ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 355−369

Journal of Clinical Epidemiology

BAYESIAN PRIORS

Methods to elicit beliefs for Bayesian priors: a systematic review

Sindhu R. Johnson[a,b,*], George A. Tomlinson[b,c,d], Gillian A. Hawker[b,e], John T. Granton[f], Brian M. Feldman[b,c,g]

[a]*Division of Rheumatology, Department of Medicine, University Health Network, Toronto, Ontario, Canada*
[b]*Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*
[c]*Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*
[d]*Division of Clinical Decision Making and Health Care, Toronto General Research Institute, Toronto, Ontario, Canada*
[e]*Division of Rheumatology, Department of Medicine, Women's College Hospital, Toronto, Ontario, Canada*
[f]*Divisions of Respirology and Critical Care Medicine, Department of Medicine, University Health Network, Toronto, Ontario, Canada*
[g]*Division of Rheumatology, Department of Paediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada*

Accepted 9 June 2009

Abstract

Objective: Bayesian analysis can incorporate clinicians' beliefs about treatment effectiveness into models that estimate treatment effects. Many elicitation methods are available, but it is unclear if any confer advantages based on principles of measurement science. We review belief-elicitation methods for Bayesian analysis and determine if any of them had an incremental value over the others based on its validity, reliability, and responsiveness.

Study Design and Setting: A systematic review was performed. MEDLINE, EMBASE, CINAHL, Health and Psychosocial Instruments, Current Index to Statistics, MathSciNet, and Zentralblatt Math were searched using the terms (prior *OR* prior probability distribution) *AND* (beliefs *OR* elicitation) *AND* (Bayes *OR* Bayesian). Studies were evaluated on: design, question stem, response options, analysis, consideration of validity, reliability, and responsiveness.

Results: We identified 33 studies describing methods for elicitation in a Bayesian context. Elicitation occurred in cross-sectional studies ($n = 30$, 89%), to derive point estimates with individual-level variation ($n = 19$; 58%). Although 64% ($n = 21$) considered validity, 24% ($n = 8$) reliability, 12% ($n = 4$) responsiveness of the elicitation methods, only 12% ($n = 4$) formally tested validity, 6% ($n = 2$) tested reliability, and none tested responsiveness.

Conclusions: We have summarized methods of belief elicitation for Bayesian priors. The validity, reliability, and responsiveness of elicitation methods have been infrequently evaluated. Until comparative studies are performed, strategies to reduce the effects of bias on the elicitation should be used.

Keywords: Belief elicitation; Bayesian; Validity; Reliability; Bias; Priors

1. Background

Bayesian analysis is an increasingly common method of statistical inference used in clinical research [1]. Within this statistical inferential paradigm, there are different schools of thought among statisticians who use a Bayesian approach [2]. The empirical Bayesian approach is one where parameters of the prior distribution are estimated by using the same data used in the main analysis. When no prior information is available, investigators use a vague prior so that new data will dominate. The fully Bayesian approach is one that considers all sources of preexisting knowledge admissible for the analysis. One advantage of the fully Bayesian approach over the traditional ''frequentist'' approach to statistical inference or the empirical Bayesian approach is the ability to incorporate beliefs into models that estimate treatment effects. Once beliefs are elicited from a sample (e.g., experts in a field), the elicited beliefs (e.g., regarding the probability of a treatment effect) can be graphically expressed as a prior probability distribution. This distribution can be used to document clinical equipoise (a prerequisite for clinical trials) [3], for sample size calculation [3], interim study monitoring [3,4], and can be incorporated with treatment effect estimates obtained from trials [5]. In a fully Bayesian analysis, when

* Corresponding author. Division of Rheumatology, University Health Network, Ground Floor, East Wing, Toronto Western Hospital, 399 Bathurst Street, Toronto, Ontario M5T 2S8, Canada. Tel.: +416-603-6417; fax +416-603-4348.
E-mail address: Sindhu.Johnson@uhn.on.ca (S.R. Johnson).

0895-4356/10/$ − see front matter
doi: 10.1016/j.jclinepi.2009.06.003

What is new?

What this adds to what was known?

- This article summarizes methods that have been applied for belief elicitation;
- Reviews the published measurement properties of each method;
- Presents a conceptual framework for the belief-elicitation process;
- Identifies pragmatic methodologic strategies to reduce the effect of bias in belief-elicitation studies.

What should change now?

- Strategies to reduce the effect of bias include sampling from groups of experts, use of clear instructions and a standardized script, provision of examples and training exercises, avoidance of scenarios or anchoring data, provision of feedback and opportunity for revision of the response, and use of simple graphical methods.

no prior information is available, investigators will use a vague prior so that the new data will dominate.

''Prior belief'' is often a combination of fact-based knowledge with subjective impressions based on clinical experience.[6] Critics of use of the fully Bayesian paradigm in clinical trials are concerned that the inclusion of prior belief is too subjective [7] and lacking in methodologic rigor [6]. Bayesian methodologists have been challenged to take a ''stand for disciplined research methodology''[6]. Therefore, to apply Bayesian prior probability distributions of existing belief about a treatment effect in clinical trials, clinical researchers would benefit from knowledge of existing belief-elicitation methods and identification of methods that have demonstrable methodologic rigor. In particular, belief-elicitation methods should be valid, reliable, responsive to change, and feasible. Thus, the primary objectives of this study were: (1) to review methods of eliciting prior beliefs for a Bayesian analysis; and (2) to review the measurement properties (validity, reliability, responsiveness, and feasibility) of these methods to determine if one method had incremental value over another. To better understand the processes by which experts formulate a belief, as well as the processes by which investigators can elicit this belief, and the potential biases that may affect the validity, reliability, and responsiveness of these methods, the secondary objectives of this study were: (1) to develop a conceptual framework for the belief–elicitation process and biases that may affect the elicited response through review of the literature; and (2) to identify methodologic strategies that may reduce the effect of bias on elicitation process.

2. Methods

2.1. Search strategy

Eligible studies were identified using MEDLINE (1950 to week 2, June 2008), EMBASE (1980 to week 25, 2008), CINAHL (1982 to week 2, June 2008), Health and Psychosocial Instruments (1985 to March 2008), Current Index to Statistics (1974 to June 2008), MathSciNet (1940 to June 2008), and Zentralblatt Math (1868 to June 2008) using the search terms (prior *OR* prior probability distribution) *AND* (beliefs *OR* elicitation) *AND* (Bayes *OR* Bayesian). Mapping of term to subject heading was used, where appropriate. Titles and abstracts were screened to exclude ineligible studies. Included studies were entered in the Science Citation Index and PUBMED (with use of the ''related articles'' tool) to search for other potentially eligible studies. In addition, the bibliographies of included studies and published reviews were searched.

2.2. Inclusion and exclusion criteria

Eligible articles included published observational studies, randomized controlled trials, book chapters, and technical reports, which describe elicitation of beliefs in a Bayesian context. Studies using human and nonhuman subjects were included. Non–English language studies were excluded.

2.3. Data abstraction and methodologic assessment

Using a standardized form, the following data were abstracted: sample size, study design (cross-sectional, longitudinal, unspecified), level of elicitation (individual, group), questionnaire-administration format (in person, telephone interview, mail, Delphi consensus, other), questionnaire format (article, computer assisted, other), question format (scenario with/without data provided in stem, predictive question, both, other), response options (visual analog scale, distribution of probabilities or proportions into bins, other), response rate (percentage, not specified, not applicable [methodologic or simulation papers]), analysis (point estimate with group-level variation, point estimate with individual-level variation), and graphical display (none, probability density function, cumulative distribution function, other). Often respondents are asked to make a probability estimate for an event which is not definitively known (e.g., probability of survival at 3 years). There may be some uncertainty around the reported point estimate. ''Group-level variation'' was used to characterize analyses that reported the variability for the groups' point estimate. ''Individual-level variation'' was used to characterize analyses that reported the variability around the point estimate for each individual study participant.

2.4. Measurement properties

Articles describing elicitation methods were evaluated for consideration of the following properties:

1. *Validity. Face validity* evaluates if the elicitation method appears to measure what it purports to measure. *Content validity* evaluates if the elicitation method captures all the relevant aspects of the belief [8,9]. *Criterion validity* evaluates the correlation of an elicitation method with the "gold standard." Under the assumption that there is no gold standard for the truth or belief, *construct validity* evaluates the relationship of two different methods of measuring the same belief. *Convergent construct validity* evaluates the correlation between two related aspects of the elicited belief, whereas *divergent construct validity* evaluates the ability of an elicitation method to correctly distinguish between dissimilar beliefs [10].
2. *Reliability.* Reliability refers to the reproducibility of the measure. Intrarater reliability is evaluated when the elicitation method is applied to the same participant(s) on two different occasions, whereas interrater reliability is evaluated when the elicitation method is applied to different participants on the same occasion. In the context of belief measurement, interrater reliability is of lesser importance. Measures of reliability include the method of Bland and Altman, intraclass correlation coefficient, or Cohen's kappa [10].
3. *Responsiveness* refers to the ability of an elicitation method to accurately detect a meaningful change in belief over time when it has occurred [10,11]. Measures of responsiveness may include Cohen's effect size or the standardized response mean [10].
4. *Feasibility* refers to the ease of usage of the elicitation method [12]. Determinants of feasibility include time, cost, and need for equipment or personnel.

Consideration of validity, reliability, responsiveness, and feasibility by investigators was categorized as commented on, evaluated (measure of association or change recorded), or not specified. The measures of validity, reliability, and responsiveness cited earlier (e.g., correlations, kappa) are appropriate when elicitation yields a single value per respondent. When each respondent provides an entire probability distribution, it is not clear how validity, reliability, and responsiveness should be measured.

2.5. Statistical analysis

Summary statistics were calculated using R 2.4 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1. Search strategy

Systematic review of the literature identified 33 articles which described unique methods for belief elicitation in a Bayesian context (Fig. 1).

3.2. Study characteristics

Table 1 summarizes the study characteristics. Belief elicitation mostly occurred in cross-sectional studies (91%), at the level of the individual (97%), using small sample sizes (median of 11 participants). Questionnaires were largely administered in person (58%) or on paper (52%), and to derive a point estimate with individual-level variation (58%).

3.3. Elicitation methods

Question stems (the question asked of the participant) and response options are summarized in Table 2. Investigators had asked participants about the mean [32,33,40], median [27,30,45] and mode [16,20,37] for a parameter. Participants had been asked to estimate the probability of an outcome/event [14,15,19,24,28,42–44,46], the proportion of individuals who will have an outcome [3,17,47], the relative risk of an outcome [35,36], the value for a dependent variable given specified values for independent variables [22,23], and their weight of belief [5,38,39,41]. Commonly used response options include direct probability estimates [4,15], visual analog scale [18,29,42,43], sketching of a graph [16,20], and use of "bins and chips" (participants are asked to put the weight of their belief expressed as percentages into discrete intervals [Fig. 2]) [25,38,41,48]. Methods used to illustrate the elicited beliefs include line graphs [3,29], histograms [13,35], probability density functions [3,17,25,32,38,41], and cumulative distribution functions [18,21,40].

3.4. Measurement properties

Of the identified studies, 64% (21 of 33) considered the validity, 24% (8 of 33) the reliability, 12% (4 of 33) the responsiveness, and 55% (18 of 33) the feasibility of the elicitation methods (Table 1). However, only four (12%) studies formally evaluated validity, two (6%) studies tested reliability, none tested responsiveness, and one (3%) study formally evaluated feasibility (Table 3).

3.5. Conceptual framework for belief formulation and elicitation

The formulation of a clinical belief, and the subsequent elicitation of the belief, is a complex process. Based on the literature [3,4,14,19,30–32,34,44,49,53–56], we have developed a conceptual framework for this process (Fig. 3). An individual's belief about the effectiveness of an intervention is influenced by his or her knowledge of the research evidence and his or her clinical experience, which are presumably both approximations of the truth. Some schools of thought suggest that an individual does not have a preexisting quantification of his or her belief "ready for the picking" [44]. Rather, when asked about his or her belief about an intervention, an individual will synthesize his or her knowledge and experience into a "quantified belief prior" [44]. Using an elicitation procedure (question and response option), the investigator tries to elicit the belief.

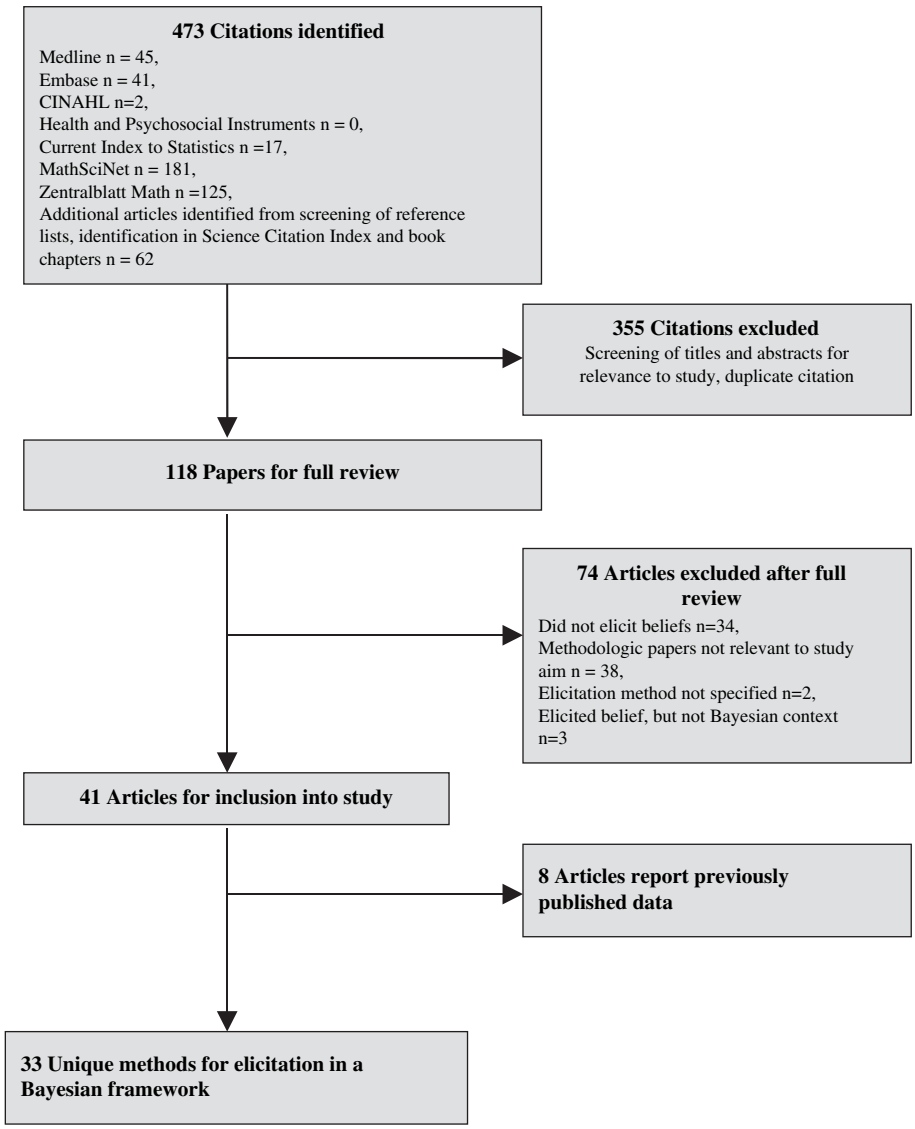

Fig. 1. Flow diagram of systematic review results.

The investigator may quantify the elicited belief, express it graphically, and then combine multiple individual priors to form a group "clinical prior" [53], which reflects a spectrum of beliefs on the subject.

Using the personalistic theory of probability, all self-consistent or coherent beliefs are admissible in a study as long as the individual feels that they correspond with his judgment [44,56]. The *elicitation procedure*, the manner in which the belief is elicited, can influence the creation of both the individual's quantified prior and the group's clinical prior [44]. A person may modify the reporting of his or her quantified belief depending on the method by which the belief was elicited. Biases that may threaten the validity of the elicited belief are summarized in Table 4 [32].

The reliability, responsiveness, and feasibility of an elicitation procedure are also important determinants of its utility. Threats to the reliability of an elicitation procedure include *lack of understanding* of the elicitation procedure, *carelessness*, *lack of interest*, and *fatigue* [49]. In the setting of a longitudinal study, an elicitation procedure should also be able to detect any important changes in belief that occur over time as new information is gained. Finally, the implementation of an elicitation method in clinical research is constrained by factors that affect its feasibility. Factors may include *costs* incurred through implementation of the method, need for *specialized personnel* or *hardware*, and the *time* required of the study participant.

3.6. Methodologic strategies to reduce bias

Methodologic strategies to reduce the influence of potential biases on the validity and reliability of elicitation methods are summarized in Table 4. Strategies to minimize bias can be implemented at each stage of the elicitation procedure: identification of the sample, framing of the question, choice of the response option, and summarizing of the data.

Table 1
Summary of study characteristics

Study characteristics	Number (%) ($N = 33$)
Article	
Methodological	4 (12)
Applied	26 (79)
Both methodological and applied	3 (9)
Study design	
Study design	
Cross-sectional study	30 (91)
Longitudinal study	2 (6)
Not applicable	1 (3)
Level of elicitation	
Individual	32 (97)
Small group	0 (0)
Not applicable	1 (3)
Use of consensus methods	4 (12)
Sample	
Sample size median (range)	11 (1–298)[a]
Questionnaire	
Format	
Paper	17 (52)
Computer	7 (21)
Combined	1 (3)
Other	3 (9)
Not specified	5 (15)
Administration	
In person	19 (58)
Telephone	2 (6)
Mail	7 (21)
Combined	1 (3)
Not specified	3 (9)
Not applicable[b]	1 (3)
Response rate	
Rate median (range)	100% (50–100)[c]
Not specified	10 (30)
Analysis	
Level of analysis	
Point estimate with group-level variation	8 (24)
Point estimate with individual-level variation	19 (58)
Other	6 (18)
Measurement properties[d]	
Consideration of validity	21/33 (64)
Consideration of reliability	8/33 (24)
Consideration of responsiveness	4/33 (12)
Consideration of feasibility	18/33 (55)

[a] Excluding studies where $n = 0$ or not specified.
[b] Belief elicitation was conducted in hypothetical participants.
[c] Excluding studies where $n = 0$ or 1.
[d] Each measurement property may occur more than once.

3.6.1. The sample

The inclusion of clinical experts [40] rather than generalists in an elicitation procedure improves the validity and reliability of the elicited beliefs for a number of reasons [22,30,50,51]. The training of a clinical expert generally extends over a period of time—years rather than weeks. During that time, the expert gains extensive experience with the specific events in question and with the factors that affect them [52]. An expert encounters the condition in a repetitive manner and receives relatively immediate feedback for the consequences of their therapeutic decisions [52]. Thus, an expert is one who has thought more deeply and over a longer period of time about the subject than others have [32]. As a result, experts are able to predict events about which they have special training, and tend to be more consistent in their beliefs than nonexperts [51,52]. Overconfidence, which underestimates realistic doubt [26], occurs among inexperienced individuals. Experienced individuals are more willing to admit to uncertainty [26,54]. Inexperienced individuals tend to overuse round numbers and label events as impossible rather than assign small probabilities to them [44]. This results in the elicited probability distributions being truncated at hard and perhaps unrealistic boundaries rather than extending to include extreme tail areas with very small probabilities [44,56]. Clinical experience reduces these tendencies [44].

3.6.2. The question

Investigators have asked participants about measures of central tendency [16,20,27,30,32,33,37,40,45], probability [14,15,19,24,28,42–44,46], proportion [3,17,47], relative risk [35,36], value for a dependent variable given specified values for independent variables [22,23], and their weight of belief [5,38,39,41]. Insufficient normative goodness (statistical understanding) and insufficient understanding of the elicitation question threaten the validity of the belief elicited [44]. Strategies that have been shown to decrease the influence of these biases include the provision of an *example* [5,28,41] or *training exercises* [44,45]. Study participants have reported that examples are helpful [28]. A training exercise improves both normative goodness [49] and reliability [44,54,56], and thus, has been recommended [15,34,37]. Other strategies to improve reliability include the use of clear instructions [34] and standardized script [17].

Investigators have provided a summary of research data [18,40,47] or a scenario [14,35,46] with the elicitation question. Although this may have the advantage of preventing a radical opinion [4], it may result in anchoring bias where their reported belief is influenced by the data [14]. Study participants give explicit attention to data to which they have been cued [19]. Strategies to reduce anchoring bias include avoidance of data presentation or scrambling the sequence of data presentation between participants [37].

3.6.3. The response option

The use of a dichotomous response option (e.g., I believe this intervention is effective. Yes/No) has insufficient content validity, as clinicians often have beliefs about the magnitude of the effect and varying degrees of certainty in the strength of their belief [28,57]. A software for belief

Table 2
Summary of elicitation methods

Authors	Question	Response option
Errington et al., 1991 [13]; Abrams et al., 1994 [67]	(a) Express your belief about neutron therapy compared with an expected 12-month failure rate of 50% in the photon arm of the trial	Given 20 counters, place 2 of them at the upper and lower limits of belief. Place the remaining 18 counters so as to express their remaining prior beliefs about the neutron failure rates
Bergus et al, 1995 [14]	(a) Estimate the probability of 3 diagnostic alternatives (b) Given additional information, give the post test probability of 3 diagnostic alternatives (c) Estimate the false negative rate and true negative rate of a normal CT scan (d) Estimate final probability estimates for the 3 diagnoses	Specify values
Chaloner et al., 1993 [15], Carlin et al., 1993 [4]—modified from Freedman and Spiegelhalter, 1983 [16]	(a) Estimate the probability of experiencing toxoplasmosis within 2 years of treatment on placebo, clindamycin, and pyrimethamine respectively (b) Guess the upper and lower quartiles of the probability's distribution	Probability on placebo = *X%* Probability on clindamycin = *Y%* Probability on pyrimethamine = *Z%*
Chaloner, 1996 [17]—modified from Chaloner et al., 1993 [15]	(a) What is your best guess of the percentage of people assigned to daily trimethoprim-sulfamethoxazole (TMS) group who will experience pneumocystitis pneumonia (PCP) 2 years after enrollment? (b) Think about the people on thrice weekly arm and think about an interval estimate for what you would expect for the percentage of people on the thrice weekly TMS arm who will experience PCP in 2 years given that the proportion experiencing PCP on the daily TMS arm is what you guessed. Please specify the interval by an upper and lower number within which you think that the percentage of people experiencing PCP on the 3 times a week arm will lie	(a) *X%* (b) *Y%* and interval
Chaloner and Rhame 2001 [3]—modified from Chaloner, 1996 [17]	(a) What is your estimate of the percent of subjects randomized to daily TMS who will experience PCP during the 2 years after entry? (b) What is your estimate of the percent of subjects randomized to thrice weekly TMS who will experience PCP during the 2 years after entry? (c) Write down the difference between the two estimated percents (d) What is your estimate of the 95% probability interval of this difference?	(a) *X%* (b) *Y%* (c) *X% – Y%* (d) 95% probability interval from — to —
de Vet et al., 1993 [18]	State belief about the hypothesis, "A high intake of beta-carotene protects against cervical cancer"	10-cm VAS: 0–100%
Dumouchel, 1988 [58]	(a) Specify parameters to be assessed and range for each parameter (b) Specify the log relative risk and uncertainty	Specify values
Evans et al., 2002 [19]	(a) 40% of students are in the Engineering faculty. What is the probability that a member of the Drama society is also in the Engineering faculty?	(a) *X%*
Freedman and Spiegelhalter, 1983 [16], 1986 [26], Spiegelhalter et al., 1993 [20][a]	(a) What is the most likely level of improvement to be gained from Thiopeta? (b) Choose upper and lower bounds which are very unlikely to be exceeded. (c) Define very unlikely (d) Estimate the chance of exceeding intermediate points	(a) Point estimate(b), (c), and (d) Sketch graph
Flournoy, 1994 [21]	(a)Sketch a 95% probability interval for the dose response curve	Graph with probability of death 0–100% on vertical axis, and medication dose 20–240 mg/kg on horizontal axis
Garthwaite and Dickey, 1991 [22]	(a) Specify name and range for independent variables (b) Estimate experimental error (c) Estimate parameters	Specify values

Garthwaite and Dickey, 1992 [23]	(a) Specify name and range for independent variables (b) Estimate experimental error (c) Estimate parameterss	Specify values
Gustafson et al., 2003 [24]	(a) Suppose you were asked to predict whether a project would be successfully implemented. You can ask me any question you want about the project and I will find the answer for you. What questions would you ask of me? (b) Please give me examples of answers that would make you optimistic and pessimistic about the chances of success (c) Estimate the prior probability of implementation success using an "estimate–talk–estimate" approach	Specify parameters and estimates
Hughes, 1991 [25]—based on Spiegelhalter and Freedman, 1986 [26]	(a) Define the lower and upper extremes of belief in relative reduction/increase in mortality. (b) Place an adhesive dot above the most likely value and then add 19 stickers to indicate their beliefs for the outcome of the trial	Graph with adhesive dots simulating a histogram
Hutton and Owens, 1993 [27]	(a) Estimate the minimum, lower quartile, median, upper quartile, and maximum prevalence of child abuse in children under the age of 10 years	Specify values
Johnson et al., 2006 [28]	(a) Please give your best estimate of the relative probability of pregnancy in the 6 months following a lipiodal hysterosalpingogram, compared with "no intervention" probability of pregnancy being 1.0 (b) Please give 95% confidence limits to this estimate. (c) What is the minimum relative probability of pregnancy following a lipiodol hysterosalpingogram that would justify, in your opinion, this being used as a standard for some women with unexplained fertility?	(a) Relative probability = X (b) Lower limit = Y, upper limit = Z (c) Relative probability
Jones et al., 1998 [29]—based on de Vet et al., 1993 [18]	Estimate degree of belief that magnesium sulfate is effective in eclampsia before and after publication of trial results	Linear analog 10-cm scale
Kadane et al., 1980 [68]	(a) Identify factors associated with fatigue cracking (b) Estimate the predictive distribution of the dependent variable given fixed values of the independent variables	Specify values
Kadane, 1986, 1994 [30,31]	(a) In a patient with this set of characteristics, which therapy would you choose? (b) In a patient with this set of characteristics, estimate the median, 75th and 90th percentile of the dependent variable on each therapy	(a) X or Y
Kadane, 1992 [69]	(a) How did you vote in the first ballot? (b) What was the distribution of the votes on the first ballot?	Specify values
Kadane and Wolfson, 1998 [32]	(a) Estimate the prior mean (b) Estimate the degrees of freedom parameter (c) Specify the range of each of the covariates (d) Specify the 50th, 75th, and 90th percentiles of y for each vector x	Specify values
Lehmann and Goodman, 2000 [33]	(a) Specify mean difference between 2 therapies and 95% Bayesian confidence interval	Specify values
Li and Krantz, 2005 [34]	(a) What is your guess of the percentage of the 758 "first words" in this particular edition of "Of Human Bondage" that have six or more letters? (b) Imagine you were allowed to draw a sample of 10 randomly selected first words out of 758 pages. What weight (in decimal numbers) do you assign to a random sample of 10? (c) What weight do you assign to the data if you were allowed to randomly select a larger sample of 50 pages from a total of 758?	(a) The percentage is X% (b) My weight placed on a sample of 10 is — (c) My weight placed on a sample of 50 is —

(Continued)

Table 2
(Continued)

Authors	Question	Response option
Lilford, 1994 [35]	(a) What is the relative risk of permanent morbidity likely to be in a hypothetical and infinitely large randomized trial of similar patients? (b) What would you consider a surprisingly good or bad result in a hypothetical trial?	Analog dial 1 = no difference between immediate delivery; 0.5 = chance of morbidity is halved by immediate delivery; 2 = chance of morbidity is doubled
Lilford and Braunholtz, 1996 [36]	(a) Estimate relative risk (b) Estimate a 95% credible interval for the relative risk	Specify values
O'Hagan, 1998 [37]	(a) Specify upper (U) and lower (L) bounds for a quantity (b) Specify the mode (M) (the most likely value) Give probabilities for the following intervals: (c) L,M (d) L, (L + M)/2 (e) (M + U)/2, U (f) L, (L + 3M)/4 (g) (3M + U)/4, U	Specify values
Parmar et al., 1994, 2001 [38,39]	We are interested in your expectations of the difference in 2 year which might result from using CHART rather than the standard radical radiotherapy for eligible patients. Enter your weight of belief in each of the possible intervals. The stronger you believe that the difference will truly lie in a given interval the greater should your weight for that interval. If you believe that it is impossible that the difference lie in a given interval your weight should be zero. Your weights should add up to 100	*X*% entered in boxes
Ramachandran, 2001 [40]	Specify the distribution, mean and relative standard deviation or lower and upper bound of distribution for each parameter	Specify values
Tan et al., 2003 [41]—modified from Parmar et al., 2001 [39]	We are interested in your expectations of the difference in 2 year survival rate which might result from using treatment X rather than the standard Y for eligible patients. Enter your weight of belief in each of the possible intervals. The stronger you believe that the difference will truly lie in a given interval the greater should your weight for that interval. If you believe that it is impossible that the difference lies in a given interval your weight should be zero. Your weights should add up to 100	*X*% entered in boxes
Ten Centre Study Group, 1987 [70]	(a) Estimate the percentage reduction in mortality of artificial surfactant in babies of 25 to 29 weeks gestation.	Specify values.
Van Der Wilt et al., 2004 [42]; Rovers et al., 2005 [43]	Estimate the probability of complete hearing recovery and normal language recovery within a year, in a situation without treatment and in a situation with ventilation tube insertion	VAS (10 cm): 0–100%
White et al., 2005 [5]—modified from Parmar	We are interested in your expectations of the difference in rates of death or hospitalization which might result from using treatment X rather than the standard Y for eligible patients. Enter your weight of belief in each of the possible intervals. The stronger you believe that the difference will truly lie in a given interval the greater should your weight for that interval. If you believe that it is impossible that the difference lies in a given interval your weight should be zero. Your weights should add up to 100. Suppose the annual event rate on placebo is 18%, what is your expectation for the annual event rate on X?	*X*% entered in boxes

Winkler, 1967 [44]	Cumulative distribution function: (a) What is the probability that a random student at the university is male? (b) Can you determine a point such that it is equally likely that p is less than or greater than this point? (c) Now suppose that you were told that p is less than I2. Determine a new point such that it is equally likely that p is less than or greater than this point (d) Now suppose that you were told that p is less than I3. Determine a new point such that it is equally likely that p is less than or greater than this point. Probability density function: (a) What do you consider the most likely value of p? (b) Can you determine 2 values of p (one on each side of p) which are about half as likely as the value in a? (c) Can you determine a point such that 1/2 the area under the graph of the density function is to the left of the point and half of the area is to the right of the point? (d) Such that 1/4 of the area is to the left of the point and 3/4 is to the right? (e) Such that 3/4 of the area is to the left of the point and 1/4 is to the right? (f) Such that 1/100 of the area is to the left of the point and 99/100 is to the right? (g) Such that 99/100 of the area is to the left of the point and 1/100 is to the right?	(a) p = *A*% (b) I2 = *B*% (c) I3 = *C*% (d) I4 = *D*%

Questions have been paraphrased for space.
Abbreviations: CT, computed tomography; VAS, visual analog scale.
[a] Used same method.

elicitation has been developed [28,57], and some studies have been computer assisted [22,33,35,37].

Strategies can be used to reduce the threat to the validity and reliability of the elicited belief of limited normative goodness, or the respondents' insufficient understanding of the elicitation procedure. *Provision of feedback* to the participant about the elicited belief allows for self-correction [30], and has been shown to improve probability assessment [46] and reliability [34]. An opportunity for *verification and revision* of the elicited response allows the participant to detect and revise inconsistencies in their response [37,47,56]. The use of a response option that requires *betting or utilizes penalties* also improves validity and reliability. Participants will reflect more deeply when provided a disincentive, as there is a sense of potential loss associated with their response (e.g., an approach where a study participant has to wager his own money based on his assessed probability of an outcome) [56]. Bias introduced by base-rate neglect (which occurs when participants fail to take account of the prevalence of the outcome among untreated patients) may be reduced by asking the participant to state the baseline rate or describe the outcome in both untreated and treated patients [19].

3.6.4. Aggregation of data

There are a variety of methods by which individual priors are aggregated to form a group clinical prior. Although some studies have used consensus methods to derive a group clinical prior [13,21,24,45,47], most studies have combined individually elicited priors. Biases introduced by overoptimism or overconfidence may be reduced by the use of *averaging methods* for the group clinical prior [26]. Methods for pooling priors have been proposed [30,49,59]. It has also been suggested that the elicited belief could be weighted by occupation, level of experience, self-confidence, or other personal characteristics [4]. However, the value of these pooling and weighting methods remains uncertain and requires evaluation.

Graphical presentation of the combined clinical prior has been used to express the degree of variability of the elicited belief, illustrate the existence of clinical uncertainty, and demonstrate the amount of evidence that would be required from data to convince optimistic and skeptical clinicians. In general, people more easily comprehend normal distributions than fractiles, relative densities, or cumulative distribution functions [44]. A probability density function is more intuitive than a cumulative distribution function, and its use is associated with improved feasibility and validity [44]. The use of a concomitant histogram is useful for individuals who are less familiar with probability distributions. The use of simple graphical representations is preferred as the trade-off of more information is busier figures where patterns are harder to see [31].

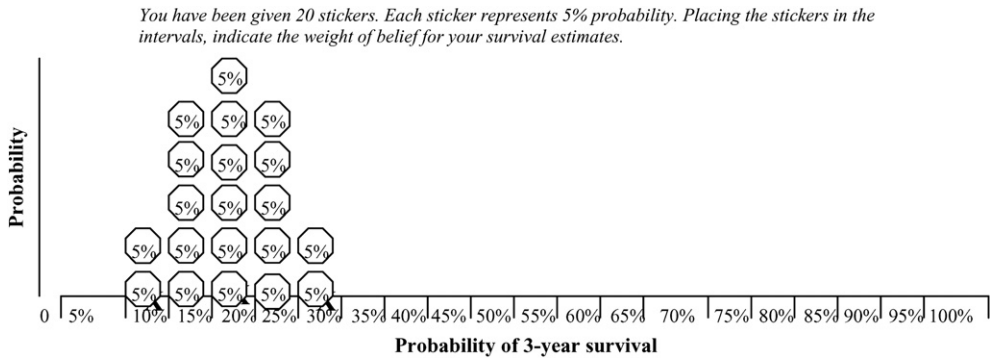

Fig. 2. Example of a "bins and chips" belief-elicitation method.

4. Discussion

This systematic review summarizes methods of belief elicitation for use in a Bayesian analysis. The validity, reliability, and responsiveness of the methods have not been adequately evaluated. Identification of the "best" method based on the principles of measurement science is limited by the paucity of data. With the increasing use of Bayesian analysis in clinical research [1], evaluation of the measurement properties of elicitation methods is required in order for researchers to be

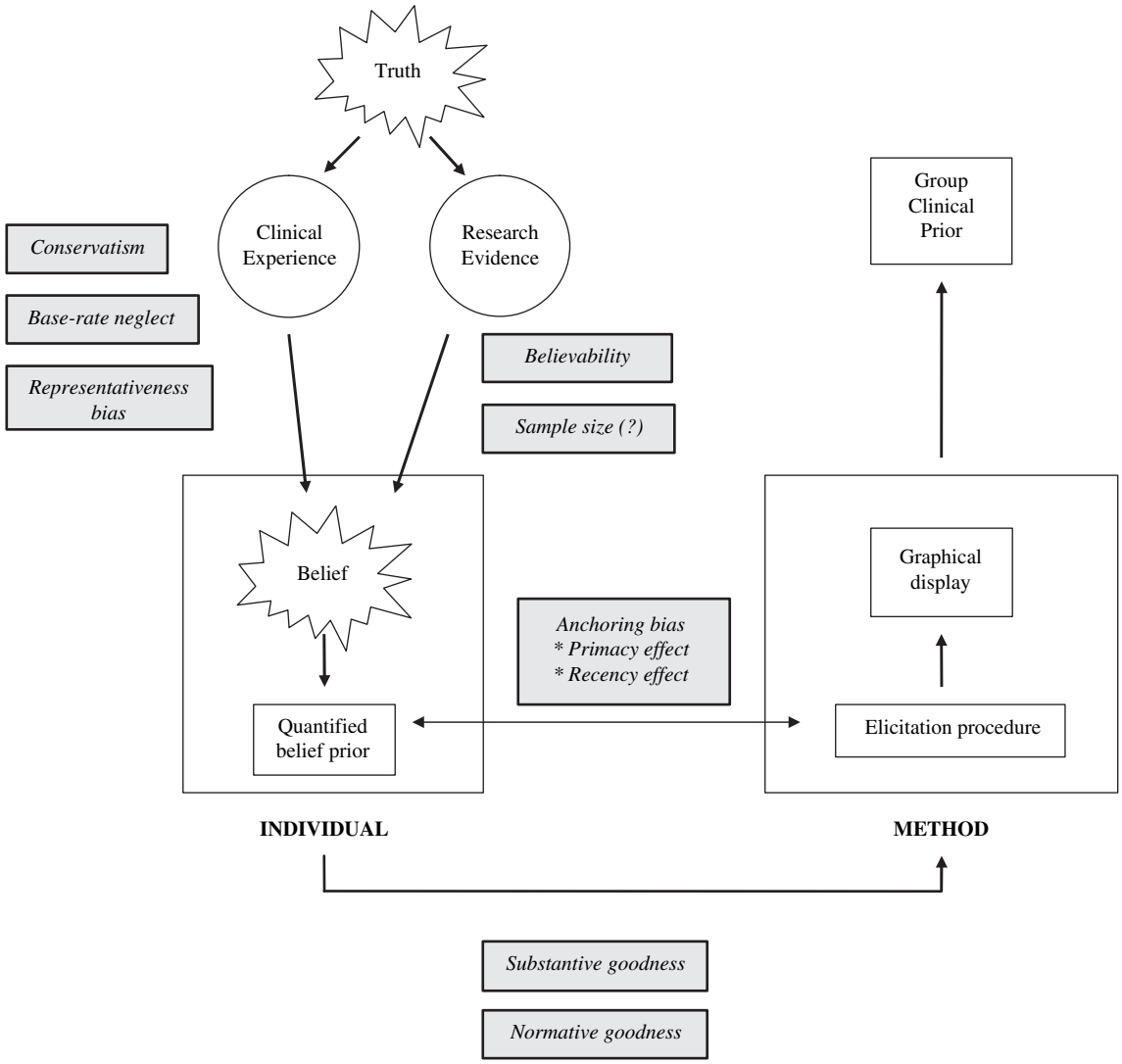

Fig. 3. Biases affecting the validity of belief elicitation.

Table 3
Summary of studies which considered validity, reliability, responsiveness, and feasibility

Authors	Validity	Reliability	Responsiveness	Feasibility
Errington et al., 1991 [13], Abrams et al., 1994 [67]	NS	NS	NS	NS
Bergus et al., 1995 [14]	Commented	NS	Commented	NS
Chaloner et al., 1993 [15], Carlin et al., 1993 [4]	Commented	NS	NS	Commented
Chaloner 1996 [17]	Commented	Commented	NS	Commented
Chaloner and Rhame 2001 [3]	Commented	NS	NS	Commented
de Vet et al., 1993 [18]	Commented	NS	Commented	NS
Dumouchel. 1988 [58]	Commented	NS	NS	Commented
Evans et al., 2002 [19]	NS	NS	NS	NS
Freedman and Spiegelhalter, 1983 [16], 1986 [26]; Spiegelhalter et al., 1993 [20]	Commented	NS	NS	Commented
Flournoy, 1994 [21]	NS	NS	NS	Commented
Garthwaite and Dickey, 1991 [22]	Commented	Commented	NS	Commented
Garthwaite and Dickey, 1992 [23]	NS	NS	NS	Commented
Gustafson et al., 2003 [24]	Literature review to ensure content validity. Concurrent validity: correlation coefficient = 0.77	Commented	NS	Commented
Hughes, 1991 [25]	NS	NS	NS	NS
Hutton and Owens, 1993 [27]	NS	NS	NS	NS
Johnson et al., 2006 [28]	Commented	NS	NS	Evaluated
Jones et al., 1998 [29]	NS	NS	Commented	Commented
Kadane et al., 1980 [68]	NS	NS	NS	NS
Kadane, 1986, 1994 [30,31]	NS	Commented	NS	Commented
Kadane, 1992 [69]	NS	NS	NS	NS
Kadane and Wolfson, 1998 [32]	Commented	Commented	NS	NS
Lehmann and Goodman, 2000 [33]	Commented	NS	NS	Commented
Li and Krantz, 2005 [34]	Poor accuracy, calibration $<30\%$ for 80% confidence	Intrarater reliability: correlation coefficient = 0.63	NS	NS
Lilford, 1994 [35]	NS	NS	NS	NS
Lilford and Braunholtz, 1996 [36]	NS	NS	NS	NS
O'Hagan, 1998 [37]	Commented	NS	NS	Comment
Parmar et al., 1994, 2001 [38,39]	Commented	NS	NS	Commented
Ramachandran, 2001 [40]	Criterion validity $R^2 = 0.5–0.6$	Interrater reliability: $R^2 = 0.9$	NS	NS
Tan et al., 2003 [41]	NS	NS	NS	Commented
Ten Centre Study Group, 1987 [70]	NS	NS	NS	NS
Van Der Wilt et al., 2004 [42]; Rovers et al., 2005 [43]	Commented	Commented	Commented	NS
White et al., 2005 [5]	Commented	NS	NS	Commented
Winkler, 1967 [44]	Concurrent validity: 2 methods were consistent 65/75 of the time	NS	NS	Commented

Abbreviation: NS, not specified.

confident that the methods meet methodologic standards. In particular, evaluation of the validity and reliability of methods is needed. If belief elicitation is to be used in a longitudinal setting where new information is gained over time, research on the responsiveness of the methods is warranted.

Through review of the literature, we have developed a conceptual framework outlining the process by which beliefs about treatment effects are formulated by experts and the process by which investigators may elicit beliefs. We have also identified potential biases which may threaten the validity, reliability, and responsiveness of the elicited belief, and incorporated these findings into the conceptual framework. Conceptual frameworks are increasingly being used to guide our thinking [60]. This framework is meant to lay down a foundation on which we synthesize the existing knowledge about the belief-elicitation process. It is not meant to be static, but rather meant to be modified as additional insights are gained. We summarize pragmatic methodologic strategies to reduce the effect of potential biases until comparative validity, reliability, and responsiveness studies are conducted. Strategies to minimize bias can be implemented at each stage of the elicitation procedure.

In an attempt to be comprehensive, we included all studies that elicited belief in a ''Bayesian context.'' Although some studies elicited prior beliefs and then incorporated it with new data in a fully Bayesian analysis, other studies did not. For example, Bergus et al. evaluated diagnostic clinical reasoning of family physicians by comparing their

Table 4
Biases in belief elicitation and methodologic strategies to their effect

Potential biases	Methodologic strategy
Identification of the sample	
Substantive goodness: knowledge of the clinical context [49]. Participants with more contextual experience provide more valid and reliable quantitative descriptions of their belief [49–51]	Include experts
Overconfidence may bias the validity of the elicited belief where some clinicians provide very little uncertainty around their estimate, corresponding to strong beliefs [2] and do not reflect realistic doubt [52]	Include experts, sample size greater than 1
Representativeness bias may occur when clinicians give more credence to study findings that conform to what they believe the results should look like [21]	Include representation of the spectrum of belief
Conservatism may occur when clinicians' beliefs confer less certainty to their belief than is justified by the data [43]	Include representation of the spectrum of belief
Believability: clinicians are more likely to be influenced by study findings that are concordant with their preconceived beliefs about the disease process or treatment effect [21]	Include representation of the spectrum of belief
Framing the question stem	
Normative goodness: knowledge of probability and statistics [49]. Participants with more mathematical experience provide more valid and reliable quantitative descriptions of their belief [49–51]	Provide an example [4,24,39] or training exercise [27,43]
Ease of use, clarity	Use clear instructions [31] and/or standardized script [44]
Anchoring bias: the reported belief is influenced by presentation of data/ scenario [37]	Avoid scenarios or summary of data
Ordering: participants' probability estimates are influenced by data presented at the beginning of the question stem (*primacy effect*) while others are influenced by data presented at the end of the question stem (*recency effect*) [37]	Avoid scenarios or summary of data or scramble the sequence of data presentation between participants [20]
Choice of response option	
Normative goodness	Provision of feedback, verification, opportunity for revision [17,20,53]
Base-rate neglect: occurs when participants fail to take account of the prevalence of the outcome among untreated patients [15,21]	State baseline rate or outcome in untreated patients [15]
Summarizing the data	
Overoptimism, overconfidence	Use averaging methods for the group clinical prior [52]
Normative goodness	Use simple figures

elicited probabilities of different diagnoses with Bayesian-derived probabilities [14]. This study was conducted in a Bayesian context, but did not use the elicited beliefs in a Bayesian analysis.

Future investigators are reminded that the term ''probability elicitation'' has been used in the literature with two different meanings [2,61]. Using Bayesian inference, subjective probabilities are not uncertain and are not estimated. A probability is stated and used to describe one's uncertainty. However, probability elicitation is also used to estimate proportions or frequencies [61]. For example, investigators may ask participants to estimate their probability of being struck by lightening, when investigators are actually asking for an estimate of the proportion of individuals who are struck by lightening. Estimating the probability of the event does not allow one to consider uncertainty. Using a Bayesian paradigm, investigators could elicit both an estimate of this proportion and the individual's uncertainty about this proportion.

One area of uncertainty is the number of participants required for a belief-elicitation study [3,4]. We found the median sample size of participants in belief-elicitation studies to be 11. Some investigators have advocated for the inclusion of more than one expert [3,4], as groups of experts are thought to perform better than the average solitary expert [37,50]. A group of participants is less likely to be dominated by a radical opinion [4]. The number of experts to include in a study is also constrained by the cost of information (time [26,37], administration [40], personnel [26]). Indeed, the addition of an expert with beliefs identical to one already elicited does not add to the range of beliefs collected in the study [30].

The correct method of sampling experts is also uncertain. The selection of a group of experts to participate in a belief-elicitation study is intended to yield some knowledge about the population of experts. It may not be possible to study the whole population. One option is simple random sampling. However, experts are not likely to be statistically independent. It may be preferable to include experts chosen nonrandomly (e.g., purposive expert sampling) and capture a range of opinions of the target population [62].

Software for belief elicitation has been developed [28,57], and some studies have been computer assisted [22,33,35,37]. This has the advantage of instant graphical presentation of the elicited belief. However, these technologies have been criticized for their lack of usability and intuitiveness [58]. This is

likely to be related to the software in question. Computer-assisted elicitation studies have been performed one-on-one. Internet-based, computer-assisted belief-elicitation surveys may be an option for future studies.

Evaluation of the validity of a belief-elicitation method for Bayesian priors is challenged by the lack of a "true objective" probability that represents subjective uncertainty about a fixed, unknown quantity. In the psychology literature, there have been studies that measure the calibration of elicited distributions compared with the true value that has been verified by the investigator (e.g., population of a country, dates of historical events, meaning of words) [63]. The participants in these studies are usually nonexperts (e.g., university students, League of Women Voters) [63]. The use of these calibration methods in studies evaluating the probability of an intervention's treatment effect is limited as the "true" treatment effect is not known. Preexisting clinical trials or observational studies may provide estimates of the treatment effect but the "truth" remains unknown. In the setting where the gold standard is not known, an alternative option would include the evaluation of construct validity. For example, one study examined intensive care unit physicians' judgments for the probability of survival for patients compared with probabilities generated by a logistic model derived from the Acute Physiology And Chronic Health Evaluation (APACHE) II illness severity index [64]. The physicians had greater discrimination than the model and identified those who were likely to die [64,65]. Whether it is better to include experts or nonexperts remains a subject of controversy. The results of this review suggest that the inclusion of clinical experts rather than generalists in an elicitation procedure improves the validity and reliability of the elicited beliefs.

Whether prior beliefs should be included in a Bayesian analysis is also controversial. Proponents of the empirical Bayesian approach do not use information external to the data at hand. We argue that the fully Bayesian approach, whether priors are informative or vague, more closely approximates true medical practice. Often, there is no published evidence available to guide physicians' ability to make a diagnosis, prognosis, or decision to institute a therapy. In these settings, clinicians will use other sources of knowledge (education, experience, expert opinion) to guide their beliefs. The fully Bayesian approach allows quantification and incorporation of these beliefs into statistical models. The onus remains on clinical investigators to use belief-elicitation methods that have demonstrable methodologic rigor. In addition, Hiance et al. have demonstrated that elicitation of prior beliefs is not only feasible, but allows for insights to be gained into the variability of experts' beliefs [66]. Consideration of a variety of prior distributions allows for the approximation of the posterior distributions held by all types of readers [66]. They suggest that elicitation from a set of experts should be considered as part of the design of future trials [66].

By summarizing methods that have been applied for belief elicitation, reviewing whatever is known about the measurement properties of each method, developing a conceptual framework for the belief-elicitation process, and identifying pragmatic methodologic strategies to reduce the effect of bias, we have synthesized the current state of knowledge for clinical researchers. This study lays the necessary groundwork for future research by highlighting areas requiring investigation. Through the use of measurement properties as criteria to assess the utility of belief-elicitation methods, we are rising to the challenge of using disciplined research methodology [6] when applying the Bayesian paradigm to clinical trials.

Our ability to comparatively evaluate the identified elicitation methods is limited by the paucity of data evaluating their measurement properties. It should be noted that for most of the studies, evaluation of the methodologic properties of the elicitation method was not the intent of the investigators. Furthermore, evaluation of the measurement properties of the methods may not have been considered necessary. In an era of evolving and more rigorous methodologic standards [8], evaluation of the measurement properties of the methods is needed, and will provide objective criteria based on which the comparative utility of the various methods could be decided.

5. Conclusion

This systematic review of the literature summarizes methods of belief elicitation for a Bayesian analysis. The measurement properties of the methods have not been adequately evaluated. Further evaluation of the validity, reliability, and responsiveness of elicitation methods is needed. Until comparative studies are performed, methodologic strategies to reduce the effect of bias on the validity and reliability of the elicited belief should be used. Based on the results of this systematic review, we recommend the following strategies: include sampling from groups of experts, use clear instructions and a standardized script, provide examples and/or training exercises, avoid use of scenarios or anchoring data, ask participants to state the baseline rate in untreated patients, provide feedback and opportunity for revision of the response, and use simple graphical methods.

Acknowledgments

Dr. Sindhu Johnson has been awarded a Canadian Institutes of Health Research Phase 1 Clinician Scientist Award. Dr. Brian Feldman is supported by a Canada Research Chair in Childhood Arthritis.

References

[1] Berry DA. Bayesian clinical trials. Nat Rev Drug Discov 2006;5: 27–36.

[2] Spiegelhalter DJ, Abrams KR, Myles JP. An overview of the Bayesian approach. Chichester: John Wiley & Sons Ltd; 2004. Bayesian approaches to clinical trials and health-care evaluation 49–120.

[3] Chaloner K, Rhame FS. Quantifying and documenting prior beliefs in clinical trials. Stat Med 2001;4:581−600.
[4] Carlin BP, Chaloner K, Church T, Louis TA, Matts JP. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. Statistician 1993;42:355−67.
[5] White IR, Pocock SJ, Wang D. Eliciting and using expert opinions about influence of patient characteristics on treatment effects: a Bayesian analysis of the CHARM trials. Stat Med 2005;24:3805−21.
[6] Moye LA. Bayesians in clinical trials: asleep at the switch. Stat Med 2008;27:469−82.
[7] Spiegelhalter DJ. Incorporating Bayesian ideas into health-care evaluation. Stat Sci 2004;19:156−74.
[8] Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, et al. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55:348−52.
[9] Johnson SR, Hawker GA, Davis AM. The health assessment questionnaire disability index and scleroderma health assessment questionnaire in scleroderma trials: an evaluation of their measurement properties. Arthritis Rheum 2005;53:256−62.
[10] Streiner DL, Norman GR. Health measurement scales. 3rd edition. NewYork: Oxford University Press; 2003. A practical guide to their development and use.
[11] Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care 2000;38(9 Suppl):II84−90.
[12] Feinstein AR. The theory and evaluation of sensibility. In: Feinstein AR, editor. Clinimetrics. New Haven: Yale University Press; 1987. p. 141−65.
[13] Errington RD, Ashby D, Gore SM, Abrams KR, Myint S, Bonnett DE, et al. High energy neutron treatment for pelvic cancers: study stopped because of increased mortality. BMJ 1991;302:1045−51.
[14] Bergus GR, Chapman GB, Gjerde C, Elstein AS. Clinical reasoning about new symptoms despite preexisting disease: sources of error and order effects. Fam Med 1995;27:314−20.
[15] Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. Statistician 1993;42:341−53.
[16] Freedman LS, Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. Statistician 1983;32:153−60.
[17] Chaloner K. Elicitation of prior distributions. In: Berry DA, Stangl DK, editors. Bayesian biostatistics. New York: Marcel Dekker Inc; 1996. p. 141−56.
[18] de Vet HC, Kessels AG, Leffers P, Knipschild PG. A randomized trial about the perceived informativeness of new empirical evidence. Does beta-carotene prevent (cervical) cancer? J Clin Epidemiol 1993;46: 509−17.
[19] Evans JS, Handley SJ, Over DE, Perham N. Background beliefs in Bayesian inference. Mem Cogn 2002;2:179−90.
[20] Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. Stat Med 1993;12:1501−11.
[21] Flournoy N. A clinical experiment in bone marrow transplantation: estimating a percentage point of a quantal response curve. In: Gatsonis C, Hodges JS, Kass RE, Singpurwalla ND, editors. Lecture notes in statistics. New York: Sringer-Verlag; 1994. p. 324−35.
[22] Garthwaite PH, Dickey JM. An elicitation method for multiple linear regression models. J Behav Decis Making 1991;4:17−31.
[23] Garthwaite PH, Dickey JM. Elicitation of prior distributions for variable selection problems in regression. Ann Stat 1992;20:1697−719.
[24] Gustafson DH, Sainfort F, Eichler M, Adams L, Bisognano M, Steudel H. Developing and testing a model to predict outcomes of organizational change. Health Serv Res 2003;38:751−76.
[25] Hughes MD. Practical reporting of Bayesian analyses of clinical trials. Drug Inf J 1991;3:381−93.
[26] Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. Stat Med 1986;5(1):1−13.
[27] Hutton JL, Owens RG. Bayesian sample size calculation and prior beliefs about child sexual abuse. Statistician 1993;42:399−404.
[28] Johnson NP, Fisher RA, Braunholtz DA, Gillett WR, Lilford RJ. Survey of Australasian clinicians' prior beliefs concerning lipiodol flushing as a treatment for infertility: a Bayesian study. Aust NZ J Obstet Gyn 2006;4:298−304.
[29] Jones P, Johanson R, Baldwin KJ, Lilford R, Jones P. Changing belief in obstetrics: impact of two multicentre randomised controlled trials. Lancet 1998;352:1988−9.
[30] Kadane JB. Progress toward a more ethical method for clinical trials. J Med Philos 1986;11:385−404.
[31] Kadane JB. An application of robust Bayesian analysis to a medical experiment. J Stat Plan Infer 1994;40:221−32.
[32] Kadane JB, Wolfson LJ. Experiences in elicitation. Statistician 1998;47:3−19.
[33] Lehmann HP, Goodman SN. Bayesian communication: a clinically significant paradigm for electronic publication. J Am Med Technol 2000;3:254−66.
[34] Li Y, Krantz DH. Experimental tests of subjective Bayesian methods. Psychol Rec 2005;55:251−77.
[35] Lilford R. Formal measurement of clinical uncertainty: prelude to a trial in perinatal medicine. The Fetal Compromise Group. BMJ 1994;308:111−2.
[36] Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996;7057:603−7.
[37] O'Hagan A. Eliciting expert beliefs in substantial practical application. Statistician 1998;47:21−35.
[38] Parmar MK, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. CHART Steering Committee. Stat Med 1994;13:1297−312.
[39] Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der SE. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. Lancet 2001;358:375−81.
[40] Ramachandran G. Retrospective exposure assessment using Bayesian methods. Ann Occup Hyg 2001;45:651−67.
[41] Tan S-B, Chung Y-F, Tai B-C, Cheung Y-B, Machin D. Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. Control Clin Trials 2003;2:110−21.
[42] Van Der Wilt GJ, Rovers M, Straatman H, Van Der BS, Van Den BP, Zielhuis G. Policy relevance of Bayesian statistics overestimated? Int J Technol Assess 2004;4:488−92.
[43] Rovers MM, Van Der Wilt GJ, Van Der BS, Straatman H, Ingels K, Zielhuis GA. Bayes' theorem: a negative example of a RCT on grommets in children with glue ear. Eur J Epidemiol 2005;1:23−8.
[44] Winkler RL. The assessment of prior distributions in Bayesian analysis. J Am Stat Assoc 1967;62:776−800.
[45] Van der Fels-Klerx IH, Goossens LH, Saatkamp HW, Horst SH. Elicitation of quantitative data from a heterogeneous expert panel: formal process and application in animal health. Risk Anal 2002;22:67−81.
[46] Carter BL, Butler CD, Rogers JC, Holloway RL. Evaluation of physician decision making with the use of prior probabilities and a decision-analysis model. Arch Fam Med 1993;2:529−34.
[47] Normand SL, Frank RG, McGuire TG. Using elicitation techniques to estimate the value of ambulatory treatments for major depression. Med Decis Making 2002;22:245−61.
[48] Parmar MKB, Ungerleider RS, Simon R. Assessing whether to perform a confirmatory randomized clinical trial. J Natl Cancer I 1996;22:1645−51.
[49] Winkler RL. Probabilistic prediction: some experimental results. J Am Stat Assoc 1971;66:675−85.
[50] Clemen RT, Wolmark N. Combining probability distributions from experts in risk analysis. Risk Anal 1999;19:187−203.
[51] Murphy AH, Winkler RL. Reliability of subjective probability forecasts of precipitation and temperature. Appl Statist 1977;26:41−7.
[52] Wallsten TS, Budescu DV. Encoding subjective probabilities: a psychological and psychometric review. Manage Sci 1983;29:151−73.
[53] Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Statist Soc A 1994;157:357−416.

[54] Hogarth RM. Cognitive processes and the assessment of subjective probability distributions. J Am Stat Assoc 1975;70:271−94.
[55] Evans JS, Brooks P, Pollard P. Prior beliefs and statistical inference. Br J Psychiatry 1985;76:469−77.
[56] Winkler RL. The quantification of judgement: some methodological suggestions. J Am Stat Assoc 1967;62:1105−20.
[57] Savage LJ. Elicitation of personal probabilities and expectations. J Am Stat Assoc 1971;66:783−801.
[58] Dumouchel W. A Bayesian model and a graphical elicitation procedure for multiple comparisons. Bayesian Stat 1988;3:127−45.
[59] Genest C, Zidek JV. Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1986;1:114−48.
[60] Hawker GA, Gignac MA. How meaningful is our evaluation of meaningful change in osteoarthritis? J Rheumatol 2006;33:639−41.
[61] O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Fundamentals of probability and judgement. Chichester: John Wiley & Sons Ltd; 2006. Uncertain judgements. Eliciting experts' probabilites 1−24.
[62] Trochim WM. The research methods knowledge base. 2nd edition. Cincinnati, Ohio: Atomic Dog Publishing; 2006.
[63] Morgan MG, Henrion M. Human judgement about and with uncertainty. Cambridge: Cambridge University Press; 1990. Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis 102−140.
[64] McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9: 125−32.
[65] O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. The psychology of judgement under uncertainty. Chichester: John Wiley & Sons Ltd; 2006. Uncertain judgements. Eliciting experts' probabilites 33−60.
[66] Hiance A, Chevret S, Levy V. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. J Clin Epidemiol 2009;62:431−7.
[67] Abrams K, Ashby D, Errington D. Simple Bayesian analysis in clinical trials: a tutorial. Control Clin Trials 1994;5:349−59.
[68] Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC. Interactive elicitation of opinion for a normal linear model. J Am Stat Assoc 1980;75:845−54.
[69] Kadane JB. Subjective Bayesian analysis for surveys with missing data. Statistician 1992;42:415−26.
[70] Ten Centre Study Group. Ten centre trial of artificial surfactant (artificial lung expanding compound) in very premature babies. Br Med J (Clin Res Ed) 1987;294:991−6.