4

Elicitation of Prior Distributions

Kathryn Chaloner University of Minnesota, St. Paul, Minnesota

ABSTRACT

Research on methods for helping experts to specify subjective prior distributions is briefly reviewed and discussed. Specific methods for elicitation for clinical trials are also reviewed. Some suggestions are made and an example is given.

Keywords: probability assessment, subjective probability distribution

1 INTRODUCTION

There is a large volume of psychological literature on how people make judgments about uncertainty. This review is not comprehensive and only a few key references are given. Some general suggestions are provided in Section 5. These suggestions are based on my personal observational experiences working with physicians and researchers and are not based on scientific experiments or psychological theories. I would be delighted if better recommendations were developed. An example is given in the Appendix of eliciting a distribution for a human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) clinical trial in progress.

2 OVERVIEW OF ELICITATION METHODS

Some of the initial work on elicitation can be found in Winkler (1967a,b, 1971) and Savage (1971). Hogarth (1975) gives an overview of early work incorporating much of the psychological and behavioral science literature into the discussion. Hogarth (1987) provides a more recent perspective, and Appendix B of his book is a useful tutorial on assessing probabilities. Other recent references on psychological and behavioral aspects of subjective probability assessment are found in Kahneman *et al.* (1982); this is a collection of papers by numerous authors about how people process uncertainty. Another important collection of papers is Kyberg and Smokler (1980). Wallsten and Budescu (1983) review the psychological aspects of subjective probability assessment and argue that measures of reliability and validity, as defined in measurement theory, should be applied to subjective probability assessments. They divide the literature into two parts: (1) studies of subjective probability assessment from nonexperts who have no expertise in either probability or the subject matter and (2) studies from experts, either subject matter experts or experts in probability or decision theory. They report that there is a lack of experiments investigating reliability and validity of experts. Von Winterfeldt and Edwards (1986, Chapter 4) also give a general discussion of elicitation and some of the basic issues.

Some probabilities are easier to assess than others. O'Hagan (1994, p. 107) gives the example that it is easy to assess the probability of a coin landing heads when tossed, but to assess the probability of 4 heads in 10 independent tosses is much harder. Ravinder *et al.* (1988) describe a technique they call *decomposition*. Rather than elicit a probability of an event A directly, they give circumstances in which it may be advantageous to elicit a series of marginal and conditional probabilities, $P(\mathrm{B})$ and $P(\mathrm{A} \mid \mathrm{B}_i)$ where the events B_i, $i = 1, \ldots, n$, form a partition of the sample space and

$$P(\mathrm{A}) = \sum P(\mathrm{A} \mid \mathrm{B}_i)P(\mathrm{B}_i)$$

Lindley (1985, pp. 39–41) calls this *extending the conversation* and also suggests its usefulness in elicitation.

Despite the large volume of psychological literature on probability assessment, few of the ideas, theories, and empirical results have been applied to develop operational methodology for eliciting prior distributions for specific statistical models and problems. The linear regression problem has received some attention: Kadane *et al.* (1980) suggested and implemented a method of elicitation based on specifying predictive distributions. They restricted beliefs to lie in the normal-gamma conjugate family and elicited quantiles of the predictive distributions at several values of the explanatory variables. They argue that predictive distributions on potentially observable quantities are easier to think about than distributions on unobservable parameters. Their method asks more than the minimum number of questions required and so any inconsistencies must be reconciled. An example of this method is given in Chapter 5, by Kadane and Wolfson. Garthwaite and Dickey extended the work of Kadane *et al.*; see Garthwaite and Dickey (1985, 1988, 1991) and Garthwaite (1992) for some successful elicitation methods for regression models. See also Dickey *et al.*, (1986). Other, related methods can be found in Laskey and Black (1989) and Black and Laskey (1989) for analysis of variance models, in Chapter 5, by Kadane and Wolfson, for exponential lifetime models, and in Chaloner and Duncan (1983, 1987) for binomial and multinomial problems. A novel way of evaluating elicitation procedures by adding random error to the values specified is given in Gavasakar (1988).

2.1 Problems

People are not typically very good probability assessors. They make mistakes and are inconsistent.

In elicitation methods such as that of Kadane *et al.* (1980) an assumption is made that the expert's beliefs follow a particular parametric family; the family is chosen for its convenience or conjugateness. Inconsistencies may arise because the subject's beliefs do not follow the chosen parametric family. For example, if an expert specifies a 5th, 50th, and 95th percentile of a distribution that is assumed to be normal, there may be no normal distribution with the specified percentiles. Or, an expert may specify a normal distribution by specifying a mean and a standard deviation and then when asked for the upper and lower quartiles give values that specify a different normal distribution.

Logical inconsistencies may also be present; for example, if a subject specifies a lower quartile of a distribution which is larger than an upper quartile specified, this is clearly an impossibility.

Kadane *et al.* (1980) suggest that probability distributions be elicited in several ways and the resulting inconsistencies reconciled. These methods are not particularly satisfactory and raise many difficult issues.

The elicitation technique known as the *device of imaginary results* has been advocated and used successfully (see Good, 1983). This technique requires subjects to give their beliefs after they are told about hypothetical data. From this, prior probabilities can be deduced, assuming that the subjects' beliefs obey Bayes' theorem. This method appears to have been

uccessful, although Phillips and Edwards (1966) and others have demonstrated that people do not necessarily use Bayes' theorem to update their eliefs correctly in a situation in which Bayes' theorem should apply. ome experimental evidence indicates that people are conservative and lo not adjust their beliefs as much as the rules of probability require. or a thoughtful discussion of the related experimental evidence and the orresponding psychological theories see von Winterfeld and Edwards 1986, Section 6.5 and 13.2).

3 ELICITATION FOR CLINICAL TRIALS

3.1 Why Is Elicitation Important?

Even without a Bayesian analysis it is important, before a trial begins, to document information and beliefs about the planned treatments. If the evidence is such that one treatment is firmly believed to be inferior, then the ethics of enrolling patients into the trial are questionable. As stated by, for example, Byar *et al.* (1990), "a trial should be open only to patients for whom the choice of recommended treatment remains substantially uncertain." Although Byar *et al.* do not quantify probabilistically what "substantially uncertain" means, it would clearly be helpful to document beliefs probabilistically in reviewing the ethical aspects of a trial.

Prior distributions can also used in the design of a clinical trial and, of course, prior distributions are also required for Bayesian monitoring and analysis of a clinical trial. Carlin *et al.* (1993, 1995) use the prior distributions elicited in Chaloner *et al.* (1993) to illustrate Bayesian monitoring for a toxoplasmosis prophylaxis trial.

Kadane (1986) suggests that subjective prior distributions be elicited from a number of experts, including a number of clinicians and patients, to form a "community" of prior distributions. Inferences can then be based on a consensus of posterior conclusions.

Spiegelhalter *et al.* (1994) recommend consulting a large number of experts and subsequently constructing a number of distributions, namely:

1. A "clinical" prior distribution by averaging prior distributions elicited from a large number of experts
2. A "vague" prior distribution leading to a posterior distribution proportional to the likelihood
3. A "skeptical" prior distribution representing a clinician unenthusiastic about the new therapy centered at the new therapy having no effect

4. An "enthusiastic" prior distribution centered at the new therapy having a large effect

Greenhouse and Wasserman (1995) (following Huber, 1973, and Berger, 1984) suggest that a single prior distribution π_0 be specified and a class of prior distributions be considered that are close to π_0 in some sense. A popular and tractable class is the "ϵ-contaminated class," which are mixtures with probability $(1 - \epsilon)$ on π_0 and ϵ on some other distribution from a specified class. This approach cannot easily, however, reflect a large amount of variability between opinions.

An excellent answer to the question "Why is elicitation important?" was given by Garthwaite and Dickey (1991), who said that "expert personal opinion is of great potential value and can be used more efficiently, communicated more accurately, and judged more critically if it is expressed as a probability distribution."

3.2 Elicitation for Clinical Trials

Freedman and Spiegelhalter (1983) describe a clinical trial of whether using a particular drug immediately following surgery for superficial bladder cancer is an improvement over the standard treatment of not using the drug. They elicit probability distributions from 18 clinicians on the magnitude of the effect on probability of nonrecurrence for 2 years. They ask clinicians about the difference in proportions in the control and treatment group. They describe an elicitation procedure in which clinicians are asked for a mode and then upper and lower bounds that are thought to be "very unlikely to be exceeded," and then the clinicians are asked to assess the probability of the effect being larger than a series of intermediate points. Freedman and Spiegelhalter report a diversity of opinions among the 18 clinicians. They also report results of asking the clinicians for a "range of equivalence." These ranges are based on the belief that before adopting a new treatment clinicians would demand not just that the new treatment is efficacious but also that it is enough of an improvement to counterbalance increased toxicity. The upper limit of the range is the lowest treatment effect required for the clinician to adopt the drug as standard care, and the lower limit of the range is the largest treatment effect for which the clinician would not use the drug as standard care. Within this range the treatments are, effectively, equivalent. The 18 clinicians reported a diversity of opinions. Freedman and Spiegelhalter suggest that such a diversity may be a particular phenomenon of a multicenter study.

Spiegelhalter and Freedman (1986, 1988) give an another example of a trial of chemotherapy in which seven consultant oncologists were interviewed and their prior distributions elicited using 20-minute structured interviews. Figure 2 of Spiegelhalter and Freedman (1988) gives the seven probability distributions, and there is a remarkable degree of consensus. The authors say "we note the consistency in judgement among "naive" subjects who had never undergone a similar exercise nor discussed it amongst themselves." As the opinions were so consistent and there were no clearly discordant opinions, taking a simple average seems reasonable. Spiegelhalter and Freedman (1988) also report the ranges of equivalence elicited for the treatments in this trial. The seven oncologists provide remarkably consistent beliefs with similar intervals.

Kadane (1994) describes a trial investigating two drugs designed to reduce hypertension after open heart surgery. As the response variable used was the deviation of arterial systolic blood pressure from a target of 75 mm Hg, regression models were used with several covariates. The elicitation method of Kadane *et al.* (1980) was used to elicit normal-gamma prior distributions from five experts. The experts had different beliefs: two preferred one drug for all types of patients considered and one expert always preferred the other.

Elicitation by interviewing experts individually is time consuming. Spiegelhalter *et al.* (1994) report an alternative approach using postal elicitation in several trials currently under way. The method is extremely simple and allows a large number of clinicians to report their beliefs. They ask clinicians for their probabilities that the effect falls in different intervals. Results from this method have yet to be reported.

Berry *et al.* (1992) take a different approach and describe subjectively assessing prior distributions for a large vaccine trial using primarily historical information. They use the collaborative expertise of the authors of the paper. Prior distributions are restricted to follow parametric distributions: gamma and F distributions. The authors mention that a computer program was written to represent graphically the parameters to be specified but provide little detail. They describe the prior distributions elicited as "reasonably open minded" and report the historical data and their experience upon which their prior distributions are based.

Chaloner *et al.* (1993) take yet another approach for a trial of a prophylactic treatment for toxoplasmosis in advanced HIV and AIDS patients. Like Freedman and Spiegelhalter (1983) they use a two-year probability: but this time it is the two-year probability of getting toxoplasmosis on the active treatment conditional on a placebo probability. They also make the assumption that the proportional hazards Cox regression model is

appropriate. They describe eliciting beliefs from a group of five AIDS experts: three physicians, an epidemiologist, and a person working on AIDS research. The experts are asked to specify a guess for the 2-year probability on placebo, and then conditional on that probability a distribution is elicited on the probability for the active treatment arm. Initially a parametric distribution is specified, using upper and lower quartiles, and this distribution is shown to the expert on the screen of a workstation. The expert then adjusts the distribution by hand, using the mouse, to represent his or her beliefs. Like Freedman and Spiegelhalter (1983), Chaloner *et al.* (1993) report a wide diversity of opinions: Figure 1 of their paper is a plot of the five distributions. The distributions are not all unimodal and could not easily be represented by a simple parametric class. Chaloner *et al.* also report that although all five experts were enthusiastic about the treatment effect and had high probabilities on a large efficacious effect, the subsequent trial data proved them all wrong. The treatment had no effect in preventing toxoplasmosis and, in fact, those receiving the active drug had a higher death rate, possibly due to a harmful toxic effect of the drug. All the experts assigned this outcome little or no probability. Chaloner *et al.* (1993) describe a simple computer program, written in the xlispstat environment of Tierney (1990), using dynamic graphics and mouse input. They also describe repeating the elicitation for three of the five experts using 3-year rather than 2-year probabilities to give a check on the assumptions made. They do not suggest methods for reconciling inconsistencies.

By comparison with other problems, a variety of methods are available for elicitation for clinical trials and relatively wide experience. It would be valuable to compare the different methods and examine whether the more complicated method of Chaloner *et al.* is an improvement over the simple methods of Spiegelhalter *et al.*

An example is given in the Appendix of using the method of Chaloner *et al.* to elicit opinions from three experts about a trial of prophylaxis for *Pneumocystis carinii* pneumonia (PCP) in patients with HIV infection or AIDS.

4 OTHER BIOMEDICAL APPLICATIONS

4.1 Medical Diagnosis

In their review, Wallsten and Budescu (1983) include a section reviewing experiments on probabilistic assessments in medical diagnosis. See also Spiegelhalter (1987).

4.2 Design of Experiments

In designing experiments clinicians have often provided opinions as to the magnitude of the effects they expect. These point predictions have traditionally been used for sample size and power calculations. For example, it is often the practice that the sample size of a clinical trial be chosen to have 80% power to detect a specified effect: either a "smallest clinically meaningful effect," which might be the upper limit of Freedman and Spiegelhalter's range of equivalence, or an "expected" effect. As clinicians are familiar with providing guesses for the purpose of making sample size calculations, it is a natural step to provide probability distributions representing uncertainty so that uncertainty can be incorporated into the design process.

Methods for elicitation can also be used to elicit prior distributions to be used in design. Flournoy (1994) describes just such an example for the design of a phase I clinical trial. Her elicitation required a group of physicians to discuss and collectively provide a sketch, by hand, of an upper and lower response curve. She describes the historical information available to the physicians. She also describes in detail how the physicians "grappled together with their priors for this interval, sketching and resketching them" and describes how the final curves were agreed upon. She reports some inconsistencies such as an instance of an assessed 2.5th percentile above the 97.5th percentile. She also reports that although the physicians were asked for 50% probability intervals, it was clear from listening to them that they were describing something much closer to a 95% interval. Interestingly, the physicians also described it later as providing "maximum and minimum" values. Flournoy describes how a normal and a gamma prior distribution were specified using the resulting sketches.

This kind of detail is rare in the literature, and it is refreshing to see an elicitation processes described—and especially refreshing to see all the problems with the process laid out clearly and honestly. Only by sharing and describing these kinds of experiences will good methods for elicitation be developed.

5 GENERAL RECOMMENDATIONS

The following subjective recommendations are based on my experience, primarily as reported in the Chaloner *et al.* (1993) study.

1. *Interactive feedback.* Forming beliefs is an iterative process and subjects like to be able to change their minds and work toward specifying a distribution. Presenting them with interactive feedback helps them formulate their ideas probabilistically. It could potentially enable the expert to reconcile inconsistencies subjectively, which seems preferable to having an automatic algorithm for doing so.
2. *Scripted interview.* A written script to structure the interview was found to be helpful in Chaloner *et al.* (1993). Freedman and Spiegelhalter (1983) also report using a "structured interview." In an interview experts will ask questions and it will be necessary to deviate from the script, but the script provides some uniformity across experts.
3. *Review.* The expert should be provided with the results of a systematic literature review.
4. *Percentiles.* Asking for quartiles is difficult. Like Flournoy (1994), Chaloner *et al.* (1993) report that even though experts were asked for 25th and 75th percentiles they interpreted these more like 2.5th and 97.5th percentiles to give intervals with probability content 0.95. A reasonable supposition is that experts are familiar with 95% confidence intervals, tend to interpret them as probability intervals, and tend to think about 95% probability intervals rather than upper and lower percentiles. It is probably better to ask for the 2.5th and 97.5th percentiles.
5. *Lots of experts.* Elicit opinions from as many experts as is practically feasible and from a variety of sources of expertise. Doctors, nurses, patients, scientists, and researchers all have opinions.

Other general comments are:

1. Postal elicitation (Spiegelhalter *et al.* 1994) seems promising but perhaps a little dangerous as so little is known about elicitation. There is no opportunity for discussion for clarification (such as when the expert does not understand what exactly a 75th percentile is) and no opportunity to deviate from the questions asked (such as when, as in Chaloner *et al.*, 1993, an expert prefers to use a different end point definition than the one planned). It also eliminates the opportunity for interactive feedback. It does have the advantage of allowing a large number of clinicians to have their beliefs easily elicited.
2. Avoid restricting beliefs to parametric families. These families may be useful as a starting point or may even be necessary in high-dimensional problems but artificially constrain beliefs.

6 CONCLUSIONS

Elicitation not only enables the many potentially useful Bayesian methods to be used in practice but also aids discussion and provides valuable documentation of the expectations of a clinical experiment before it begins. There is an urgent need for operational methods for elicitation of subjective probability distributions.

ACKNOWLEDGMENT

I am grateful to George Duncan and Dalene Stangl for helpful comments and to Winston Cavert, Carlton Hogan, and Frank Rhame for their opinions. Support was provided in part by NIAID contract NO1-AI-05073.

APPENDIX

A description is now given of eliciting a prior distribution for a trial of PCP prophylactic therapy. The method of Chaloner *et al.* (1993) is used and the script was adapted from the script described in that paper. The method assumes that data from the trial will be analyzed using the Cox (1972) proportional hazards regression model and the prior distribution will be noninformative on all but the regression parameter corresponding to the treatment assignment. A probability distribution is elicited on the proportion of patients experiencing the end point on the treatment arm conditional on a guess for the proportion on the control arm. That is, the distribution elicited is a conditional subjective probability distribution.

The experts had available the protocol document of the trial in question: "A randomized, comparative, prospective study of daily trimethoprim/sulfamethoxazole (TMS) and thrice weekly TMS for prophylactic therapy against PCP in HIV-infected patients" (CPCRA 006). The protocol document contained a review of relevant information in the literature. The trial began enrollment in 1992 and at the time of elicitation in September 1994 results remain confidential.

In each of the three elicitations performed the experts asked many questions and entered into a lengthy discussion about what they thought the trial would show. Their distributions, together with a summary of the discussions, are given later. The script, which gave a structure to the process, is given below.

The Script

People with AIDS often develop *Pneumocystis carinii* pneumonia (PCP). Trimethoprim/sulfamethoxazole (TMS) has been shown to fight PCP but it is not without side effects. The PCP-TMS study is designed to show which dose of TMS is safest and which works best. Daily TMS has been shown to be effective in preventing a recurrence of PCP in people who have already had it. There is also evidence that taking TMS three times a week may be effective against PCP and may cause fewer side effects than a daily dose.

The purpose of this exercise is to quantify your beliefs about the efficacy of the two treatment arms in the PCP-TMS trial. One treatment is one double-strength (DS) tablet of TMS daily and the other treatment is one DS tablet three times a week (Monday, Wednesday, Friday). The protocol document describes the eligibility criteria for the trial as "designed to include all patients for whom a primary care physician would prescribe prophylaxis, while excluding some patients for safety reasons."

Do you need to know more about the trial? Or about the eligibility criteria? The protocol document provides a literature review of relevant studies on PCP prophylactic therapy and AIDS/HIV and is here if you want to study it or refer to it.

The end point of the trial is an episode of *Pneumocystis carinii* pneumonia (PCP). Deaths are not included in the end point. Some patients will reach an end point and some will not. Only patients experiencing PCP will reach an end point. Patients who die without experiencing PCP will be removed from the analysis.

Think about a large number of people enrolled in the trial for 2 years and think about the proportion of people who will reach the PCP end point.

What is your best guess of the percentage of people assigned to the daily group who will experience PCP 2 years after enrollment? Please give your best guess—you will have uncertainty about this guess—but if you had to make a guess what would your guess be? Choose the value you think most likely. Your guess should relate to people similar to those expected to be enrolled in the trial—some of whom may supplement their study drug or reduce their dosage, some may fail to comply with the treatment, and some may develop intolerance to TMS—but think of the entire group of people assigned to an arm.

Now suppose your guess of this proportion turns out to be correct—the percentage of people on placebo experiencing PCP is exactly your best guess. Think about the people on the 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 two years *given* that the proportion experiencing PCP on the daily TMS arm is what you guessed.

Please specify an interval, by an upper number and a lower number, within which you think that the percentage of people experiencing PCP on the three times a week arm will lie. The interval should be such that you have probability 0.95 that the proportion will lie in the interval: probability 0.025 that the percentage will be higher than the upper limit and probability 0.025 that the percentage will be lower than the lower limit. In other words, the upper and lower limits are the 2.5th and 97.5th percentiles of your probability distribution on the percentage.

You will now be shown a plot of a probability distribution with the properties you have given. It is a smooth estimate of a distribution with the percentiles you specified. You can adjust the curve smoothly using the slider-dialogs to something that you think is reasonable. The curve is constrained to be of a particular shape.

You can now adjust the distribution to more accurately reflect your beliefs using the hand adjustment by changing the mouse mode to hand drawing. (If at any point you want to start again from the beginning then we can do so.)

The Results

Beliefs were elicited from three experts: two infectious disease MDs treating AIDS patients (experts A and C) and a person involved in the running and design of AIDS clinical trials (expert B). The three values for the best guess of the proportion experiencing PCP on the daily dose arms were 0.07, 0.15, 0.10. The probability distribution, conditional on this percentage, of the corresponding proportion of patients experiencing PCP on the three times a week arm is plotted on the top line of Figure 1. On each of the three plots the expert's best guess of the proportion for the daily arm is indicated by a cross (×). The corresponding distribution on the regression coefficient in a proportional hazards regression model was calculated for each expert and is plotted on the second row of Figure 1.

Expert A thought that either dose would be completely effective if the patient complied with the dose. That is, the only patients experiencing PCP would be those who either did not comply with treatment or became intolerant to TMS and could not comply. This expert thought that if a patient on the three times a week arm forgot to take a dose, the patient might become susceptible to PCP. On the daily dose, however, forgetting to take one dose occasionally would probably not increase susceptibility. Because of this, expert A typically prescribes TMS to be taken daily,

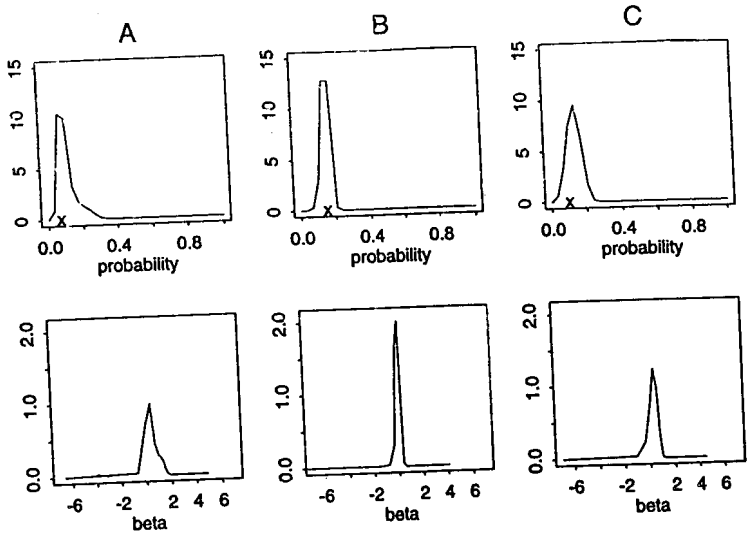

Figure 1 Prior distributions on the probability (top row) and on the regression coefficient (bottom row).

although expert A believes more people will develop intolerance on the daily dose. Expert A's guess for the proportion of end points in 2 years on the daily arm was 7% with a corresponding initial conditional interval for the three times a week arm of 3% to 20%. Expert A adjusted the initial plot to give the plot shown in Figure 1.

Expert B had the opinion that the two arms were equivalent. This expert guessed that the proportion of patients experiencing PCP on the daily arm would be about 15% and had a 95% interval of 11.25 to 18.75 (which is 15% ± 25% of 15%) for the percentage on the thrice weekly arm conditional on 15% on the daily arm. Although this expert claimed he had great uncertainty, his distribution was, in fact, the least variable of the three.

Like expert B, expert C also believed the two arms to be equivalent and indeed, in clinical practice, typically gives patients a choice of either daily TMS or TMS three times a week. Unlike expert A, expert C believes that developing intolerance to TMS will occur equally often on the two doses. Expert C's guess for the daily arm was 10% with an initial conditional interval of 5% to 20% for the thrice weekly arm.

All three experts had uncertainty about which of the two treatments would be best. This should be reassuring to the designers of the trial, as clinical trials should be designed to answer questions about which there is uncertainty.

REFERENCES

Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics*, ed. Kadane, J. B., pp. 64–144. North-Holland, Amsterdam.

Berry, D. A., Wolff, M. C., and Sack, D. (1992). Public health decision making: A sequential vaccine trial. In *Bayesian Statistics 4*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., pp. 79–96. Oxford University Press, Oxford, UK.

Black, P. and Laskey, K. B. (1989). Models for elicitation in Bayesian ANOVA: Implementation and application. *ASA Proceedings of Statistical Computing Section*, pp. 247–252.

Byar, D. P., Schoenfeld, D. A., Green, S. B., Amato, D. A. (1990). Design considerations for AIDS trials. *New England Journal of Medicine* 323: 1343–1348.

Carlin, B., Chaloner, K., Church, T., Matts, J. P., and Louis, T. A. (1993a). Bayesian monitoring of an AIDS clinical trail. *The Statistician* 42: 355–367.

Carlin, B. P., Chaloner, K. M., Louis, T. A. and Rhame, F. S. (1995). Elicitation, monitoring and analysis of an AIDS clinical trial. In *Case Studies of Bayesian Statistics in Science and Industry, 2*, eds. Gatsonis, C., Hodges, J., Kass, R., and Singpurwalla, N., pp. 48–89, Springer-Verlag, New York.

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

Chaloner, K. M. and Duncan, G. T. (1987). Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics A* 16: 511–523.

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

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Royal Statistical Soc. Ser. B* 34: 187–220.

Dickey, J. M., Dawid, A. P. and Kadane, J. B. (1986). Subjective-probability assessment methods for multivariate-t and matrix-t models. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. Goel, P. K., and Zellner, A. pp. 177–195. North-Holland, Amsterdam.

Flournoy, N. (1994). A clinical experiment in bone marrow transplantation: Estimating a percentage point of a quantal response curve. In *Case Studies in Bayesian Statistics*, eds. Gatsonis C., pp. 324–336, Springer-Verlag, New York.

Freedman, L. S. and Spiegelhalter, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician* 32: 153–160.

Freedman, L. S. and Spiegelhalter, D. J. (1992). Application of Bayesian statistics to decision making during a clinical trial. *Statistics in Medicine* 11: 23–35.

Garthwaite, P. H. (1992). Preposterior expected loss as a scoring rule for prior distributions. *Communications in Statistics—Theory Meth.* 21(12): 3601–3619.

Garthwaite, P. H. and Dickey, J. M. (1985). Double- and single-bisection methods for subjective probability assessment in a location-scale family. *Journal of Econometrics* 29: 149–163.

Garthwaite, P. H. and Dickey, J. M. (1988). Quantifying expert opinion in linear regression models. *Journal of the Royal Statistical Society Ser. B* 50: 462–474.

Garthwaite, P. H. and Dickey, J. M. (1991). An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making* 4: 17–31.

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

Good, I. J. (1983). *Good Thinking*. University of Minnesota Press, Minneapolis.

Greenhouse, J. B. and Wasserman, L. (1995). Robust Bayesian Methods for Monitoring Clinical Trials. *Statistics in Medicine* 14: 1379–1391.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association* 70: 271–294.

Hogarth, R. M. (1987). *Judgement and Choice*, 2nd ed. Wiley, New York.

Huber, P. (1973). The use of Choquet capacities in statistics. *Bull. Internat. Statist. Inst.* 45: 181–191.

Kadane, J. B. (1986). Progress toward a more ethical method for clinical trials. *The Journal of Medicine and Philosophy* 11: 385–404.

Kadane, J. B. (1994). An application of robust Bayesian analysis to a medical experiment (with discussion). *Journal of Statistical Planning and Inference* 40: 221–232.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association* 75: 845–854.

Kahneman, D., Slovic, P., and Tverskey, A., eds. (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK.

Keeney, R. and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York.

Kyberg, H. E. Jr. and Smokler, H. E., eds. (1980). *Studies in Subjective Probability*. Krieger, New York.

Laskey, K. B. and Black, P. (1989). Models for elicitation in Bayesian analysis of variance. *Proceedings of Computer Science and Statistics: 21st Annual Symposium on the Interface*, Florida pp. 242–247.

Lindley, D. V. (1985). *Making Decisions*, 2nd ed. Wiley, New York.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 2B, *Bayesian Inference*. Arnold, London.

Phillips, L. D. and Edwards, W. (1966). Conservatism in simple probability inference tasks. *Journal of Experimental Psychology* 72: 346–357.

Ravinder, H. V., Kleinmuntz, D., and Dyer, J. S. (1988). The reliability of subjec-

tive probabilities obtained through decomposition. *Management Science* 34: 186–199.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66: 783–801.

Spiegelhalter, D. J. (1987). Probability expert systems in medicine: practical issues in handling uncertainty (with discussion). *Statistical Science* 2: 25–34.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* 5: 1–13.

Spiegelhalter, D. J. and Freedman, L. S. (1988). Bayesian approaches to clinical trials. In *Bayesian Statistics 3*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., pp. 453–477. Oxford University Press, Oxford, UK.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. (1993). Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine* 12: 1501–1511.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. (1994). Bayesian approaches to randomized trials. *J. Royal Statistical Soc. Ser. A*, 157: 357–416.

Tierney, L. (1990). *LISP-STAT, an Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.

von Winterfeldt, D. and Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press, Cambridge, UK.

Wallsten, T. S. and Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science* 34: 186–199.

Winkler, R. L. (1967a). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association* 62: 776–800.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association* 66: 675–685.

Winkler, R. L. (1967b). The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association* 62: 1105–1120.