Elicitation of a Beta Prior for Bayesian Inference in Clinical Trials

Yujun Wu[*,1], **Weichung J. Shih**[2], and **Dirk F. Moore**[2]

[1] Biostatistics and Programming, Sanofi-Aventis, BWX200-4-1642, 200 Crossing Blvd., Bridgewater, NJ 08807, USA

[2] Division of Biometrics, The Cancer Institute of New Jersey, and Department of Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey, NJ, USA

Received 17 July 2007, revised 22 August 2007, accepted 28 August 2007

Summary

When making Bayesian inferences we need to elicit an expert's opinion to set up the prior distribution. For applications in clinical trials, we study this problem with binary variables. A critical and often ignored issue in the process of eliciting priors in clinical trials is that medical investigators can seldom specify the prior quantities with precision. In this paper, we discuss several methods of eliciting beta priors from clinical information, and we use simulations to conduct sensitivity analyses of the effect of imprecise assessment of the prior information. These results provide useful guidance for choosing methods of eliciting the prior information in practice.

Key words: Bayesian inference; Prior and posterior distribution; Imprecise assessment; Beta distribution; Clinical trials.

1 Introduction

Ascertaining and setting up the prior information is an essential step for obtaining the posterior distribution in the Bayesian inference framework. While the dominant references in this area appear to be in the engineering field, Bayesian methods are also increasingly used for designing, monitoring and analyzing clinical trials. In a clinical trial application, subjective assessment of the prior density based on the clinical investigator's expert opinion has been studied by Chaloner et al. (1993), and Chaloner and Rhame (2001). The process typically starts by asking the investigator a number of questions, which are expected to provide sufficient information to construct the density function representing what he or she actually believes. Two issues to consider are (1) the quantities to ask the expert and (2) the mathematics required to derive the distribution parameters from these quantities (Garthwaite, Kadane, and O'Hagan, 2005). Not only should these quantities provide sufficient information to be converted into the distribution's parameters, but more importantly they also need to be easily understood and assessed by the investigator (Kadane and Wolfson, 1998).

For continuous variables with a normal distribution as the prior, it is relatively easy to ascertain the mean and standard deviation, as these are familiar concepts to most practitioners. For binary variables, the family of beta distributions with parameters α, and β, or equivalently, with mean $\mu = \alpha/(\alpha+\beta)$ and variance $\mu(1-\mu)/(\alpha+\beta-1)$, is a good choice for a prior as it is the conjugate of the binomial

[*] Corresponding author: e-mail: yujun.wu@sanofi-aventis.com, Phone: +1 732 235 8816, Fax: +1 732 235 8808

distribution. But the asymmetry of the beta distribution makes selection of a mean, and especially of a standard deviation, much less intuitive to our medical colleagues. To ascertain the beta prior for a binary variable, several methods have been developed based upon eliciting observable quantities such as the mode and percentiles. We will review some the previous work, which mainly have appeared in engineering applications, provide our assessment of these methods, and apply them to clinical trials. Furthermore, we consider the possible uncertainty problem in the elicitation process.

As noted by Oakley and O'Hagan (2005), a critical problem in the elicitation process is that experts can seldom specify the quantities with precision, and, psychological aspects of the investigators can adversely influence their ability to provide precise numerical quantities such as probabilities (Daneshkhah, 2004). While these researchers emphasize engineering applications, additional variability inherent in human studies of the type we consider here further exacerbates errors in specifying prior quantities. These difficulties complicate the elicitation process and impact the choice of an elicitation method, an issue that has been largely ignored in the elicitation literature. In this paper, we use simulation studies to assess the sensitivity of different methods of beta prior elicitation. We also provide guidance for choosing the elicitation methods in practice to reduce the impact of misspecification error.

The article is organized as follows. We briefly review several methods of eliciting beta priors in Section 2, and discuss choices of quantities to elicit in connection with the beta distribution properties. In Section 3 we conduct simulation experiments to study the effect of imprecise assessment on elicitation methods. Section 4 presents a pilot clinical study as an example. More comments and discussion are given in Section 5 including some methods based on subjective predictive probability.

2 Methods of Eliciting Beta Prior

The elicitation of the two parameters α and β of the beta prior distribution has received good coverage in the engineering literature, and a comprehensive reviews were provided by Hughes and Madden (2002) and Jenkinson (2005). Briefly, we can classify the elicitation methods into two categories according to the quantities to elicit from the investigator: (a) "location and interval" methods, and (b) "percentiles only" methods. We discuss these in turn.

2.1 Location and interval methods

This category of elicitation methods involves the specification of a location parameter value of the beta prior distribution and an interval with a probability that the true location value falls in the interval. The location parameter may be the mean, as in Gross (1971), or the mode, as in Fox (1966). In our experience with clinical trials, when assessing a binary variable, a medical investigator's prior estimate of the proportion, such as the response rate (p) in an efficacy study, usually represents his or her belief of the most likely value, i.e., the mode of the distribution, rather than the mean, which has no immediate intuition in the case of a skewed beta distribution. We thus recommend eliciting the mode when asking for "the response rate that is most likely to occur".

After obtaining the mode estimate, denoted by $\hat{m}$, the next question to ask is then the medical investigator's subjective probability, say μ, that the true rate lies within a given interval (r_1, r_2): "What is the chance that the response rate falls between r_1 and r_2?" Of course, it is also possible to elicit this information the other way around, by pre-specifying a probability μ and asking the investigator to assess the corresponding interval (r_1, r_2): "What is the range of values within which the response rate will have a chance of $100\,\mu\%$ to occur?"

Given the above information, the statistician is able to determine the parameters of the beta prior by solving the following two equations,

$$\frac{\alpha - 1}{\alpha + \beta - 2} = \hat{m} \qquad \text{and} \qquad \int_{r_1}^{r_2} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, r^{\alpha-1}(1-r)^{\beta-1}\,\mathrm{d}r = \mu\,. \tag{1}$$

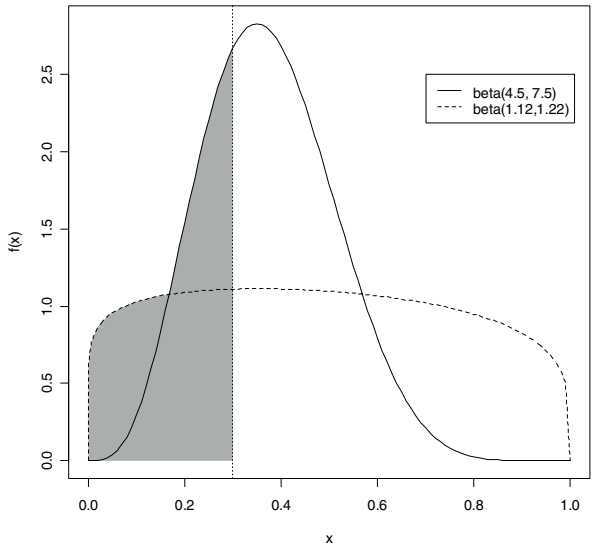

Figure 1 Two beta distributions satisfying the same conditions: mode = 0.35 and Prob ($X < 0.3$) = 0.31.

These methods are commonly used in an efficacy problem, where a unimodal (strictly between 0 and 1) beta distribution ($\alpha > 1$ and $\beta > 1$) is often assumed as a prior for the response rate (p) and medical investigators are comfortable with providing a prior estimate of the most likely rate.

We consider three types of interval (r_1, r_2). In Gross (1971) and Duran and Booker (1988), the interval is (0, $K\hat{p}$), where $0 < K < 1$ and $\hat{p}$ is the estimate of the elicited mean value. In Weiler (1965), the interval is ($2\hat{p}$, 1), requiring $\hat{p} < 0.5$. Another type is the interval with approximately the highest probability. In Fox (1996), the interval is ($\hat{m} - K\hat{m}$, $\hat{m} + K\hat{m}$), where $0 < K < 1$ is a pre-specified constant.

We caution, however, that an inappropriate interval may lead to non-unique solutions. One such a situation is an interval (0, r_2) with r_2 less than the mode $\hat{m}$, when $\hat{m} < 0.5$ (i.e., when the distribution skews to the left), as $\Pr(\text{beta}(\alpha, \beta) < r_2)$ may not be a monotone function of α with $\beta = ((1 - \hat{m})\,\alpha + 2\hat{m} - 1)/\hat{m}$. For instance, assuming that the investigator supplies a mode of 0.35, if the lower interval (0, 0.3) is chosen and a subjective coverage probability of 0.31 is provided, we will then have two solutions, beta(4.5, 7.5) and beta(1.12, 1.22), mathematically satisfying the given conditions (see Figure 1).

Therefore, when $\hat{m} < 0.5$, to ensure a unique solution, the lower interval (0, r_2) should be chosen with $r_2 > \hat{m}$, or the upper interval (r_1, 1) with $r_1 > \hat{m}$. This is because, for a left-skewed beta distribution with the mode $\hat{m}$ fixed, the density becomes more concentrated on the mode as α increases and $\beta = ((1 - \hat{m})\,\alpha + 2\hat{m} - 1)/\hat{m}$. When r_2 is greater than the mode, the probability $\Pr(\text{Beta}(\alpha, \beta) < r_2)$ is a strictly increasing function of α and we then have an unique solution of (α, β). Similarly, when $\hat{m} > 0.5$ is elicited, we should always choose $r_2 < \hat{m}$ for the lower interval (0, r_2), or choose $r_1 < \hat{m}$ for the upper interval (r_1, 1).

In addition, the length of the interval (r_1, r_2), or equivalently the coverage probability of the interval, is also an important factor to consider. As discussed by O'Hagan (1998), asking for too wide an interval, which has a high coverage probability, will yield unreliable answers that give a very poor indication of the variance. This is because, in effect, assessing a high probability interval requires the knowledge of the small tails of the distribution, and small tails are often difficult to quantify. For example, it is hard to justify why a person thinks that the probability for a small tail area is 0.01 rather than 0.02 when the exact information is lacking. Moreover, the beta distribution parameters are

very sensitive to small tail area probabilities. For instance, assume that the true distribution is Beta(4.5, 7.5), with the mode at 0.35, and that we ask for the coverage probability for the interval (0.05, 0.8). If the investigator misjudges the probability as 97% while the true coverage is 99.89%, we would obtain a distribution Beta(2.28, 3.37) from the elicited information, which has a quite different parameter values compared to the true values. Therefore, we suggest eliciting either the upper or the lower quartiles (e.g., León et al. (2003)), or a highest density interval with 50–80% coverage probability.

2.2 Percentiles only methods

Based on the beta distribution function, another class of elicitation methods is through the specification of two or more percentiles, without asking for the location information. Mathematically, two percentiles are enough for determining the two parameters of a beta distribution. The basic theme is to obtain the information of two percentiles K_1 and K_2 accompanying the corresponding probabilities μ_1 and μ_2. The beta parameters can then be obtained from the numerical solution of

$$\int_0^{K_1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, r^{\alpha-1}(1-r)^{\beta-1}\,\mathrm{d}r = \mu_1 \qquad \text{and} \qquad \int_0^{K_2} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\,\mathrm{d}r = \mu_2\,.$$

See, Duran and Booker (1988). Weiler (1965) used a special case, $\mu_2 = 1 - \mu_1$ and $K_1 < K_2$.

Questions one could ask the investigator would be, "What is the value below which the true response rate will have a chance of $100\,\mu_1\%$ to lie?" and "How about for a chance of $100\,\mu_2\%$?" A natural alternative is to ask for the reverse information, providing two percentiles K_1 and K_2 and then letting the investigators give their prior assessments of the probabilities μ_1 and μ_2, e.g., "What is the chance that the true response rate is less than K_1?" and "What is the chance that the true response rate is less than K_2?" When asking for the two percentiles, we should not choose μ_1 too close to 0, or μ_2 too close to 1. Equivalently, K_1 and K_2 should not be close to the two tail ends. This is to avoid misjudging the probabilities for small tail areas and the adversely magnified effect on the determination of the distribution parameters.

For these percentiles only methods, it is not necessary to elicit a single point estimate such as the mode. Hence, they are expected to perform better than those methods in the previous category for some pilot clinical trials where the investigator is uncertain about a single (the most likely) value. The methods are also particularly useful when a non-toxic drug or vaccine is under study, where the toxicity rate is believed to be more likely nil and the chance is smaller for a higher toxicity rate. In this case, a strictly decreasing beta distribution function ($0 < \alpha \leq 1$ and $\beta < 1$) is then a good choice for a prior for the toxicity rate, where the distribution has a mode at 0.

3 Compare Elicitation Methods Under Imprecise Prior Assessment

In studying the various methods discussed above, previous authors assume that the investigators' inputs always precisely represent their prior knowledge or beliefs. However, we know that the investigators are often unable to give their prior judgments with absolute precision in an elicitation process. How do the methods behave and compare when imprecision is taken into account? We investigate this question with simulations in term of the parameters of the prior distribution and quantities of the posterior distribution.

3.1 Sensitivity analysis 1: Assessing prior distribution parameters with imprecision

We have studied the eliciting methods based on the mode and interval (MI) and the percentiles only (PO). For the MI methods, choices of either the lower interval or the highest density interval are compared. For the PO methods and MI methods when percentiles or intervals are involved, we com-

Table 1 Eliciting methods based on the mode and interval (MI) and the percentiles only (PO).

Method	Eliciting from investigator		Given to investigator
MI_Low_a	Mode	Coverage probability of $(0, r_2)$	A lower interval $(0, r_2)$
MI_Low_b	Mode	The lower interval $(0, r_2)$	Coverage probability of a lower interval $(0, r_2)$
MI_HD_a	Mode	Coverage probability of (r_1, r_2)	A highest density interval (r_1, r_2)
MI_HD_b	Mode	The highest density interval (r_1, r_2)	Coverage probability of a highest density interval (r_1, r_2)
PO_a	$100\,\mu_1\%$	$100\,\mu_2\%$	*Percentiles* K_1 and K_2
PO_b	K_1	K_2	$100\,\mu_1\%$ and $100\,\mu_2\%$

pare the two ways of elicitation: one is to give a probability, and ask for the subjective assessment of the boundaries of the interval or the percentile; the other is to specify the two ends of the interval or the percentile, then ask for a subjective assessment of the probability. Based on the quantities being elicited, these methods are now summarized in Table 1 and described as follows.

- MI_Low: Ask for the mode m and
  a. Give a lower interval $(0, r_2)$ and ask for the coverage probability μ.
  b. Give a probability μ and ask for the lower interval $(0, r_2)$.
- MI_HD: Ask for the mode m and
  a. Give a highest density interval (r_1, r_2) and ask for the coverage probability μ.
  b. Give a probability μ and ask for the highest density interval (r_1, r_2).
- PO: Ask for the $100\,\mu_1\%$ percentile K_1 and the $100\,\mu_2\%$ percentile K_2.
  a. Give K_1 and K_2 and ask for the μ_1 and μ_2.
  b. Give μ_1 and μ_2 and ask for the K_1 and K_2.

Throughout the simulation, a beta(4.5, 7.5) is assumed to be the true prior distribution which has a mode at 0.35 and mean $4.5/12 = 0.375$. Hence, the distribution is moderately skewed to the left. We have studied other parameter values, and observed the same overall results. For the method of MI_Low, r_2 was set at 0.45, and accordingly μ was equal to 0.714. For the method of MI_HD, we chose the interval (0.2, 0.5) with the coverage probability μ equal to 0.723. For the method of PO, points 0.3 and 0.5 were chosen, corresponding to the 31% and 82% percentiles, respectively. We simulated the subjective assessments with imprecision using an additive error to the true value. Specifically, let $P^* = P + \varepsilon$, where P is the true value of the quantity to elicit and $\varepsilon \sim N(0, \tau^2)$ with the variance parameter τ^2 describing the imprecision in the assessment of P. For example, P is m or μ for the method of MI_HD_a. The P^* then falls within the interval $P \pm Z_{0.025} \cdot \tau$ with 95% confidence, i.e., the P^* has an approximate imprecision error of $Z_{0.025} \cdot \tau$, where $Z_{0.025}$ denotes the $100(1 - 0.025)$ percentile of the standard normal distribution ($= 1.96$). Considering that the imprecision should be smaller for quantities that are near 0 or 1, we set $\tau^2 = [\sigma(P(1 - P))]^2$, as suggested in Oakley and O'Hagan (2005). To investigate the effect of different imprecision magnitude, we specified $\sigma = 0.05$ and 0.1, and accordingly, the P^* deviates from the true value with a conservative error of 0.025 ($= \max\{1.96 \times 0.05 \times (P(1 - P)),\ 0 < P < 1\}$) and 0.05 ($= \max\{1.96 \times 0.1 \times (P(1 - P)),\ 0 < P < 1\}$), respectively. We simulated 1000 replications, and evaluated the averages of the elicited distribution parameters, their variances and the corresponding mean square errors (MSE).

The simulation results are summarized in Tables 2 and 3 for $\sigma = 0.05$ and 0.1, respectively. The most striking result is that all the "a" methods perform substantially better than the corresponding "b" methods, producing less biased estimates with considerably smaller variance, thus much smaller MSE. The performance of the "b" methods is more adversely affected by the increased imprecision com-

Table 2 The elicited prior distribution parameters when $\sigma = 0.05$.

Method	α (= 4.5)			β = (7.5)		
	Average	SD	MSE	Average	SD	MSE
MI_Low_a	4.63	0.92	0.86	7.68	1.37	1.90
MI_Low_b	4.80	1.44	2.12	8.00	2.39	5.96
MI_HD_a	4.52	0.28	0.08	7.55	0.32	0.11
MI_HD_b	4.56	0.48	0.24	7.62	0.82	0.69
PO_a	4.49	0.26	0.07	7.49	0.41	0.17
PO_b	4.62	0.78	0.62	7.71	1.26	1.63

Table 3 The elicited prior distribution parameters when $\sigma = 0.1$.

Method	α (= 4.5)			β = (7.5)		
	Average	SD	MSE	Average	SD	MSE
MI_Low_a	5.29	2.82	8.60	8.61	4.01	17.33
MI_Low_b	5.85	4.10	18.59	9.65	6.63	48.57
MI_HD_a	4.64	0.62	0.39	7.70	0.72	0.56
MI_HD_b	4.88	1.86	3.59	8.11	2.79	8.17
PO_a	4.52	0.52	0.28	7.53	0.84	0.70
PO_b	4.86	2.03	4.26	8.12	3.30	11.27

pared to the "a" methods. This indicates that, when eliciting a beta prior distribution in the presence of possible misjudgment, the approaches of giving the interval or the percentiles and then asking for the probability are favorable, compared to the reverse approaches of asking for the interval limits or the percentiles for a given probability.

For the two MI methods, MI_Low and MI_HD, it is also clear from the Tables 2 and 3 that MI_HD is superior to MI_Low, as reflected by the smaller bias and over 70% reduction in the variance, thus dramatically smaller MSE. The out performance of MI_HD becomes more obvious when the imprecision increases.

As far as the comparison between MI_HD_a and PO_a is concerned, a competitive result was observed in Tables 2 and 3. We continue their comparisons in term of the posterior distribution as follows.

3.2 Sensitivity analysis 2: Effect on posterior distribution

In order to further differentiate the two optimal methods from the previous section, MI_HD_a and PO_a, we now study their sensitivity to the possible error in elicitation in terms of the posterior quantities of interest. Both one-sample and two-sample problems have been studied for these sensitivity analyses, and similar results have been obtained. In the following, we present only the simulation study from the one sample problem.

We keep the same design settings as in the previous section with a beta(4.5, 7.5) as the true prior (with true mode = 0.35), except that σ is now set at 0.1, 0.2, and 0.3 to show the impact of moderate to large errors. Considering that in practice investigators often have relatively small error in judging the mode, we fix σ at 0.1 for the mode assessment. The posterior quantities to be evaluated are the posterior mode and the posterior probability $\Pr(R > r \mid n, y)$ as r varies from 0 to 1, where R denotes

Table 4 The length of the 95% percentile intervals for the posterior mode.

σ	*Y*	MI_HD_a	PO_a
0.1	10	0.024	0.009
	20	0.026	0.028
0.2	10	0.027	0.018
	20	0.048	0.056
0.3	10	0.027	0.028
	20	0.071	0.085

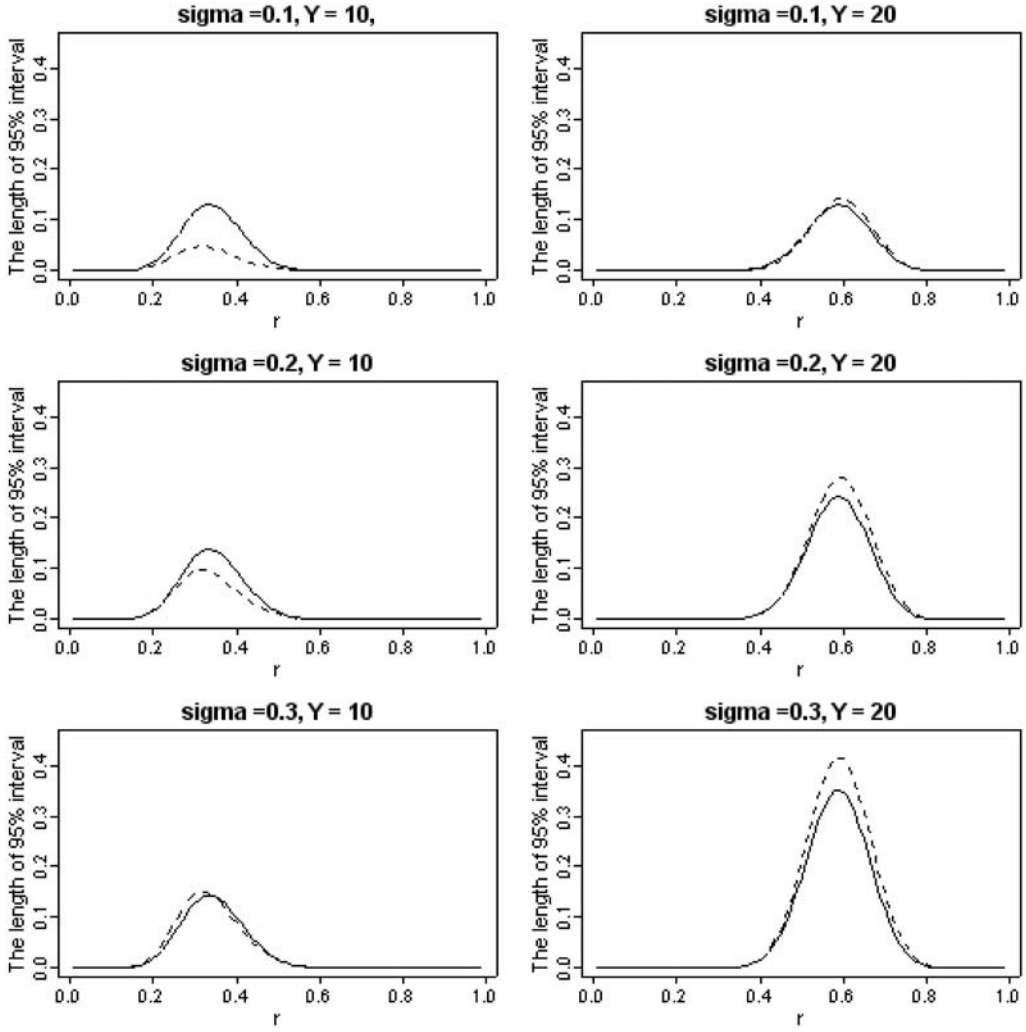

Figure 2 The length of the 95% percentile intervals for the posterior probabilities. The solid line represents the method MI_HD_a. The dashed line indicates the method PO_a.

the true response rate. We let the planned sample size $n = 30$, and assume the observed responses $y = 10$ and 20 to represent the two scenarios where, respectively, the observed response rate (= 1/3, for $y = 10$) is close to and (= 2/3, for $y = 20$) is largely deviated from the true prior beta mode. We examine how the posterior quantities of interest change when the prior information elicited with methods MI_HD_a and PO_a are contaminated with an error described by σ. Specifically, we simulate the elicitation process 5000 times, and compute their lower and upper bounds of the 95% percentile interval, i.e., the 2.5% and 97.5% percentiles.

Table 4 summarizes the length of the 95% percentile intervals of the posterior mode for different σ and *Y*. Figure 2 depicts the length of the 95% percentile intervals for the posterior probabilities $\Pr(R > r \mid n, y)$ as r varies from 0 to 1. From both the table and the figure, we see that, when the observed rate is close to the prior beta mode (i.e., when $Y = 10$), PO_a is less sensitive to the error than MI_HD_a is for $\sigma = 0.1$ and 0.2, but its sensitivity increases, as measured by the length of the interval as σ becomes large, and catches up that of MI_HD_a when $\sigma = 0.3$, while MI_HD_a's sensitivity remains relatively stable. When the observed rate deviates from the prior beta mode (i.e., when $Y = 20$), both methods have increased sensitivity in calculating the posterior quantities compared to the case where $Y = 10$. They also become more sensitive as σ increases. In this case, PO_a is under performed compared to MI_HD_a. In addition, it is revealed in the figure that, for both methods, the posterior probabilities $\Pr(R > r \mid n, y)$ are more variable for r closer to the observed response rate (y/n), and the less so for r different from the observed rate.

4 An Example to Illustrate Application in Monitoring Clinical Trials

We give an example to illustrate the use of the elicitation methods in monitoring clinical trials with a pilot study of green tea for patients with indolent chronic lymphocytic leukemia (CLL) or low grade non-Hodgkin's lymphoma (NHL) who are not receiving cytotoxic therapy (Strair et al., 2005). The purpose of this trial is to study if green tea extract has any toxicity or disease impact in patients with CLL/NHL. Toxicity and disease impact are the primary endpoints. The study will accrue up to 24 patients, who will be monitored every 4 months while taking green tea extract. Each clinical evaluation will be scored as clinically improved, stable disease, or progressive disease. The study will be discontinued for toxicity or futility if we are at least 90% certain that more than 30% of patients will progress within 1 year, or at least 80% certain that that less than 15% of patients will have a clinically improved response.

For such a pilot study with a small sample size, a Bayesian approach is particularly useful for formulating the stopping rule based on the posterior probability. Specifically, let q denote the probability of having progressive disease within 1 year and p denote the probability of patients having a clinically improved response. The stopping rule applies if either the posterior probability of $\{q > 30\%\}$ greater than 90% or the posterior probability of $\{p < 15\%\}$ greater than 80% given the collected data. We assume the beta distribution as a prior for q and p.

The MI_HD_a method was used to elicit the prior distribution parameters for one-year progressive disease rate q. We first asked the investigator: "What do you suppose is the most likely one-year progressive disease rate?" After having heard the answer of 15% (i.e., the assumed mode), we then asked: "What do you suppose is the chance that the progressive disease rate falls between 5% and 30%?" The answer was 70%. From this information, a prior distribution for q was then determined as beta(2.15, 7.53). We listed in Table 5 the values of $P(q > 0.3 \mid n, y)$ for $n = 1, \ldots, 24$ patients and $y = 0, 1, \ldots, n$ events, with the shaded areas indicating the situations where the posterior probability of $\{q > 30\%\}$ is greater than 90%. If any of the cases of (n, y) in the shaded areas occurs, the study will be stopped. As for the clinically improved response rate p, the investigator was not able to provide a single prior estimate of the rate (the most likely) due to lack of information about the effect of green tea extract. We thus used the PO_a method to do the elicitation. The question asked the investigator was "What are the chances that response rate is less than 15%? And less than 50%?" The answers were 30% and 80%, respectively. Based upon this information, a beta(1.06, 2.43) was elicited

Table 5 Posterior probability that the progressive disease rate at 1 year is greater than 30% given (n, y).

(n, y)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0.19	0.45																							
2	0.15	0.37	0.63																						
3	0.11	0.31	0.55	0.77																					
4	0.09	0.25	0.48	0.71	0.87																				
5	0.06	0.20	0.41	0.64	0.82	0.93																			
6	0.05	0.16	0.35	0.57	0.77	0.90	0.96																		
7	0.04	0.13	0.29	0.50	0.71	0.86	0.94	0.98																	
8	0.03	0.10	0.24	0.44	0.65	0.81	0.92	0.97	0.99																
9	0.02	0.08	0.20	0.38	0.58	0.76	0.89	0.95	0.98	1.00															
10	0.01	0.06	0.16	0.33	0.52	0.71	0.85	0.93	0.98	0.99	1.00														
11	0.01	0.05	0.13	0.28	0.46	0.65	0.81	0.91	0.96	0.99	1.00	1.00													
12	0.01	0.04	0.11	0.23	0.41	0.60	0.76	0.88	0.95	0.98	0.99	1.00	1.00												
13	0.01	0.03	0.08	0.19	0.36	0.54	0.71	0.84	0.93	0.97	0.99	1.00	1.00	1.00											
14	0.00	0.02	0.07	0.16	0.31	0.48	0.66	0.80	0.90	0.96	0.98	0.99	1.00	1.00	1.00										
15	0.00	0.02	0.05	0.13	0.26	0.43	0.61	0.76	0.87	0.94	0.98	0.99	1.00	1.00	1.00	1.00									
16	0.00	0.01	0.04	0.11	0.22	0.38	0.55	0.71	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00								
17	0.00	0.01	0.03	0.09	0.19	0.33	0.50	0.67	0.80	0.90	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00							
18	0.00	0.01	0.03	0.07	0.16	0.29	0.45	0.62	0.76	0.87	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00						
19	0.00	0.01	0.02	0.06	0.13	0.25	0.40	0.57	0.72	0.84	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00					
20	0.00	0.00	0.02	0.05	0.11	0.22	0.36	0.52	0.67	0.80	0.89	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
21	0.00	0.00	0.01	0.04	0.09	0.18	0.32	0.47	0.63	0.76	0.86	0.93	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00			
22	0.00	0.00	0.01	0.03	0.08	0.16	0.28	0.42	0.58	0.72	0.83	0.91	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
23	0.00	0.00	0.01	0.02	0.06	0.13	0.24	0.38	0.53	0.68	0.80	0.89	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
24	0.00	0.00	0.01	0.02	0.05	0.11	0.21	0.34	0.49	0.64	0.76	0.86	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6 Posterior probability that the clinically improved response rate is less than 15% given (n, y).

(n, y)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0.40	0.07																							
2	0.49	0.12	0.02																						
3	0.56	0.18	0.03	0.00																					
4	0.62	0.24	0.05	0.01	0.00																				
5	0.68	0.29	0.08	0.01	0.00	0.00																			
6	0.73	0.35	0.11	0.02	0.00	0.00	0.00																		
7	0.77	0.41	0.15	0.04	0.01	0.00	0.00	0.00																	
8	0.80	0.46	0.19	0.05	0.01	0.00	0.00	0.00	0.00																
9	0.83	0.51	0.23	0.07	0.02	0.00	0.00	0.00	0.00	0.00															
10	0.85	0.56	0.27	0.10	0.03	0.01	0.00	0.00	0.00	0.00	0.00														
11	0.88	0.60	0.31	0.12	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00													
12	0.89	0.64	0.36	0.15	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00												
13	0.91	0.68	0.40	0.18	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00											
14	0.92	0.72	0.44	0.22	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00										
15	0.93	0.75	0.48	0.25	0.10	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00									
16	0.94	0.78	0.52	0.28	0.12	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
17	0.95	0.80	0.56	0.32	0.15	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00							
18	0.96	0.82	0.60	0.36	0.17	0.07	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00						
19	0.97	0.84	0.63	0.39	0.20	0.09	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					
20	0.97	0.86	0.66	0.43	0.23	0.10	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
21	0.97	0.88	0.69	0.46	0.26	0.12	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
22	0.98	0.89	0.72	0.50	0.29	0.14	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
23	0.98	0.91	0.75	0.53	0.32	0.17	0.07	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
24	0.98	0.92	0.77	0.56	0.35	0.19	0.09	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

as a prior distribution for p. Table 6 was then constructed to show the values of $P(p < 0.15 \mid n, y)$ and to determine the cases of $(n,\ y)$ for which the stopping rule applies. It should be noticed that Table 5 and 6 are only used at the first time when we check the data. For the following k-th monitoring, we should use the posterior beta distribution at the $(k-1)$-th monitoring as the new prior distribution, and reconstruct the tables.

5 Discussion

We have discussed several methods of eliciting beta priors for binary variables for Bayesian inference with special attention to clinical trial applications. Previously, the methods appear mostly in the engineering literature. We have used sensitivity analyses to compare them and suggested two optimal methods, MI_HD_a and PO_a. When the misspecification error is mild, MI_HD_a is favored over PO_a. The reverse is true when the error is severe. We illustrated these two methods using an example of monitoring a pilot cancer study.

In an attempt to reduce the effect of misspecification error, we may go beyond asking for only the minimum information required to derive the prior density. For example, we may elicit multiple intervals for their coverage probabilities for the MI_HD_a method and more than two percentiles for PO_a method. However, in doing so, we may find the investigator's inputs inconsistent such that no beta distribution matches all the interval probabilities or percentiles. To consolidate the inconsistent inputs, we can determine the beta parameters by minimizing the sum of squared distances between the elicited values (such as the mode and interval probabilities for MI_HD_a or percentile probabilities for PO_a) and the corresponding values implied by the beta distribution, as suggested by O'Hagan (1998). Specifically, for MI_HD_a, the beta parameters are

$$\arg\min_{(\alpha,\beta)}\left\{\left(\frac{\alpha-1}{\alpha+\beta-2}-\hat{m}\right)^2+\sum_{i=1}^{t}\left(\int_{r_{i,1}}^{r_{i,2}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,r^{\alpha-1}(1-r)^{\beta-1}\,\mathrm{d}r-\mu_i\right)^2\right\},$$

where μ_i, $i = 1, \ldots, t$ (> 2), is the elicited coverage probability associated with the i-th interval $(r_{i,1}, r_{i,2})$; for PO_a, they are

$$\arg\min_{(\alpha,\beta)}\left\{\sum_{i=1}^{t}\left(\int_{0}^{K_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,r^{\alpha-1}(1-r)^{\beta-1}\,\mathrm{d}r-\mu_i\right)^2\right\},$$

where μ_i, $\mu = 1, \ldots, t$ (> 2), is the elicited probability for the ith percentile K_i. We also performed simulation studies with $t = 4$ for both methods. Our results showed that, in the presence of misspecification error, eliciting with an increased number of intervals (i.e., (0.15, 0.25), (0.25, 0.35),(0.35, 0.45), and (0.45, 0.55)) did not improve the MI_HD_a method with respect to the mean squared error; however, when eliciting with more percentiles (i.e., 0.15, 0.3, 0.45, and 0.6), the PO_a method slightly reduced the mean squared error. (Details are available from the authors upon request.) Another approach to handling multiple beta prior estimates from inconsistent inputs would be to use a mixture for the prior of the elicited beta distributions. Further study of this approach will be required.

Notice that our approach has focused on eliciting what Tan et al. (2002) referred to as a "clinical" prior. That is, the prior is based on the medical expert's "best" subjective opinion for the particular clinical trial under study. Other priors suggested by Tan et al. are a "reference" (or non-informative) prior, a "skeptical" prior, and an "enthusiastic" prior. The latter two are not intended to be unbiased opinions. The approach with different priors aims to form an envelope information to see if the resulting Bayesian inference is robust to the different type of priors.

Besides the elicitation of a beta prior directly, Chaloner and Duncan (1983) suggested eliciting a predictive mode and successive numbers of responders in relation to this predictive mode in a hypothetical study with a given number of patients. The beta prior parameters are then solved from

equations involving these predictive quantities. Gavasakar (1988) extended Chaloner and Duncan's elicitation method to several hypothetical studies with different number of patients, serving as a measure of sensitivity. These methods are much more sophisticated and their gain over the simpler methods discussed in this paper has yet to be investigated for clinical trials.

Acknowledgements The first author was a faculty member of the University of Medicine and Dentistry of New Jersey when the paper was prepared. The work was supported by grant number NCI CA-72720-10 for all three authors.

References

Chaloner, K., Church, T., Louis, T. A., and Matts, J. P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* **42**, 341–353.

Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a beta prior distribution: PM elicitation. *The Statistician* **32**, 174–180.

Chaloner, K. M. and Rhame, F. S. (2001). Quantifying and documenting prior beliefs in clinical trials. *Statistics in Medicine* **20**, 581–600.

Daneshkhah, A. (2004). Psychological Aspects Influencing Elicitation of Subjective Probability. BEEP's report, University of Sheffield.

Duran, B. S. and Booker, J. M. (1988). A Bayes sensitivity analysis when using beta distribution as a prior. *IEEE Transactions on Reliability* **37**, 239–247.

Fox, B. L. (1996). A Bayesian approach to reliability assessment. *Memorandum RM-5084-NASA*, The Rand Coporation, Santa Monica, CA, 23 pp.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**, 680–700.

Gavasakar, U. (1988). A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Management Science* **34**, 784–790.

Gross, A. J. (1971). The application of exponential smoothing to reliability assessment. *Technometrics* **13**, 877–883.

Hughes, G. and Madden, L. V. (2002). Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence. *Crop Protection* **21**, 203–215.

Jenkinson, D. (2005). The Elicitation of Probabilities-A Review of the Statistical Literature. BEEP's report, University of Sheffield.

Kadane, J. B. and Wolfson, L. J. (1998). Experience in Elicitation. *The Statistician* **47**, 3–19.

León, C. L., Vázquez-Polo, F. J., and González, R. L. (2003). Elicitation of Expert Opinion in Benefit Transfer of Environmental Goods. *Environmental and Resource Economics* **26**, 199–210.

Oakley, J. and O'Hagan, A. (2005). Uncertainty in prior elicitations: a non-parametric approach. *Research Report No. 521/02*, Department of Probability and Statistics, University of Sheffield.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician* **47**, 21–35.

Strair, R., Rubin, A., Bertino, J., Schaar, D., Gharibo, M., Goodin, S., Krimmel, T., Grospe, S., Dudeck, L., Mills, L., Martin, J., and Wu, Y. (2005). Green Tea for Patients with Indolent non-Hodgkin's Lymphoma or Chronic Lymphocytic Leukemia not receiving Cytotoxic Therapy. *Clinical Trial Protocol*, The Cancer Institute of New Jersey.

Tan, S.-B., Machin, D., Tai, B.-C., Foo, K.-F., and Tan, E.-H. (2002). A Bayesian re-assessment of two Phase II trials of gemcitabine in metastatic nasopharyngeal cancer. *British Journal of Cancer* **86**, 843–850.

Weiler, H. (1965). The use of incomplete beta functions for prior distributions in binomial sampling. *Technometrics* **7**, 335–347.

