Statistical Science
2004, Vol. 19, No. 1, 175–187
DOI 10.1214/088342304000000044

Bayesian Statistics and the Efficiency and Ethics of Clinical Trials

Donald A. Berry

Abstract. The Bayesian approach is being used increasingly in medical research. The flexibility of the Bayesian approach allows for building designs of clinical trials that have good properties of any desired sort. Examples include maximizing effective treatment of patients in the trial, maximizing information about the slope of a dose–response curve, minimizing costs, minimizing the number of patients treated, minimizing the length of the trial and combinations of these desiderata. They also include standard frequentist operating characteristics when these are important considerations. Posterior probabilities are updated via Bayes' theorem on the basis of accumulating data. These are used to effect modifications of the trial's course, including stopping accrual, extending accrual beyond that originally planned, dropping treatment arms, adding arms, etc. An important aspect of the approach I advocate is modeling the relationship between a trial's primary endpoint and early indications of patient performance—auxiliary endpoints. This has several highly desirable consequences. One is that it improves the efficiency of adaptive trials because information is available sooner than otherwise.

Key words and phrases: Bayesian updating, decision analysis, predictive probabilities, clinical trials, adaptive designs, clinical ethics, auxiliary endpoints, extraim analyses.

1. INTRODUCTION

The purpose of clinical trials is to learn about therapies or interventions under consideration. The experimental units are people. The Belmont Report (1979) provides guidance for conducting medical research with human subjects. It contrasts "clinical practice" and "clinical research." The former "refers to interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success. The purpose of medical or behavioral practice is to provide diagnosis, preventive treatment or therapy to particular individuals. By contrast, the term 'research' designates an activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge (expressed, e.g., in theories, principles, and statements of relationships). Research is usually described in a formal protocol that sets forth an objective and a set of procedures designed to reach that objective."

This view of clinical research is neither universally held nor appreciated, even among researchers. In a survey of 547 U.S. oncologists, Joffe and Weeks (2002) report that many "viewed the main societal purpose of clinical trials as benefiting the participants rather than as creating generalizable knowledge to advance future therapy." They indicate that "this view ... conflicts with the established principles of research ethics" and that "there may be a belief among some cancer specialists that clinical trials seamlessly unite research and therapy. This belief, which challenges conventional doctrines of research ethics, has important implications for how we conduct, review, and regulate clinical trials, as well as how we present them to patients."

I have a different perspective on this question. Clinical trials can and should be designed both to lead

Donald A. Berry is Professor and Chair, Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, Texas 77030, USA (e-mail: dberry@mdanderson.org).

to generalizable knowledge and to benefit participants in the trials.

There is an inevitable tension between research and practice. If there were no clinical trials but instead patients were assigned therapy that they and their physicians thought best, then many potentially useful therapies would go untested. Benefits of therapies would be confounded with patients' characteristics and desires. For example, if a therapy that was overlooked by most of the medical profession were to be used by a few physicians and the patients did relatively well, we could not be sure that the patients were typical or that their physicians had given correlative care that was typical. Moreover, reports comparing clinical outcomes of patients choosing different therapies would be based on small sample sizes. Many of the most important advances in modern medical care would never have been recognized had there not been randomized clinical trials that enrolled thousands and even tens of thousands of patients.

Randomization eliminates bias in the assignment of patients to competing therapies. It is the only generally accepted way to eliminate bias. Physicians who regard different therapies being considered as having different overall effectiveness cannot ethically enroll patients in the trial. That is to say, physicians who participate in a randomized clinical trial must be in "equipoise." The patients who enroll in clinical trials enable us to learn about the relative benefits and risks of the competing therapies. Their participation means that later patients will receive better care, on average. This in turn implies an overall positive drift in the effectiveness of medical therapies—which is not to say that *every* putative advance is in fact an advance!

So it seems to be a win–win situation. Patients in the trial get therapies regarded as among the best available and as being similar to other therapies in the trial. Patients who come after the trial benefit from whatever is learned during the trial. But there is a rub. Consider a randomized trial comparing therapies A and B on the basis of a short-term endpoint. The final results indicate that A is significantly better than B. Consider the last patient in the trial who was randomized to therapy B. The results from the trial at the time this patient was treated must have been at least suggestive that therapy B was inferior. Such a circumstance is unavoidable for short-term endpoints because the data start pointing toward the eventual winning therapy before the end of the trial when the results are taken to be conclusive. Should not that patient's therapy be switched to A? If so, where do we stop in moving backward in time through the trial? If we go way back to the time at which the results started pointing toward a particular therapy and switch therapeutic assignments, then there is a good chance the designated winner is in fact inferior. We may never learn the truth.

The standard resolution of this conundrum is to avoid looking at the results of ongoing trials, or at least to restrict the looking to a data monitoring committee that is operating under prespecified guidelines concerning when a trial can be stopped. Such committees carry the onerous burden of ensuring that patients in the trial are not ill-treated while protecting the scientific integrity of the trial (Ellenberg, Fleming and DeMets, 2002). This burden epitomizes the tension between research and delivering good therapy. I will not discuss the decision-making process of data monitoring committees, except to say that the most useful set of guidelines for such committees involves (Bayesian) predictive probability calculations of future results on the basis of results that are currently available. The many data monitoring committees upon which I have served and to whom I report rely on such calculations. An example is reported in Lewis et al. (2001).

In the other extreme from the example above, suppose that a trial is nearing the end of its accrual goals. The objective is to decide whether therapy A is superior to therapy B on the basis of efficacy. Therapy A has many serious side effects and therapy B is relatively benign. The results to date are pretty clearly suggesting that the two therapies have identical efficacy. In particular, the predictive probability of achieving superiority eventually is very small. Continuing the trial to its planned sample size will expose additional patients to a toxic therapy A with little evident compensating benefit. Stopping at this point was not considered in the protocol. But it is difficult to imagine that continuing the trial is reasonable, regardless of whether stopping was considered explicitly in the trial's design.

A culprit in the above examples is striving to have a fixed sample size. Sample sizes are usually determined via power considerations by averaging over the possible results. Two types of results that are available toward the end of the trial are considered in sample size determinations. Some results are very clear in pointing to one conclusion or another, including the conclusion of the equivalence of the therapies. For other results, uncertainty about the conclusions will still exist and additional data will help us to understand which hypothesis is correct. In the former case, continuing is unnecessary and in the latter case it is important. The problem with fixed-sample-size

calculations is that they average over both types of results; the sample size is too large in the first case and too small in the second.

The design of modern clinical trials has moved an iota from having fixed sample sizes. They employ interim analyses at specific time points during the accrual and, in studies with time-to-event endpoints, after the accrual. Standard frequentist methods for early stopping of a trial focus on preserving the overall type I error rate. These methods tend to be very conservative and to call for stopping only when the results are quite extreme. Such methods are steps in the right direction, but they are not big enough.

In this article I address expanding the horizons of the standard frequentist approaches to the design of clinical trials. The approach I advocate is radical in medical research but common in scientific enquiry more generally: look at the data! And let it guide the future course of the trial. The flexibility of the Bayesian perspective facilitates taking this tack (Berry, 1993). In this article I consider a variety of types of actions during the course of a clinical trial. These include stopping the trial early, as discussed above. They also include adaptively assigning patients to therapies (drugs, drug combinations or different doses of a drug) that are performing better, adding and subtracting treatment arms and extending the accrual beyond that originally targeted when the answer to the question posed is still not known.

I will describe actual settings and actual trials. Many of the designs I consider have been used in clinical trials and others are being considered by investigators and pharmaceutical companies for future trials. The goals are (i) to more effectively use patient resources while treating patients in clinical trials effectively and (ii) to identify better drugs and other therapies more rapidly, moving therapies more quickly through the development process. For some of the designs these goals are made explicit through the use of decision analysis. Other designs are ad hoc and their performance characteristics are evaluated using simulation.

Consistent with the Bayesian approach, the designs presented here exploit the available evidence regarding the therapy in question and also as regards related therapies. In addition, information gleaned during an ongoing clinical trial is incorporated via Bayes' theorem into what was known at the start of the trial (Berry, 1996; Spiegelhalter, Freedman and Parmar, 1994). However, to adhere to conventional and regulatory standards, I simulate type I error rates and power characteristics of the design proposed. If the type I error rate is too high for regulatory acceptance, then I modify the design to bring it into line with the regulatory norms. In such cases the Bayesian approach serves as little more than a tool—a very handy tool!—for developing efficient and ethical designs that have acceptable frequentist properties.

Returning to the point of the opening part of this section, the tension between clinical research and clinical practice is inevitable. No approach can completely overcome this tension. A virtue of the Bayesian approach is that it reckons with the conflicting desiderata in a completely explicit way. Decision making in clinical research is never easy, even when the trade-offs are spelled out. But Bayesian decision analysis clearly lays out the pros and cons, and it balances them in such a way as to maximize the benefit to society or to a portion of society. Altering the balance alters the decision, and so sensitivity analyses with respect to the utilities and prior distributions may be necessary.

The Bayesian approach does not resolve all the ethical issues involved in clinical research. No approach can do this. However, taking a Bayesian approach can lead to *better* treatment for patients both in and out of clinical trials by providing more efficient designs of trials and of drug development programs more generally.

2. PREDICTIVE PROBABILITY

Predictive probability is an enormously important contribution of the Bayesian approach. Without it, the Bayesian approach would be much less compelling. Predictive probabilities are essential for designing clinical trials from a Bayesian perspective. As I indicated in the Introduction, it is natural and useful for monitoring ongoing trials.

This is not an appropriate forum for developing the mathematics of predictive probabilities, but I will give an example. Consider the simple case of a binomial experiment, with a sample size of n and a success rate of θ. There are $n + 1$ possible numbers of successes. Conditioning on θ, these have binomial probabilities. But θ is unknown (why else conduct the experiment?). Bayesians condition on what is known and associate probability distributions with unknown quantities. In this instance, θ has a prior probability distribution that reflects the available information. Predictive probabilities are conditional on the prior distribution (or the current distribution) of θ but they are not conditional on θ. Namely, $P(k \text{ successes}) = EP(k \text{ successes}|\theta)$, which, in the case θ has a beta distribution, are the familiar beta-binomial probabilities.

Other types of sampling, other types of prior distributions and the possibility of covariates give rise to more complicated predictive distributions. Indeed, in many realistic settings, predictive probabilities cannot be calculated analytically. However, modern computational methods easily skirt this erstwhile stumbling block.

Predictive probabilities incorporate two sources of uncertainty. One is the usual sampling variability (in the example above, that inherent in the binomial distribution for fixed θ). The other is the uncertainty about the various parameters, such as that in θ above. Considering only one source of variability overestimates the precision in one's ability to predict future results. Of the two sources, the one that is more commonly ignored is that in θ. (For example, a common—and unfortunate—attitude to making power calculations is to assume that a proportion observed in a previous study is the true population proportion in the next study.) This can lead to making the wrong decisions about a trial's sample size and about whether a trial with a particular design is appropriate.

3. DECISION ANALYSIS AND TRIAL DESIGN

Choosing a design for a clinical trial is making a decision. The decision has ramifications that can be explicitly stated and considered. Formal decision analysis is second nature to many statisticians, but it is surprisingly obscure to other scientists. Following any particular course of action—including the design of a clinical trial—leads to an outcome in a set of possibilities that is determined by the course of action. For decision purposes, a pair of numbers is associated with each possible outcome: the outcome's predictive probability and its utility. Averaging the utility with respect to the predictive probabilities gives the utility of the course of action. (However, the mere process of evaluating the predictive probabilities and utilities can be enormously beneficial to decision makers, even without averaging the utilities.)

Just as for other decisions, trial designs are ordered by their utilities—see Berry (1995, 1996), Berry and Eick (1995), Clemen (1991), Lewis and Berry (1998) and Sox, Blatt, Higgins and Marton (1988). Designs having maximal utility are called "optimal." It is possible and even likely that the utility assessment has not considered all relevant aspects. Therefore, one may reasonably select a nonoptimal trial design. However, before choosing a design that has substantially less utility than the maximum, one should endeavor to incorporate the utility aspects not previously considered and redo the analysis.

Utilities can be determined in terms of the effective treatment of patients, both those within and those outside the trial. Utilities can be economic, as when a pharmaceutical company is contemplating sponsoring a clinical trial. All trials have costs, and these are relatively easy to predict. Benefits are less clear, even when they are strictly economic. Profits will depend on the demand for the drug. This depends in turn not only on the drug being approved for marketing by regulators but also on the results of clinical trials concerning the drug's benefits—and on the perception of the medical community concerning those benefits. It also depends on the drug's side-effect profile and its cost. And it depends on the efficacies, side effects and costs of its various competitors, including competitors that emerge during the drug's development!

The next section deals with an example in which utility is determined in terms of the effective treatment of a collection of patients who have a particular disease or condition.

4. DECISION ANALYSIS AND CHOOSING SAMPLE SIZE

Choosing a sample size is a particular design issue. Consider a two-armed clinical trial. Take the utility of any particular design to be its consequent impact on patients who have the disease under study. In a decision analysis one can consider delivering good treatment to all such patients. As indicated in the Introduction, Joffe and Weeks (2002) report that "many respondents viewed the main societal purpose of clinical trials as benefiting the participants rather than as creating generalizable knowledge to advance future therapy." I disagree with Joffe and Weeks—and I disagree with the need to have the sharp distinction between clinical practice and research that is drawn in the Belmont Report (1979).

Suppose clinical trials are being planned. Let N be the size of the "patient horizon," those patients in and out of the trials who will benefit from the conclusions. [The concept of patient horizon is due to Anscombe (1963) and Colton (1963).] The numerical value of patient horizon N varies depending on the disease and the available treatments. The population of patients who have coronary artery disease and who would benefit from an advance in therapy is very large, but few patients would benefit from an advance in a rare type of children's cancer. Patients in the latter population are no less and no more important than in the former, but in the latter case the investment

associated with any knowledge obtained will have less benefit. These two extremes are addressed in the same way when choosing a clinical trial's sample size via power calculations. However, a sample size that is right for one in terms of treating as many patients with the disease in question as effectively as possible cannot be right for the other.

Suppose the sample size calculated from considerations of power turns out to be 500. In the case of small horizon N, a substantial portion of the patients—all, in the case of some children's cancers—may be in the clinical trial and so few if any patients will get to take advantage of anything learned from the trial. In the case of large N, a 500-patient trial may be too small to enable an informed choice between the two treatments and so the very large number of patients outside the trial may be treated with the inferior therapy. A conclusion of this section is that, when the goal is treating as many patients as effectively as possible, the sample sizes of the clinical trials in these two extremes should be very different.

The patient horizon N is seldom precisely known. In particular, N depends on the effectiveness and side effects of the various treatments in the trial, which are themselves unknown. However, precision in setting the value of N is not critical and only its order of magnitude need be considered. Considering the extremes, diseases or conditions that are very common (large N) call for larger trials than do rare diseases (small N).

When N is unknown, the results of this section apply reasonably well by simply replacing it with its mean. So experts could assess the size of the patient population and the potential availability of other therapies for each of the next several years. Patients presenting in the future could be discounted by the probability that they will be treated using a therapy other than one involved in the trial. This gives a mean value of N that can be used in designing the trial.

For convenience, consider dichotomous outcomes: success and failure. The goal is to treat successfully as many of the N patients as possible with one of two therapies. The utility of any trial is the number of successes over the patient horizon (including both those in the trial and those beyond). An optimal sample size n maximizes the expected number of successes over the patient horizon N. By definition, patients in the horizon are those who present after the trial and who are given the therapy that performed better in the trial.

The optimal trial sample size has order of magnitude $N^{1/2}$ (Cheng, Su and Berry, 2003). If there are two clinical trials (followed thereafter by clinical practice with the better performing therapy), then the first trial of the two should have a sample size with order of magnitude $N^{1/3}$. For example, if $N = 10^6$ and the optimal sample size is 1,000 for a single trial, then the optimal sample size for the first of two trials is 170. If, instead, $N = 1{,}000$, then the respective optimal sample sizes are 32 and 17.

In a decision analysis one can explicitly consider asymmetry in information concerning the treatment arms under consideration. The allocation proportions may then be asymmetric as well. Continue to assume a two-armed trial with the goal of effectively treating as many patients in horizon N as possible. Consider three particular forms of prior information about the unknown rates of success: beta$(1, 1)$, beta$(2, 1)$ and success rate known to equal 0.50.

Consider the specific values of N in Table 1. This table shows the optimal sample sizes for each of the arms. As indicated above, these increase with N in proportion to $N^{1/2}$ (approximately). Consider case $N = 1{,}000$ and distribution beta$(1, 1)$ (uniform on the unit interval) for both success rates, which is the leftmost case in Table 1. The table indicates that 21 patients should be assigned to one of the arms and 20 to the other. In view of symmetry, either arm 1 or

TABLE 1

Optimal sample sizes for arms 1 *and* 2 *in a two-armed clinical trial and corresponding success proportion among all N patients. "*$(1, 1)$*" and "*$(2, 1)$*" indicate beta prior distributions with those parameters; the optimal asymptotic* (*large N*) *success proportion is the prior expected maximum of the two success rates*

Patient horizon, *N*	**Arm 1 (1, 1)**	**Arm 2 (1, 1)**	**Proportion**	**Arm 1 (1, 1)**	**Arm 2 (2, 1)**	**Proportion**	**Arm 1 (1, 1)**	**Arm 2 0.50**	**Proportion**
100	6	5	0.63	4	8	0.71	9	0	0.60
1,000	21	20	0.65	16	30	0.74	29	0	0.62
10,000	70	69	0.66	56	98	0.75	99	0	0.62
Large N	$\sqrt{N/2}$	$\sqrt{N/2}$	2/3	$\sqrt{N/3}$	$\sqrt{N}$	3/4	$\sqrt{N}$	0	5/8

arm 2 could get the extra patient. (Symmetry does not extend to equality and so these two sample sizes are not equal. With 21 patients assigned to arm 1 and 20 to arm 2, adding exactly one more patient to arm 2 would never change the optimal arm for assigning to patients outside of the trial. The result for this patient could only introduce the possibility of ties in the observed success rates, in which case both arms would be optimal outside of the trial.) Using this 21 : 20 assignment, the resulting success proportion among the 1,000 patients in the horizon is 65%, which cannot be improved upon by any fixed-sample-size design. Consider the extremes, not running a clinical trial at all and running a trial entering all 1,000 patients; both have expected success proportions of 50%.

For the middle case considered in Table 1, the beta$(2, 1)$ arm is more promising than the beta$(1, 1)$ arm and so it is assigned more patients—about $\sqrt{3} - 1 = 73\%$ more for large N. For the third case considered, arm 2 is known to have a success rate of 0.5. One of the purposes of the trial—in addition to treating patients effectively—is to identify whether the unknown arm 1 success rate is greater than or less than the known arm 2 success rate. Allocating patients to the known arm 2 in the trial would be wasteful and the alternative, arm 1, has the same (unconditional) probability of success. Any known arm would be held in reserve and used after the trial should arm 1's observed success rate be less than 0.5. (Seldom is it reasonable to assume that a treatment's success rate is known since patient populations vary over time because of changing methods of diagnosis and changing attitudes about treatment, and so a treatment's success rate may vary similarly.)

This section assumed no interim monitoring. A substantial increase in the success rate is achievable if updating is possible during the trial. Such updating could be used to modify the proportions of patients allocated to the two arms and it could be used to determine when the clinical trial should end. These possibilities and other related modifications are considered in the next section. However, the next section is not explicitly decision-analytic and in particular it does not address maximizing a prespecified utility function in choosing a clinical trial design.

5. ADAPTIVE DESIGNS OF CLINICAL TRIALS

This section deals with flexible designs. Although these are not based on an explicit consideration of utilities, the goals are efficient learning and effective treatment of patients. For explicit decision-analytic generalization of some parts of this section, see Berry and Fristedt (1985).

Consider a trial having a particular design. Finding the predictive probabilities of the trial's results is always possible, even for the most complicated of designs. Similar calculations allow for finding a variety of the design's attributes, including the probability of achieving a statistically significant benefit of one therapy over another, the expected number of patients in the trial and the expected number of patients in the trial who successfully respond to their assigned treatment. Comparing attributes for different designs facilitates choosing one design over another.

The focus of this section is a family of designs that are dynamic in the sense that observations made during the trial can affect the subsequent course of the trial. The general class of designs is *adaptive* or *sequential*. Adaptation means examining the accumulating data periodically—or even continually—with the goal of modifying the trial's design. These modifications depend on what the data show about the unknown hypotheses. Among the modifications possible are stopping early, restricting eligibility criteria, expanding the accrual to additional sites, extending the accrual beyond the trial's original sample size if its conclusion is still not clear, dropping arms or doses and adding arms or doses. All these possibilities are considered in the light of the accumulating information. Adaptive designs also include unbalanced randomization where the degree of imbalance depends on the accumulating data. For example, arms that give more information about the hypothesis in question or that are performing better than other arms can be weighted more heavily (Berry and Fristedt, 1985).

Adaptation is not limited to the data accumulating in the trial. Information that is reported from other ongoing trials can also be used. This is easier to effect if one takes a Bayesian approach, possibly using hierarchical modeling (Berry and Stangl, 1996; Spiegelhalter, Myles, Jones and Abrams 2000).

Adaptive designs are used increasingly in cancer trials. This is true for trials sponsored by pharmaceutical companies and more generally. For example, a variety of trials at my home institution, The University of Texas M. D. Anderson Cancer Center (MDACC), are prospectively adaptive. I will describe some of them here.

5.1 Adaptive Dose-Finding in Phase II Clinical Trials

The standard phase II dose-finding design allocates a fixed number of patients to each dose in a grid. In retrospect the investigators usually wish they had assigned patients in some other fashion. Perhaps the dose–response curve is shifted more to the left or right than anticipated. If so, then the assignment of many patients to one end or the other was a wasted effort. Or perhaps the slope of the dose–response curve is greater than anticipated and the response of patients assigned to the flat regions of the curve would have been more informative if the assigned doses were in the region where the slope is apparently greatest. Or perhaps results for the early patients made it clear that the dose–response curve was flat and that the trial could have stopped earlier. Or perhaps the results of the trial show that the standard deviation of response is greater or less than anticipated and so the trial should have been larger or could have been smaller.

The approach of Berry et al. (2002) is to proceed sequentially, analyzing the data as it accumulates; see also Malakoff (1999) and Farr-Jones (2001). There are two stages of the trial, first a dose-ranging stage and then a confirmatory stage, if the latter is warranted. The dose-ranging stage continues until a decision is made that the drug is not sufficiently effective to pursue future development or that the optimal dose for the confirmatory stage (phase III) is sufficiently well known. (Switches to phase III can be effected seamlessly; see below.) The example trial of Berry et al. (2002) involves a neuroprotective agent for stroke. Accrual began in November 2000 and ended in November 2001. This type of trial is designed to assign each entering patient the dose (one of 16, including placebo in this example) that maximizes information about the dose–response relationship, given the results observed so far. This dose could be in the region of the greatest apparent slope, or it could be the placebo or a high dose. However, patients are not assigned doses in regions where evidence suggests that the dose–response curve is flat.

In the dose-ranging stage, neither the number of patients assigned to any particular dose nor the total number of patients assigned in this stage was fixed in advance. The dose-ranging sample size can be large when the drug has marginal benefit, when the dose–response curve is gently sloping or when the standard deviation of the responses is moderately large. It tends to be small if the drug has substantial benefit, if the drug has no benefit, if the dose–response curve rises over a narrow range of doses or if the standard deviation of the responses turns out to be small. (In addition, and somewhat nonintuitively, the dose-ranging stage is small if the standard deviation of responses turns out to be very large. The reason is that a sufficiently large standard deviation implies that a very large sample size is required to show a beneficial drug effect. The required sample size may be so large that it makes it impossible to study the drug and so the trial stops in the dose-ranging phase before substantial resources go down the drain.)

In the stroke trial considered by Berry et al. (2002) the ultimate endpoint is the improvement in the stroke scale from baseline to 13 weeks. If the accrual rate is large, then the benefit of adaptive assignment is limited by delays in obtaining endpoint information. To minimize the effects of delayed information, each patient's stroke scale is assessed weekly between baseline and week 13. Within-patient measurements are correlated, with correlations greater if they are closer together in time. We incorporate a longitudinal model into the analysis of the trial and do Bayesian predictions (using multiple imputation) of the ultimate endpoint based on the current patient-specific information, and we update the probability distributions of the treatment effect accordingly.

In comparison to a standard design, adaptive dosing is more effective in identifying the right dose, and it usually identifies the right dose with a smaller sample size than when using fixed-dose assignments. Another advantage is that many more doses can be considered in an adaptive design. (Even though some doses will be little used and some might never be used, which ones they are cannot be predicted in advance.) An adaptive design therefore has some ability to distinguish abrupt changes and other nuances in the dose–response curve.

The circumstances of the stroke trial of Berry at al. (2002) are similar to those in many other types of trials. Finding the right dose is a ubiquitous problem in pharmaceutical development, and it is done neither well nor efficiently. The adaptive nature of the stroke trial would be less advantageous if we did not exploit early endpoints. Many diseases and conditions are characterized by the availability of such early endpoints: information about how a patient is doing (local control of the disease, biomarkers, etc.) before reaching the primary endpoint. Finally, the possibility of moving seamlessly into phase III depending on the phase II results exists for many types of drugs.

The results of the stroke trial have been published (Krams et al., 2003). The dose-assignment–trial-termination algorithm was a great success—but the drug was not! The algorithm recommended stopping the trial for futility as soon as the predetermined minimum number of patients had accrued, and the data monitoring committee agreed.

5.2 Seamless Phases II and III

An unfortunate convention is categorizing drug development into phases. We go from one phase to the next when we think we know something: the maximum tolerated dose from phase I or the "right dose" from phase II to be used in phase III. In the Bayesian perspective one never regards a quantity to be perfectly known. Instead, the Bayesian design carries uncertainty along with whatever knowledge is available. Phases of drug development are arbitrary labels that describe a process that is—or should be—continuous.

One of the consequences of partitioning drug development into arbitrary phases is that there are delays between them. For example, there is a pause between phases II and III to set up one or more pivotal studies. As mentioned above, the design of the stroke trial allows for avoiding such a pause. At each time point, say weekly, the algorithm that guides the conduct of the trial does a decision analysis. It provides a recommendation to (i) continue the dose-ranging stage of the trial, (ii) stop the trial for lack of evidence of efficacy (inadequate slope of the dose–response curve) or (iii) shift into a confirmatory stage. The shift in (iii) can be made seamlessly, with no break in the accrual. Indeed, it is theoretically possible to effect such a shift without informing the investigators: they would continue to randomize doses, but, unbeknownst to them, the only doses being assigned would be the phase III dose and the placebo. (Although this seamless switch was an option of the algorithm in the stroke trial, it was not made available in the actual trial.)

At MDACC we designed a trial (Inoue, Thall and Berry, 2002) that encompasses both phases II and III. If there is a switch to phase III, it is seamless. The trial compares a single dose of a drug with placebo, both on top of standard chemotherapy. (It could incorporate other doses in a manner similar to that of the stroke trial, but it is easier to make progress in the scientific community by effecting one innovation at a time.) The anticipated effect of the drug is via local control. We model survival as it depends on local control and as it depends on treatment. Though remote, we allow for the possibility that the drug has a beneficial effect on survival that is not mitigated by local control. So local control is an auxiliary endpoint in a way similar to the role of the early stroke score in the stroke trial. (See Section 6 for more discussion about auxiliary endpoints.) However, the clear focus is on survival as the primary endpoint, and the utility of the auxiliary endpoint must be demonstrated by the results actually observed in the trial. We exploit any relationships that exist, but do not assume such relationships. We analyze the data in the trial frequently and adapt to the accruing evidence.

The seamless aspect is as follows. Initially, only MDACC patients are accrued to the trial. Think of this as phase II. If the accumulating data are sufficiently strong in suggesting that the drug has no effect on local control or survival, then the trial stops. If the data suggest that the drug may have an impact on local control and that this impact translates into a survival benefit, then the trial will be expanded to include other centers and the accrual rate will increase accordingly. During such an expansion, patients continue to accrue at MDACC so that there is no downtime in the local accrual while other centers gear up for joining the trial. This is efficient use of patient resources because the responses of patients accrued early at MDACC contribute to the eventual inferences about survival. These patient responses are the most informative of those enrolled in the two phases because their follow-up times are the longest.

The trial continues until (i) stopping occurs for futility, (ii) the maximum sample size of 900 is reached or (iii) the Bayesian predictive probability of eventually achieving statistical significance becomes sufficiently large. Should (iii) occur, the accrual ceases and the drug company submits an application for marketing approval to regulatory agencies.

The sample size of a conventional phase III trial with the desired operating characteristics is 900. We take this to be the maximum sample size in the seamless design. The actual accrual is very likely to be much less than this maximum sample size and on average it will be about half as large. On the other hand, incorporating the same number of interim analyses in a conventional design using an O'Brien–Fleming stopping boundary allows for only a slight decrease in the average sample size. Under any hypothesis, null or alternative, the Bayesian design occasionally leads to a relatively large trial (close to 900 patients). However, a pleasant aspect of a Bayesian design is that the sample size is large precisely when a large trial is necessary. Conventional trials may well (and sometimes do!) come to their

predetermined end with an ambiguous conclusion. In a Bayesian approach one may choose to continue such a trial to resolve the ambiguity, and this option has substantial utility. (Carrying this argument to the maximum sample size, there may be times for which stopping at 900 is ill advised, but for logistical reasons we felt the need to specify a maximum size.)

Reductions in sample size result from two characteristics of the seamless design described above. First are the frequent analyses to assess the predictive probability of eventual statistical significance. The second is the explicit modeling of the possible relationship between local control and survival. Of the two, the second is more important.

A conventional drug development strategy involves running a phase II trial that addresses local control, digesting the results and, if the results are positive, starting to develop phase III trials with survival as the primary endpoint. As indicated above, in comparison with such a strategy, a seamless approach can greatly reduce the sample size. In addition, a seamless design minimizes the pauses between phases and so the total drug development time is greatly shortened.

5.3 Adaptive Allocation

The adaptive designs discussed so far are motivated by the desire to learn as efficiently and as rapidly as possible. Another kind of adaptive design aims to treat patients as effectively as possible. These designs use adaptive allocation in which patients are more likely to be assigned to treatments that are performing better. In addition to making clinical trials more attractive to patients and thereby increasing participation in clinical trials, such strategies have the interesting side effect of efficient and rapid learning!

As of this writing more than a dozen trials at MDACC have been designed and are being conducted using adaptive allocation. Our standard approach is to randomize treatment assignment, but to shift the weights toward better performing arms as the trial proceeds and the results accumulate. Many of these trials have more than two arms. The arms are sometimes distinct therapies, and sometimes they are closely related. An example of the latter is an MDACC trial involving five doses (including 0) of a drug. For reasons I will not go into, the dose–response curve is probably not monotone. In particular, efficacy may increase for small doses and then decrease. Initially we assign doses in a graduated fashion, climbing the dose ladder slowly. But as doses become "admissible," we assign patients to those that have been performing well.

Consider a patient who qualifies for the trial. To decide which dose to assign we calculate the current (Bayesian) probabilities that each admissible dose is better than the placebo. This calculation uses all available information. We allocate doses randomly, with weights proportional to these probabilities. [The first version of this idea is 70 years old, dating to Thompson (1933). An historical footnote is that Thompson's 10-page paper focused almost exclusively on the computational problem of evaluating the probability that $\theta_1 > \theta_2$, where these parameters have independent beta distributions. The era of computers renders such a calculation trivial, but in Thompson's time it was the great hurdle.] We consider other allocation algorithms, including assigning in proportion to the powers of these probabilities. The assignments we consider involve some amount of randomization, but patients are more likely to receive doses that are performing better. Doses that are doing sufficiently poorly become inadmissible in the sense that their assignment weight becomes 0. When and if we learn that the drug is effective, we stop the trial. When and if we learn that the drug is ineffective, then again we stop the trial. Patients in the trial benefit from data collected *in the trial*.

Our explicit goal is to treat patients more effectively, but a happy side effect is that we learn efficiently. We evaluate a design's frequentist operating characteristics and modify the sample size if necessary.

Thompson-like designs are similar to play-the-winner designs in that their goal is to treat patients in the trial more effectively. They are also similar in that both designs are ad hoc and neither is optimal in any sense that I know. Various optimality criteria have been suggested for evaluating adaptive trials. The most common is maximizing an expected sum of observations over the course of the trial, such as the total number of successes, possibly with later observations discounted relative to the next observation (Berry, 1972, 1978; Berry and Fristedt, 1985). Decision problems of this type are called bandits. For example, two-armed bandits involve two treatments. Finding optimal bandit strategies means solving dynamic programming problems. The solutions can require intensive computation, but finding on-line solutions during a trial is not out of the question, even for some rather complex settings.

A characteristic of optimal strategies for commonly considered objectives in bandit problems is that they are deterministic. One of the vagaries of clinical trials is that the prognosis of patients may fluctuate over the trial's course. This would not be a serious problem if prognoses were well understood and easy to measure.

Usually, neither is true. Or, more precisely, we seldom know whether they are true. A treatment may do relatively well during a period when it happened to be assigned to patients who had relatively good prognoses. When following a deterministic strategy, such a treatment may be used for some number of patients in a row and its performance may therefore be artificially inflated. An easy fix is to mix a randomized assignment with an optimal assignment: assign the optimal arm with probability r and an arm chosen randomly with probability $1 - r$. At MDACC we have not yet progressed to this point. However, we recognize the importance of having some level of randomization, which is why we use Thompson-like strategies.

5.4 Process or Trial? Evaluating Many Drugs Simultaneously Using Adaptive Allocation

The greatest room for innovation and for improving drug development is effectively dealing with the enormous numbers of molecules that are available as potential drugs. The notion of screening drugs one at a time is ingrained in the pharmaceutical culture. This notion is inefficient in the extreme. Only those companies that are able to process many drugs simultaneously and efficiently will survive.

Many different drugs should be evaluated in the same preclinical experiments. During the evaluation process, there should be continuous updating (via Bayes' theorem) with some candidates dropped and others added. The degree of focus on any particular drug will depend on the available data. Drugs that are apparently more promising will move faster through the preclinical setting. Drugs that give disappointing data will languish.

These ideas apply as well to clinical trials. As an example, at MDACC we are building the foundation for a phase II trial for evaluating drugs that is more a process than a trial. The idea is a straightforward extension of the adaptive assignment strategies described above. We start with a number of treatment arms plus a control—possibly a standard therapy. We randomize patients to the arms and learn on-line about their relative efficacy. Arms that perform better get used more often. An arm that performs sufficiently poorly gets dropped. When an arm does well enough it graduates into phase III; if it does sufficiently well, it might even replace the control. As more arms become available, we add them to the mix.

The result is that better arms move through quickly and poorer arms get dropped. An advantage to patients in the trial is that they receive better treatment (provided the arms are not the same). The advantage to patients outside the trial is that they get access to better drugs more rapidly.

5.5 Extraim Analyses

Many clinical trials end without a clear conclusion. For example, a statistical significance level of 5% in the primary endpoint may be required for drug registration and the p-value may turn out to be 6%. The regulatory agency suggests that the trial was "underpowered" and that the company should carry out another trial. It would be much more efficient to simply increase the sample size in the present trial with the goal of resolving the issue. But the possibility of extending the accrual increases the type I error rate. The principle is identical to that for interim analyses.

A solution is to build into the design the possibility of continuing the trial depending on the results, suitably adjusting the significance levels. In contrast to adjustments for interim analyses, I adjust for extraim analyses in the opposite direction, with much of the overall significance level "spent" at the originally planned sample size. For example, taking equal significance levels at each possible termination point is preferable to O'Brien–Fleming stopping boundaries because the latter are overly conservative. Allowing for extending the trial increases the maximal sample size and also the average sample size. However, a modest increase in average sample size (such as 10%) comes with a substantial increase in statistical power (such as 80% increasing to 95%). The reason for such a beneficial trade-off is that the trial is extended only when an extension will help resolve marginal results.

The "penalty" in significance level can be either partially or fully offset by including futility analyses as part of the design. Namely, the trial would be stopped for sufficiently negative results at preset interim time points. The reason such analyses offset the penalty for extraim analyses is that the null hypothesis is never rejected when the trial stops for futility. Decreasing the opportunity for a type I error also decreases the power of the trial. However, this decrease is usually quite modest and in any case is more than compensated for by the increase in power due to the extraim analyses.

The increment in sample size depends on the available data at the time the decision is made to continue the accrual. It also depends on the number of possible extensions. I base each extension on the *predictive power*. The usual definition of power assumes a particular value of the parameter of interest, say θ. Predictive power considers all possible values of θ. The

data available at the time of the extraim analysis play two roles. First, they count in the final results of the trial. Second, they are used to update the (Bayesian) probability distribution of θ. We fix the total sample size n and calculate the power for detecting each possible value of θ. We average this power with respect to the probability distribution of θ to give the predictive power for sample size n, and extend the accrual to give a total sample size having a prespecified predictive power. If there is no such value of n, then continuing the accrual may be unwise.

There is an aspect of the above development that may seem unrealistic. The problem is that endpoints for those patients treated in the trial may not be available at the time of the extraim analysis. Even if the endpoint is tumor response, there is a delay in obtaining this information. However, endpoint results for patients in the trial can be predicted along with that of patients not yet accrued. If there is some early information (biomarkers, performance status, etc.) that is correlated with the endpoint of interest, then this can be used to inform the prediction. A special and important case is when the endpoint is time to an event. The fact that a patient has not yet reached the event is useful information that is accommodated in Bayesian updating. But if there is no patient-specific early information, then patients treated but not yet assessed for response are handled in the same way as patients not yet treated. (This set of issues is sufficiently important that they deserve being addressed separately—see Section 6.)

The above process is complicated, but it can be completely and precisely described. That means the process can be simulated. The simulations can be carried out under various assumptions about the parameter of interest. In particular, the false-positive rate can be calculated. If there is a target significance level (such as 5%), then the various inputs into the design (number and type of extraim analyses, number and type of futility analyses, etc.) can be varied until achieving that target. An advantage of simulations is that each iteration provides a fully accrued trial. So it is possible to check any characteristic of interest regarding the trial's design by calculating the proportion of the trials that have the characteristic. Among the characteristics of interest are power, actual sample size and the probability of extending the accrual.

6. AUXILIARY ENDPOINTS

The adaptive designs considered in Section 5 are based on information that accrues during the trial on the primary endpoint. If the primary endpoint is delayed and the accrual is rapid, then adaptive methods are of limited value. The present section addresses statistical procedures for designs that exploit accumulating information on other than the primary endpoint.

Suppose that the endpoint is time to disease progression and a patient has not yet progressed. The absence of an event is information that can be used for Bayesian updating of the distributions of the parameters involved.

Information accrues as well about each patient's condition. Whether the patient's tumor has responded is information, and this is so even if tumor response is not the primary endpoint. The point is that tumor response *may* be related to the primary endpoint. A patient's performance status can change over time (or not!) and that information too is important to assess and incorporate into the analysis. There are many such variables that are candidates for consideration. I call them *auxiliary endpoints* because they may contain information about the primary endpoint even though they are not themselves primary.

It is important to take advantage of the wealth of information that accrues in a trial. The approach is to model the possibility of a relationship between the accumulating information and the primary endpoint.

There are several benefits of modeling auxiliary information. One benefit was considered in Sections 5.1 and 5.2. Waiting for long-term endpoints limits the ability to modify the design of a clinical trial during its course. Using auxiliary endpoints makes adaptation possible. Another benefit of modeling is that the relationship between the primary endpoint and auxiliary endpoints may allow for announcing trial results earlier or for getting earlier regulatory approval of an experimental drug. For example, suppose that survival is the primary endpoint and that modeling its relationship with an auxiliary endpoint was considered explicitly in the design of the trial. Accrual to the trial has ended and all the patients have been treated. There is insufficient information to conclude the drug's benefit on the basis of survival alone, in part because many patients' outcomes are censored. However, the drug has a positive impact on the auxiliary endpoint, and it turns out that in both drug and control groups there is a clear relationship between the auxiliary endpoint and survival. A model that utilizes auxiliary information may conclude a survival benefit.

An auxiliary endpoint may or may not be a "surrogate endpoint." This distinction is critically important.

An auxiliary endpoint is not a surrogate, or more accurately, whether it is a surrogate is not relevant for the design and analysis. The focus of the definitive analysis is the primary endpoint and not the auxiliary endpoint. The conclusion of the trial is whether the drug improves survival.

A model incorporating early information can be arbitrarily complicated. In particular, it can contain many variables. However, especially as regards registration trials, one should tiptoe into model development by considering one auxiliary endpoint at a time. Especially important will be to explicitly consider the possibility that any relationship between auxiliary and primary endpoints depends on the treatment. Should it happen that there is an interaction between the auxiliary endpoint and the treatment—such as that tumor response is related to survival in the control group but not in the treatment group—then an appropriate model automatically discounts the auxiliary endpoint in the treatment group.

Modeling loses little, and much can be gained, as indicated in the seamless phase II–III trial design presented in Section 5.2. Again, the gains and any losses can be assessed by simulation.

A special type of auxiliary endpoint is a biomarker. Models relating to primary endpoints can be based on longitudinal models that incorporate biomarker information that accrues over time. An example longitudinal model is described in Berry et al. (2002), which is set in the context of a stroke trial.

7. DISCUSSION AND CONCLUSION

The Bayesian philosophy is rather different from the standard frequentist approach to clinical trial design and analysis. It is ideally suited to clinical research. In particular, it is more flexible. And it is consistent with the scientific principle of paying heed to the accumulating data.

A possible use of the Bayesian approach is to expand the range of clinical trial designs considered in the standard approach, but then to evaluate the frequentist operating characteristics (i.e., those that are functionally related to the parameter values). Is this approach Bayesian or frequentist? If the design is changed to have "good" frequentist characteristics, then it is frequentist. Otherwise it is Bayesian. But the distinction is obviously blurred—suppose someone other than the designer (such as a regulatory agency) checks the frequentist characteristics and finds that they are adequate. It may be Bayesian to one person and frequentist to another.

A fully Bayesian approach is decision-analytic. That is, one considers the consequences of an experiment and evaluates their worth (or utilities). The usual approach to designing clinical trials does not formally address such consequences. Clinical researchers may take into consideration the prevalence of the disease, for example, but it is not formally considered in the design. This has unfortunate consequences.

Consider a rare cancer, with two possible therapies. Researchers would like to conduct a clinical trial. They contact their local statistician, who calculates that the sample size necessary (based on power considerations) requires that all patients who have this disease in the next 10 years participate in the trial. So no trial gets done. An alternative is a decision-analytic approach with the goal of treating as many patients with the disease as effectively as possible. Such a trial can always be conducted, and its sample size will be modest—as indicated in Section 4.

This example serves as a prototype for this article. The ethics of clinical research are problematic. The Belmont Report (1979) resolves the conflict by separating clinical research from clinical practice. Such a separation is artificial—and unnecessary. A Bayesian decision-analytic approach resolves the issue by facing it head-on. The goal is to treat patients as effectively as possible, whether they are in the trial or will present later.

Much of the focus of this article is on adaptive designs. There are obvious benefits to be derived from updating one's state of knowledge as relevant evidence accumulates, and using this information to guide the course of the trial. So why are adaptive designs not more common? Part of the reason is the dominance of the frequentist approach in medical research. The Bayesian view has made occasional inroads into attitudes among medical researchers, but only recently has its influence been felt in pharmaceutical development. Further changes will not be drastic or immediate. However, in the next few years we will see Bayesian approaches used increasingly. At least for the near future they will be used as tools, with justifications following a more or less traditional frequentist course. As time passes and as researchers and regulatory folk become more accustomed to Bayesian ideas, they will be increasingly accepted on their own terms.

REFERENCES

ANSCOMBE, F. J. (1963). Sequential medical trials (with discussion). *J. Amer. Statist. Assoc.* **58** 365–387.

BELMONT REPORT (1979). Ethical principles and guidelines for the protection of human subjects of research. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Available at http://ohrp.osophs.dhhs.gov/humansubjects/guidance/belmont.htm.

BERRY, D. A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.

BERRY, D. A. (1978). Modified two-armed bandit strategies for certain clinical trials. *J. Amer. Statist. Assoc.* **73** 339–345.

BERRY, D. A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine* **12** 1377–1404.

BERRY, D. A. (1995). Decision analysis and Bayesian methods in clinical trials. In *Recent Advances in Clinical Trial Design and Analysis* (P. Thall, ed.) 125–154. Kluwer, Boston.

BERRY, D. A. (1996). *Statistics*: *A Bayesian Perspective*. Duxbury, Belmont, CA.

BERRY, D. A. and EICK, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine* **14** 231–246.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems*: *Sequential Allocation of Experiments*. Chapman and Hall, London.

BERRY, D. A., MÜLLER, P., GRIEVE, A. P., SMITH, M., PARKE, T., BLAZEK, R., MITCHARD, N. and KRAMS, M. (2002). Adaptive Bayesian designs for dose-ranging drug trials. In *Case Studies in Bayesian Statistics V. Lecture Notes in Statist.* **162** 99–181. Springer, New York.

BERRY, D. A. and STANGL, D. K., eds. (1996). *Bayesian Biostatistics*. Dekker, New York.

CHENG, Y., SU, F. and BERRY, D. A. (2003). Choosing sample size for a clinical trial using decision analysis. *Biometrika* **90** 923–936.

CLEMEN, R. T. (1991). *Making Hard Decisions*. PWS-Kent, Boston.

COLTON, T. (1963). A model for selecting one of two medical treatments. *J. Amer. Statist. Assoc.* **58** 388–400.

ELLENBERG, S. S., FLEMING, T. R. and DEMETS, D. L. (2002). *Data Monitoring Committees in Clinical Trials*: *A Practical Perspective*. Wiley, New York.

FARR-JONES, S. (2001). Better statistics. *BioCentury, The Bernstein Report on BioBusiness* **9** (No. 12, March 12.)

INOUE, L. Y. T., THALL, P. F. and BERRY, D. A. (2002). Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* **58** 823–831.

JOFFE, S. and WEEKS, J. C. (2002). Views of American oncologists about the purposes of clinical trials. *J. National Cancer Institute* **94** 1847–1853.

KRAMS, M., LEES, K. R., HACKE, W., GRIEVE, A. P., ORGOGOZO, J.-M. and FORD, G. A. (2003). Acute stroke therapy by inhibition of neutrophils (ASTIN): An adaptive dose–response study of UK-279,276 in acute ischemic stroke. *Stroke* **34** 2543–2548.

LEWIS, R. J. and BERRY, D. A. (1998). Decision theory. In *Encyclopedia of Biostatistics*. Wiley, New York.

LEWIS, R. J., BERRY, D. A., CRYER, H., FOST, N., KROME, R., WASHINGTON, G. R., HOUGHTON, J., BLUE, J., BECHHOFER, R., COOK, T. and FISHER, M. (2001). Monitoring a clinical trial conducted under the new FDA regulations allowing a waiver of prospective informed consent: The DCLHb traumatic hemorrhagic shock efficacy trial. *Annals of Emergency Medicine* **38** 397–404.

MALAKOFF, D. (1999). Bayes offers a "new" way to make sense of numbers. *Science* **286** 1460–1464.

SOX, H. C., BLATT, M. A., HIGGINS, M. C. and MARTON, K. I. (1988). *Medical Decision Making*. Butterworth and Heinemann, Boston.

SPIEGELHALTER, D. J., FREEDMAN, L. S and PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials (with discussion). *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

SPIEGELHALTER, D. J., MYLES, J. P., JONES, D. R. and ABRAMS, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment* **4** 1–130.

THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.