Health. 2010;13(5):585–91.
29. Hilgerink MP, Hummel MJM, Manohar S, et al. Assessment of the added value of the Twente Photoacoustic Mammoscope in breast cancer diagnosis. Med Devices. 2011;4:107–15.