

Bayesian Methods

A clinical trial is an experiment carried out to gain knowledge about the relative benefits of two or more treatments. Typically, this is part of a gradual accrual of knowledge: a trial to confirm benefits in a large population may follow much careful work on smaller scale studies, or a study may be asking essentially the same question as several other studies. Conventionally, clinical trials are analyzed formally as an individual trial, and their contribution to accruing knowledge then assessed informally. However, increasingly the technique of **meta-analysis** is used to combine the information from similar trials into a formal summary.

CAM003-

More generally, researchers may wish to frame the following questions: “What do we think about the relative benefits of the treatments before knowing the results from this trial?” “What information can be gained from the results of this trial?” “Considering the results of this trial in the light of previous understanding, what do we now think about the relative benefits of the treatments?”

If this seems too subjective, an alternative way of casting this framework is to ask: “What is the previous evidence on the relative benefits of treatment?” “What is the current evidence from this trial?” “What is the updated evidence, once we combine the previous with the new evidence?”

The concept of updating of beliefs or evidence is the essence of Bayesian statistics. This article explains the essential concepts through a simple example, and then discusses some of the issues raised, namely the legitimate sources of previous beliefs or evidence, including the question of subjectivity, and implications for the design of trials and Bayesian reporting of clinical trials. A particular area of application is data monitoring (*see* **Data and Safety Monitoring**).

Most of the discussion is in the context of two-group parallel trials, partly for simplicity of exposition, but mainly because of the pre-eminence of this design in practice. The framework is, however, completely general, and applies to more complex designs. For the combination of results of several trials, possibly with other evidence, Bayesian meta-analysis is outlined. Clinical trials are often used as part of wider decision-making processes. Bayesian statistics is sometimes set in the context of decision-making,

and the implications of this are discussed. Finally, there is a note on computational issues.

An Example

Consider the following example: after a heart attack, thrombolysis is often indicated. There is a tension between whether this is done at home once the ambulance arrives, which confers the advantage of speed, or in hospital, which is a more optimal environment, but necessitates a delay in treatment. The GREAT trial was run to compare these two strategies [8]. When the trial reported, there were 13 deaths out of 163 patients in the home group, and 23 out of 148 in the hospital group. The authors estimated a reduction in mortality of 49%. Some commentators were skeptical that a halving of mortality was really possible. Pocock & Spiegelhalter [14] carried out a Bayesian analysis. They judged, ignoring the trial results, that home treatment probably conferred some benefit, say a 15%–20% reduction, but that a 40% reduction, let alone a halving of mortality, was fairly unlikely. These beliefs are termed the *prior distribution*. The evidence from the trial is described through the *likelihood function*. Combining these two gives the *posterior distribution* of beliefs. This gives an estimate of the reduction of mortality of about 25%, but still says that the extremes of no effect or of a halving of mortality are unlikely.

Differences from classical analyses include the incorporation of prior beliefs, the absence of P values, and the absence of any idea of hypothetical repetitions of such studies. The posterior estimate of effect and its surrounding uncertainty via a *credibility interval* is analogous to a classical point estimate and its associated confidence interval, but has a direct interpretation in terms of belief. As many people interpret a confidence interval as the region in which the effect probably lies, they are essentially acting as Bayesians.

Mathematical Formalization

We can express the preceding analysis formally as

$$P(\theta|\text{data}) \propto P(\text{data}|\theta) \times P(\theta).$$

$P(\theta)$ is the prior distribution expressing initial beliefs about the parameter of interest. In the example, this would be the difference in mortality rates,

2 Bayesian Methods

described using the Normal distribution. $P(\text{data}|\theta)$ is the likelihood function, expressing the statistical model of variability for the data given the parameters. In the example, a Normal distribution is assumed for the difference in binomial proportions. $P(\theta|\text{data})$ is the posterior distribution of beliefs. Its shape depends on the previous two distributions, but where the prior distribution and likelihood are assumed to be Normal, the resulting posterior distribution is Normal.

The equation is usually expressed and worked with in its proportional form: if needed, the constant of proportionality is obtained by integrating the right-hand side with respect to θ , which ensures that the posterior distribution is properly defined, integrating to one.

Clinicians will have views on how different these treatments need to be before it becomes unethical to randomize patients between them. For example, some may think that if the home thrombolysis is no worse than hospital and no more than 20% better than hospital treatment, that **randomization** is acceptable, whereas once there is reasonable evidence that the difference is outside this range, a decision can be reached as to which is preferable. This range is often termed the *region of equivalence*. One end is essentially the same as the “clinically important difference” used conventionally when deciding how large studies should be, and the other end may be the point of no difference. An alternative is to have the range symmetrical about the point of no difference: the range of equivalence in the GREAT trial might be that home is no more than 20% better or worse than hospital. (This is often used in **bio-equivalence studies**.)

Sources of Prior Distributions

The Bayesian approach just outlined gives a framework for updating beliefs or evidence. There are several possible sources of prior distributions. Spiegelhalter et al. [18], recommend that there is no need to select just one, and outline a *community of priors* that can be used for interpretation.

The *reference prior* represents minimal prior information. This is the least subjective, and analyses based on this act as a useful baseline against which to compare analyses using other priors. The *clinical prior* formalizes the opinion of well-informed specific individuals. The *skeptical prior* formalizes the

belief that large treatment differences are unlikely. This can be set up, for example, as having a mean of no treatment effect, and only a small probability of the effect achieving a clinically relevant value. By contrast, the *enthusiastic prior* can be specified, for example, with a mean equivalent to a clinically relevant effect, and only a small probability of no effect, or worse.

The reference prior, skeptical prior, and enthusiastic prior are essentially mathematical constructs, calibrated using points such as that of no effect, and the clinically relevant effect. By contrast, a clinical prior is intended to represent the current state of knowledge. Where possible, it should be based on good evidence, such as a meta-analysis of relevant randomized controlled trials. Where this is not possible, evidence from nonrandomized studies may be needed. Alternatively, subjective clinical opinion may form the basis of a prior distribution. Elicitation of opinion can be carried out using techniques such as interviews, questionnaires or interactive computer packages with feedback [5, 13, 21]. These are not mutually exclusive: for example, subjective judgment about relevance or changed circumstances may be needed to modify results from an objective meta-analysis.

For example, in a Bayesian analysis of a cancer clinical trial comparing high-energy neutron therapy versus the standard of photon therapy [18], priors from two sources were used. The first was from a survey of interested clinicians, which showed beliefs favoring neutron therapy; the second was from a meta-analysis of related studies of low-energy neutron therapy, which showed a detrimental effect compared with placebo (*see Blinding or Masking*). The data from an interim analysis (*see Data and Safety Monitoring*) was against neutron therapy, and, starting from either prior, the posterior belief in a worthwhile benefit was small, with the weight of posterior evidence on a harmful effect. The data monitoring committee (*see Data and Safety Monitoring Boards*) had actually stopped the trial at that stage (on classical analyses), and the Bayesian analyses express explicitly the wisdom of that decision.

Region of Equivalence

The region of equivalence is the area in which *equipoise* (*see Ethics*) exists: a patient or his/her

CAR001-

CAE002-

doctor is indifferent to which of the two treatments is used. Whilst there is a reasonable probability that the treatments are in equipoise, randomization is ethical.

There are close parallels with the specification of the alternative hypothesis in the design of clinical trials based on the classical statistical paradigm. For a Bayesian analysis of a classically designed trial, an obvious choice for a region of equivalence is to take the points associated with the null and alternative hypotheses. When Bayesian thinking is informing the design, the range of equivalence is often elicited from clinicians using similar techniques to those used for elicitation of prior beliefs. In the neutron therapy trial described above, clinicians had also been asked about how good neutron therapy would need to be before it should be routinely used. They said (on average) that one-year survival of 50% would need to be increased to 61.5%. The range of equivalence was then taken as being between no improvement and an improvement of this magnitude.

The region of equivalence provides a useful benchmark for the design of trials, for reporting of results and for data-monitoring. These are discussed in more detail below. The region of equivalence is often determined in a relatively informal fashion by clinicians. Wider questions, about whose equipoise is really relevant, and what considerations should inform this, point towards a decision-making perspective. These are considered at the end of the article.

Bayesian Power

Classical power calculations for clinical trials are carried out by specifying a null hypothesis that two treatments do not have different effects on the outcome of interest, and an alternative hypothesis that the difference in outcome is equal to some pre-specified value. The risks of wrong decisions under these two hypotheses are then fixed at chosen levels, which then determine the necessary sample size. These calculations are essentially conditional on the choice of the alternative hypothesis. There is as yet no consensus on a Bayesian approach to **sample size determination** for clinical trials. Some advocate focusing on a reasonable probability of getting a posterior interval less than a certain width, while others take an explicit decision-making perspective, with utilities either essentially “information”, or some trading-off of health benefits and cost. A wide-ranging discussion of Bayesian sample size calculation can be found in

a special issue of *The Statistician* [16]. See also [18] and [20].

Data Monitoring

In many trials, results accrue fast relative to patient recruitment. In this situation, data-monitoring committees are often set up to review the data to ensure that equipoise still exists, and it is still ethical to enter patients into the trial. Statistically, the challenge is to guard against stopping a trial too early, as an overreaction to early dramatic results, whilst protecting new trial patients from inappropriate randomization. From a classical statistical perspective, this is often formalized in terms of adjusting significance levels.

One Bayesian approach [18] formalizes it differently. At the start of the trial, a skeptical prior (see above) is used to represent the view that there is not too much difference between treatments. Only when the data dominate this sufficiently would early stopping be considered. The effect of such a prior is to put a brake on early results. For a trial of esophageal cancer comparing surgery with pre-operative chemotherapy and surgery, this approach was used [6]. It has been shown that there is a close tie-up between this approach and classical group sequential designs (*see Sequential Methods*), in that a particular design, say with five interim analyses and a Pocock boundary, corresponds to a Bayesian procedure with a particular choice of prior distribution [7].

An alternative Bayesian approach to monitoring takes a much more decision-theoretic perspective. For a trial of influenza vaccination of Navajo children, monitoring included explicit consideration of future children and their risk of influenza [4].

Complex Trial Designs

The two-group parallel trial described so far is important, but not the only trial design. For more complex designs, the same framework of prior distribution/likelihood/posterior distribution outlined above still holds, although because there are more parameters, careful specification is needed, for example in parameterization. Bayesian methods have been developed for other designs, including **crossover** trials [9] and **factorial** trials [1, 15].

CAS002-

CAC014-
CAF001-

4 Bayesian Methods

Bayesian Reporting of Clinical Trials

A good report of a trial specifies the question being addressed, describes the design and conduct of the trial, gives results, makes formal statistical inference from these, discusses sensitivity to assumptions, and then interprets the trial in the context of other relevant research. Many of these do not differ from usual good practice, but some aspects can benefit from formalization using Bayesian procedures [11, 19].

The results of the trial should be described clearly, and in enough detail that another reader could carry out alternative analyses if desired. Formal statistical inference follows the prior/likelihood/posterior analysis outlined above. The posterior distribution then represents a summary of beliefs/evidence about the parameters of interest. This is most fully represented graphically, but can be further summarized by giving a *95% credibility interval*. In addition, it is often useful to give the probability that an effect is in a particular region, for example the probability that the parameter lies above the region of equipoise, or, for a bio-equivalence study, the probability that the parameter lies inside it.

The results section of the report should certainly include an analysis that starts from an uninformative prior distribution. If there are other well-specified prior distributions, then an analysis using these can also be presented in full. For example, if Bayesian data-monitoring has been used, then analyses with relevant prior distributions are appropriate. Sensitivity analyses should also be carried out. These may be for sensitivity to the specification of the prior distribution, but also to the specification of other parameters in the model. For example, Grieve [9] present plots for a bio-equivalence study looking at sensitivity to prior beliefs on the treatment effect. Sensitivity to other choices of the region of equipoise may be needed.

The discussion section of the report often contains more speculative interpretation. This can usefully be formalized through Bayesian analysis. If other opinions can be captured, for example by a skeptical or enthusiastic prior distribution, then the appropriate posterior distributions can be presented here. If there are other similar studies, then a Bayesian meta-analysis (see below) can be used to provide a combined estimate of the effects of interest.

In all Bayesian reporting the separate elements of the prior distributions and likelihood should be clearly specified and appropriately justified, so that the posterior distribution may be clearly interpreted. The likelihood should not be controversial, since it comes from the data, but specification of prior distributions is more difficult. A good rule of thumb is that if the prior distribution is based on belief, then the posterior distribution should be interpreted as an updated statement of belief, but if the prior distribution represents a summary of hard evidence, then the posterior distribution represents an updated summary of hard evidence.

Bayesian Meta-Analyses of Clinical Trials

Bayesian statistics is essentially about the updating of evidence. So far in this article the focus has been on individual trials, but where several trials address essentially the same question, a combined analysis is desirable. Bayesian meta-analysis extends Bayesian ideas used for a single trial to multiple trials. Previous evidence is expressed through *prior distributions* about quantities of interest: in a meta-analysis of binary outcomes, this will include, for example, the log odds ratio. Current data are expressed through the *likelihood*, based on an appropriate model. The *posterior distribution* for quantities of interest can then be obtained. The Bayesian framework also allows calculation of the probability that the odds ratio is at least say 1, or at least 3, which cannot be done in the classical framework.

After a careful search for all relevant trials, it seems strange to combine objective trial data with subjective opinion. In the meta-analysis context, it may be reasonable to use noninformative priors, which give intuitively interpretable results. This is particularly true for the main comparison. However, it may be useful to bring in judgment on some of the other parameters, on which the trials are less helpful, such as the size of the random effects. It is important to carry out sensitivity analyses on assumptions made. Examples of Bayesian meta-analyses include modeling random effects in a meta-analysis in urinary tract infections [17], incorporating external evidence on heterogeneity in a trial in cirrhosis [10] and modeling heterogeneity in relation to underlying risk [22].

Decision-Making with Clinical Trials

The focus in this article has been on the estimation of effects of interest using the accrued evidence. The purpose of accruing evidence is to make decisions. Bayesian statistics leads naturally towards explicit decision-making.

There is some debate as to whether clinical trials are, in themselves, decision-making contexts. Some (Lindley and others in discussion of [18]) argue they are, whereas Spiegelhalter et al. [19] argue that an individual clinical trial can be put to a variety of purposes, and so it is better not to construe the trial as a decision in itself.

Ashby & Smith [2] argue more generally that evidence-based medicine is about making decisions and the Bayesian approach is a natural one to adopt. When a decision is to be made, the following should be identified: the decision-maker, the possible actions, the uncertain consequences, the possible sources of evidence, and the utility assessments required. For example, a patient is diagnosed with esophageal cancer. He is advised that until recently routine treatment has been surgery, but a new suggestion is to precede the surgery by a course of several weeks of chemotherapy. The *decision-maker* is the patient, who may effectively delegate to his doctor. The *possible actions* are to undergo surgery, or to opt for the combination treatment. The *uncertain consequences* are the length of his survival, and side-effects (such as severe nausea), and the delay in completion of treatment. The possible *sources of evidence* relating to his expected survival come from routine data such as cancer registries, and relating to the additional benefit of combined treatments from clinical trials. The *utility assessment* required is the patient's tradeoffs between extra survival, side-effects and time spent undergoing treatment. Within this framework, evidence from clinical trials plays a very important role.

Computation

Some of the simple analyses in this article can be done analytically, using nothing more than a hand calculator. BUGS is a general-purpose package written to facilitate the fitting of complex Bayesian models [21]. It is available from <http://www.mrc-bsu.cam.ac.uk/bugs/>, and can handle the kinds of analyses referred to in this article.

Bayesian Clinical Trials in Practice

For many years the principles of Bayesian statistics have been well understood. Implementation in practical areas such as clinical trials has been hampered, until recently, by computational complexity. However, with the growth in modern computing power, the situation is changing rapidly. Analyses of real complex studies are relatively recent, and their use as the first or primary approach even newer. A Bayesian analysis now offers an intuitive approach, combined with the power to deal with complexity when necessary. Bayesian clinical trials, and integrated summaries of them using Bayesian analyses, are finding their place in practice.

A comprehensive discussion of Bayesian clinical trials with excellent references based on systematic review, can be found in Spiegelhalter et al. [19] and several case studies of Bayesian clinical trials in Berry & Stangl [3] and Kadane [12].

References

- [1] Abrams, K.R., Ashby, D., Houghton, J. & Riley, D. (1996). Tamoxifen and cyclophosphamide – synergists or antagonists?, in *Bayesian Biostatistics*, D. Berry & D. Stangl, eds. Marcel Dekker, New York.
- [2] Ashby, D. & Smith, A.F.M. (2000). Evidence-based medicine as Bayesian decision-making, *Statistics in Medicine* **19**, 3291–3305.
- [3] Berry, D.A. & Stangl, D. (1996), *Bayesian Biostatistics*. Marcel Dekker, New York.
- [4] Berry, D.A., Wolff, B.C. & Sack, D. (1992). Public health decision making: a sequential vaccine trial, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 79–96.
- [5] Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: a review, *Statistical Science* **10**, 273–304.
- [6] Fayers, P.M., Ashby, D. & Parmar, M.K.B. (1997). Bayesian data monitoring in clinical trials, *Statistics in Medicine* **16**, 1413–1430.
- [7] Freedman, L.S. & Spiegelhalter, D.J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials, *Controlled Clinical Trials* **10**, 357–367.
- [8] GREAT Group (1992). Feasibility, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial, *British Medical Journal* **305**, 548–583.
- [9] Grieve, A.P. (1985). A Bayesian analysis of the two-period crossover design for clinical trials, *Biometrics* **41**, 979–990.
- [10] Higgins, J.P.T. & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* **15**, 2733–2749.

6 Bayesian Methods

- [11] Hughes, M.D. (1991). Practical reporting of Bayesian analyses of clinical trials, *Drug Information Journal* **25**, 381–393.
- [12] Kadane, J.B. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*. Wiley, New York.
- [13] Parmar, M.K.B., Spiegelhalter, D.J. & Freedman, L.S. (1994). The chart trials: Bayesian design and monitoring in practice, *Statistics in Medicine* **13**, 1297–1312.
- [14] Pocock, S.J. & Spiegelhalter, D.J. (1992). Grampian region early anistreplase trial, *British Medical Journal* **305**, 1015.
- [15] Simon, R. & Freedman, L.S. (1997). Bayesian design and analysis of two \times two factorial clinical trials, *Biometrics* **53**, 456–64.
- [16] Smeeton, N.C. & Adcock, C.J., eds. (1997). Sample size determination, *Statistician* **46**, 129–291 (special issue).
- [17] Smith, T.C., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine* **14**, 2685–2699.
- [18] Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials, *Journal of the Royal Statistical Society* **157**, 357–416.
- [19] Spiegelhalter, D.J., Myles, J.P., Jones, D.R. & Abrams, K.R. (2000). Bayesian methods in health technology assessment: a review, *Health Technology Assessment* **4**(38).
- [20] Tan, S.B. & Smith, A.F.M. (1998). Exploratory thoughts on clinical trials with utilities, *Statistics in Medicine* **17**, 2771–2791.
- [21] Thomas, A., Spiegelhalter, D.J. & Gilks, W.R. (1992). BUGS: a program to perform Bayesian inference using Gibbs sampling, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 837–842.
- [22] Thompson, S.G., Smith, T.C. & Sharp, S.J. (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis, *Statistics in Medicine* **16**, 2741–2758.

DEBORAH ASHBY