14 Bayesian Designs in Clinical Trials

Gary L. Rosner
B. Nebiyou Bekele

In this chapter, we discuss issues that arise when developing, writing up, and implementing clinical study designs that incorporate Bayesian models and calculations. We have had the opportunity to work with many such designs at The University of Texas M. D. Anderson Cancer Center. We feel that study designs that incorporate Bayesian models offer many advantages over traditional frequentist designs, and we will discuss these advantages in this chapter. At the same time, Bayesian models require a lot of thought and close work with the clinical investigators. Also, there may be some reluctance on the part of some clinical investigators to accept a study design that is built on Bayesian considerations. We will provide some arguments and real examples that may help statisticians overcome such reluctance. Although our examples tend to come from the field of oncology, the lessons and underlying ideas have broad application. (See Carlin and Louis [1] for a general introduction to Bayesian methods.)

WHY BAYESIAN DESIGNS

What Are Bayesian Designs? Types of Bayesian Designs

First we need to define what we mean by a Bayesian design. In the first paragraph, we specifically avoided writing the term "Bayesian design," choosing instead the phrase "clinical study designs that incorporate Bayesian models and calculations." The latter phrase allows us to include many designs that are not fully Bayesian, meaning that they do not choose the design to minimize some risk. Instead, many of these "calibrated Bayes" (2) designs incorporate a Bayesian model, possibly considering prior information, in the stopping rules of the study.

An example of this calibration is the following. The statistician and clinical investigators decide on the general form of the criteria for decisions at interim analyses, such as basing decisions on the posterior probability that the treatment's success probability exceeds a threshold value. Next, the statistician will typically carry out a large number of simulations under various scenarios. The statistician reviews the simulation results with the clinical investigators, allowing them to decide on the criteria that yield the best (to their minds) operating characteristics. This process may include changing the benchmark value against which one compares the posterior treatment-related success probability or the degree of certainty (e.g., 80% or 90%) that one will require before one will consider stopping the study.

There also exist more formal Bayesian designs for clinical trials. Berry argues for the application of decision theory in clinical trial design (3, 4). Even if one takes a fully Bayesian view, one will still find that reviewing these *a priori* simulations serves to make the

transition from frequentist designs to Bayesian ones easier for clinical investigators. Simulations under various scenarios also help reveal sensitivity of the study's decisions and inferences to prior assumptions. We discuss these ideas later in this chapter through our examples.

Advantages of Bayesian Designs

Why is one interested in Bayesian designs for clinical trials? One can view a clinical trial as an experiment that will lead to a decision (use the new treatment or do not use the new treatment) or prediction (the new treatment regimen will provide a benefit of so much over the standard treatment). Bayesian methods are ideal for decision making (i.e., minimizing risk or maximizing utility) and for prediction.

Additionally, Bayesian methods are ideal for combining disparate sources of information. Thus, one can construct a coherent probability model to combine the information from the current study with historical data and with any information available from ongoing studies using Bayesian considerations. Perhaps a further impetus to the current interest in Bayesian designs is the fact that the Bayesian inference obeys the likelihood principle. Many of the clinical studies we see include interim analyses, and when there is no provision for interim analyses, we often suggest them in our reviews. The likelihood principle is important as it relates to interim analyses of an ongoing study. One develops frequentist stopping rules, such as group sequential designs (5), in a way that preserves the overall type I error under the null hypothesis. Thus, a treatment effect that might have been statistically significant without any prior interim analyses may not be significant after accounting for the number of prior analyses. The likelihood principle, however, requires that data that lead to the same likelihood for the parameter of interest should lead to the same inference (6). A consequence of the likelihood principle is that the number of interim analyses does not affect Bayesian inference, since the likelihood is the same whether the current analysis had been the first or the most recent of several earlier analyses. All that matters to the Bayesian are the data at hand and not what happened before, unless earlier analyses somehow alter the likelihood.

Another reason more and more clinical trials are incorporating Bayesian ideas is the desire in many situations to include adaptive randomization. Such clinical trials change the randomization probabilities in light of the accruing data. The study may start randomizing patients to the different treatments with equal probability. Then, perhaps after enrolling some minimum number of patients, the randomization probabilities adapt to favor the better performing treatments. Bayesian methodology may enter the study by way of using posterior probability calculations to influence the randomization probabilities (7). The ethical idea is to reduce the number of patients who receive inferior treatment while still accruing convincing evidence within the clinical trial. (There have been interesting discussions of the ethics of randomization and adaptive randomization (8–12), but we do not discuss this aspect of clinical trial design here.) We discuss an example of Bayesian adaptive randomization later in this chapter.

REQUIREMENTS FOR A SUCCESSFUL BAYESIAN DESIGN

As with all clinical studies, considerable work has to go into the preparation of the study design. The statistician and the clinical investigators need to discuss the study's aims and objectives. Care must go into selecting endpoints for the primary and secondary aims of the study. Much of these considerations are discussed elsewhere in this volume, so we will focus more on the aspects that relate to the Bayesian part of the design. In particular, we will talk about the prior distribution and stopping rules. Additionally, if one takes a decision-theoretic approach, one will have to consider the utility function that accounts for the study's aims. If one wishes to calibrate the design, then one will have to review with the other investigators the implication of various decision-rule parameters on the operating characteristics.

Software for Real-Time Updating

Real-time updating is an important aspect of modern Bayesian trial designs. These designs incorporate early stopping rules, allowing the investigator to stop early for lack of efficacy, superiority, or excessive toxicity. For example, in a single-arm phase II study that will compare the progression free survival (PFS) associated with a new treatment to historical information for one or several standard treatments, an investigator may desire to stop the study early if there is evidence that the new treatment results in worse outcomes than the historical standard. A Bayesian approach to this problem might assume that PFS follows an exponential distribution (with rate parameter θ) and, with a conjugate gamma prior, that θ follows, *a posteriori,* a gamma distribution. A common stopping rule under this setup is to stop the trial if at any point $\Pr(\theta > \theta^* | \text{ Data}) > C$ (where θ^* usually represents some historical event rate and C is some pre-specified threshold value). One computes this probability each time a new patient (or group

of patients) enters the study or when a patient already enrolled experiences disease progression.

Typically, calculation of the above probability requires numerical integration, and one must develop statistical software to carry out the calculations necessary to monitor the accruing data and determine whether the interim stopping boundaries have been crossed. By *software,* we include R scripts or SAS macros written solely for use by the collaborating statistician, stand-alone desktop computer programs written for use by other statisticians, or even Web-based applications for use by nonstatistical research staff. The kind of application one develops is a function of who the end user will be and how often future studies may use the same sort of design. For example, it may be sufficient for the collaborating statistician to have a function that runs within a general purpose statistical or mathematical package when carrying out a single-institution study that will evaluate outcomes for a rare disease with slow accrual in which posterior updating will be necessary only every 4 to 6 weeks. In contrast, a rapidly accruing multicenter, multiarm study may require real-time updating via a Web-based application or a telephone voice-response system. Part of the work involved in implementing Bayesian methods is to determine the exact software needs of the particular study's design. Below we describe a set of commonly used trials that require real-time updating.

Types of Studies

Phase I Oncology Dose-finding Study

Many drugs used in oncology are associated with severe toxicities and have a narrow therapeutic window, meaning that there is only a small range of doses that may be efficacious without being overly toxic. Therefore, the initial step in assessing these compounds in humans usually focuses on finding a dose that has an acceptable level of toxicity. Because one of the most important constraints on the conduct of these initial trials is the desire to limit the number of patients who experience severe toxicity, these studies are conducted with dose escalation proceeding in a sequential manner. That is, the study enrolls small cohorts of patients (e.g., three to six) and does not assign a higher dose until each patient in a given cohort has been through at least one cycle of treatment and their outcomes assessed. The toxicity outcomes observed from these (and earlier) patients may enter into an algorithm that the investigators use to select the dose for the next cohort. The purpose of this sequential approach is to decrease the chance that large numbers of patients receive doses that are too toxic.

The assumption underlying this approach in oncology, at least, is that toxicity and response are correlated *through dose.* That is, higher doses lead to an increase in the toxicity risk and an increase in the probability that a patient will respond to treatment. This assumption was historically reasonable in oncology, where one defined activity in terms of killing cancer cells. Thus, phase I oncology studies have traditionally attempted to determine the highest dose that has an acceptable toxicity level, since by assumption this dose will also lead to greater efficacy than lower doses.

Bayesian phase I designs treat a patient's risk of toxicity at a given dose as a quantity about which the investigator has some degree of uncertainty. One quantifies this uncertainty via a probability distribution. Decisions to escalate the dose, continue with the current dose, or de-escalate from the current dose incorporate the most current data. Given what one has learned to date, one will treat the next patient with the dose with an expected risk of toxicity that is closest to a predefined target toxicity risk. In such a setting, Bayesian methods offer clear advantages. The Bayesian framework provides a means by which one can learn about toxicity risks at the different doses and naturally make decisions based on the data observed in a sequential manner. The increase in knowledge is reflected by a decrease in uncertainty as one moves from prior to posterior.

Phase II Adaptive Randomization Trials

Bayesian adaptive randomization designs successively (as patients are evaluated for outcome) modify the randomization probabilities based on either posterior or predictive probabilities favoring one treatment over another. In essence, data from patients previously enrolled and evaluated in a study are used so that patients currently enrolling onto the trial will have a higher probability of being randomized to the most efficacious treatments. In these types of designs, subjects are initially randomized fairly (i.e., with equal probability) to the various (at least two) treatment arms. Since many adaptive randomization trials usually have a period in which patients are equally randomized prior to the implementation of adaptive randomization, it is important that the statistician monitor the actual randomization versus expected randomization.

Other Trials

Other interesting and useful examples of successful Bayesian applications in the design of clinical trials include single-treatment phase II studies that consider efficacy and toxicity, with stopping rules based on both

end points (13–16). Another interesting innovation is the so-called seamless phase II/III design (17, 18). With this design, randomization begins within the context of a small phase II study that collects survival information but has an intermediate endpoint as the primary outcome. Based on early results with respect to the intermediate end point, however, the study may expand to a large randomized phase III study with survival as the primary outcome. Berry et al. discuss a design that simultaneously sought the best dose of a drug in an adaptive way and maintained a randomized comparison with placebo (19). Other examples exist in the literature (20).

As mentioned earlier, the Bayesian inferential machinery fits well with decision theory. Once one has determined an appropriate utility function, one can set up the design to optimize the utility. Furthermore, one can carry out sequential decision making, either fully via backward induction (21) or by looking ahead one or a few steps. In all cases, one maximizes the utility, taking into account posterior uncertainty. There also exist more formal Bayesian designs for clinical trials. Berry argues for the application of decision theory in clinical trial design (3, 4). Kadane (22) presents an interesting example of a clinical trial, describing the background and development of the study. The literature includes other examples of formal Bayesian designs (23–26). Rossell et al. (27) and Ding et al. (28) present decision-theoretic designs for phase II studies that screen out active therapies from among a sequence of new treatments.

Realistic Priors

Historical Priors

Often, there have been earlier studies with one or more of the agents under investigation in the current study. These data usually inform the study's design, either informally (as in determining the null and alternative hypotheses in frequentist designs) or formally via a prior distribution. One may find, however, that if one assumes that the current study's patients will be exchangeable with the historical information, the historical information will be extremely informative with respect to inference during the current study. In fact, in some cases, it may well be that there is little reason to embark on the current study, given the evidence in the historical information. (In many situations, it may well be appropriate to consider whether there really is a need for the current study, given the strength of historical evidence. That is a topic for another discussion, however.) Since the current study will go forward, one has to find a way to discount the historical information or choose not to assume that the patients in the current study are exchangeable with the earlier studies.

If we consider a binary outcome, such as treatment success or failure (however defined), then we might characterize the historical data by means of a beta distribution. For example, if an early study enrolled 50 patients, and 30 patients experienced a treatment success, we might characterize the uncertainty about the treatment's underlying success probability by a beta distribution with parameters equal to 30 and 20. One might think of this prior as the posterior distribution arising from an experiment that gave rise to these data and a fully noninformative beta[0,0] prior. (Alternatively, one could consider an initial uniform[0,1] prior or a Jeffreys beta[0.5, 0.5] prior and determine a posterior beta distribution with slightly different parameters[1].)

Now, one might feel that the beta[30,20] prior is too informative for this study. For example, this distribution has 95% of the central mass between 0.46 and 0.73. If one wants to entertain the possibility of smaller success probabilities than 0.4, then one may want to discount this prior data in some way. A natural way to keep the prior mean 0.6 but increase the uncertainty is to decrease the prior sample size. For example, one might choose to reduce prior information to the equivalent of a prior sample size of 5 by way of a beta[3, 2] distribution. Now the central 95% of the mass lies between 0.19 and 0.93.

A related approach for discounting the historical information is with a power prior (29, 30). The power prior extends the notion of discounting to a general class and allows for inference with respect to the degree of discounting. Briefly, one considers a parameter in the probability model that will characterize the level of discounting for the historical information. The basic idea of the power prior is that the more similar the prior and current data are, the less discounting that takes places and vice versa. Let $L(\theta \mid D)$ represent the likelihood function that will characterize the data at the end of the current study (i.e., after collecting the data represented by D). Using the same likelihood function with the historical data D_H, the power prior is $p(\theta \mid D_H, \delta) \propto L(\theta \mid D_H)^{\delta} p(\theta \mid \phi)$, where the parameter ϕ is a hyperparameter for an optional initial prior. The parameter δ will serve to discount or down-weight the information content of the historical data when one will apply this prior to carry out posterior inference in the analysis of the current study.

Another way people have discounted prior information is less direct: they have modified parameters in the stopping rules to make it more difficult to stop early. In other words, one uses the historical information to generate an informative prior but makes the

cutoff for early stopping more stringent than perhaps one would normally consider reasonable. For example, if one is basing the stopping rule on a criterion based on the posterior probability that some parameter or function of model parameters exceeds a threshold, one may require a very high probability (e.g., 99%) of this event before considering early stopping. Making the stopping rule more stringent basically provides a way to keep the prior from dominating early decision making and allows the current study to continue accumulating data.

The process of determining the boundary criteria often proceeds iteratively. One determines the criteria for early stopping by carrying out simulations under various scenarios and then deciding which stopping rules lead to satisfactory operating characteristics. Although such devices tend to make the designs acceptable to frequentists, because of the calibrated operating characteristics, they also may tend to undermine the benefit of the underlying Bayesian model. The historical information may become almost neglected or, at most, these data enter into the design as a formality without giving full consideration of their importance to the inferential question under investigation.

Elicitation of Experts

Elicitation of priors from experts would seem a reasonable approach, especially in the absence of historical data. Carlin et al. (31) describe their experience eliciting prior information for a clinical trial. Problems may occur in a clinical trial for which the *experts* may have provided a prior that subsequently appears to be at odds with the data. An informative example is discussed by Carlin et al. in the context of a randomized clinical trial evaluating the benefit of prophylaxis against possible infection with toxoplasmic encephalitis (TE) (32). In this study, the five experts whose opinions went into the prior distribution turned out to have been overly optimistic. Each expert anticipated a treatment benefit. Although there was widespread disagreement among these five individuals, none considered the possibility that the treatment would be no better than placebo, let alone worse.

The key points resulting from these investigators' experience with this study are instructive. In particular, the experts may provide point estimates, but there is underlying uncertainty in each expert's opinion. Perhaps a mixture of these separate prior distributions will be more robust to the analysis than combining the experts' point estimates into a single prior. Another point brought out in this study was that different experts might find it easier to specify priors for the effect of the treatment on different end points. For example, one expert was not able to provide a prior estimate of the effect of the treatment on the risk of death or TE, whereas the other four could and did.

In our experience, it is also important that those whose opinions one seeks see the consequences of their *a priori* estimates. Graphical displays of uncertainty distributions or of observable quantities, given prior specification, allow the experts to gain insight into the implications of their stated beliefs (16, 33). Quite often, this feedback reveals inconsistencies and leads to revisions.

Thus, one has to be careful about incorporating expert opinion into a prior distribution for a clinical trial's design.

Operating Characteristics

One of the biggest challenges to utilizing Bayesian methods when designing studies is having software available to assess the operating characteristics of a design. For any Bayesian design used in practice, the collaborating statistician must provide operating characteristics that summarize the behavior of the proposed method under a wide variety of situations (called *scenarios*). Because these designs typically involve complex models and decision rules, one has to carry out simulations to evaluate the operating characteristics of the proposed design. Some of the characteristics that one typically summarizes are the number of patients assigned to each treatment, the probability of selecting each dose as most efficacious, the probability of stopping a trial if all treatments are too toxic, etc. The statistician typically considers a wide variety of possible scenarios ranging from very pessimistic, such as the case when no treatment provides any benefit, to optimistic cases in which several of the treatments are effective.

Purpose of Checking Operating Characteristics (Calibration)

Controversy Surrounding Evaluation of Frequentist Properties

If one has chosen to demonstrate the frequentist characteristics of the Bayesian design, then one will have to simulate the design under different scenarios. It may seem odd to want to evaluate the frequentist characteristics of a proposed Bayesian design, but some reasons are as follows. First, one may want to convince the non-Bayesian audience that the proposed design offers benefits over standard frequentist designs without incurring a loss in terms of the frequentist characteristics. For example, some sequential designs base their stopping rules on posterior probability calculations, such as Prob(treatment difference > delta | Data) > cutoff. One

can certainly view these posterior probabilities as test statistics, being functions of the data, even though they differ from more common test statistics. Thus, one can evaluate the operating characteristics. Another reason one might want to estimate the operating characteristics of the proposed design is to evaluate how robust the design is under different scenarios. If one feels that the prior distribution is based on rather limited historical information, for example, then one might want to ensure that the prior does not overly dominate inference in certain situations.

Potential Pitfalls

Potential pitfalls include not stopping when one should, stopping a study and later regretting it, and the often perceived possibility that the study's Bayesian analysis will not receive widespread acceptance. The surest way to avoid these problems is to carry out simulations under many, many different scenarios.

EXAMPLES OF BAYESIAN DESIGNS

What Worked and Why

We have seen dozens of successful Bayesian clinical trials at the M. D. Anderson Cancer Center. One characteristic that has contributed to successful implementation is a schedule of regular meetings between the statisticians and the clinical research staff during the trial's design stage. The meetings serve to educate both groups to the other's needs and perspectives. After initiation of patient enrollment, meetings between the research staff and the statistician continue for the purpose of interim review of the trial's progress. Also, the statistician should provide some data management oversight to ensure that the database accurately reflects the trial data.

Clear communication between the clinical investigators and statisticians with respect to what a design can and cannot do is essential. It is also vitally important for the statistician to test the computer code and interface to ensure everything is working properly. Is the program computing the posterior probabilities correctly? Do the results and recommendations in different hypothetical situations make sense mathematically and clinically? Is the user interface (for example, a stand-alone graphical user interface or a Web-based application) intuitive and easily navigated by the individuals who will be using it? Does the interface perform appropriately? These are important questions to address while preparing the protocol and well before the study enrolls the first patient if one wants to realize the full potential of the Bayesian design. When clinical studies with Bayesian designs work well, the benefits of these designs are very much appreciated by the collaborating investigators. Below we give three examples of clinical studies from our institution (from a potential list of dozens).

Correlated Ordinal Toxicity Monitoring in Phase I

In this example, investigators used a Bayesian design within a new statistical framework for dose-finding based on a set of qualitatively different, ordinal-valued toxicities (34). The objective of this trial was to assess the toxicity profile associated with the anticancer drug gemcitabine when combined with external beam radiation to treat patients with soft-tissue sarcoma. The study's design allowed for possible evaluation of a total of 10 gemcitabine doses, combined with a fixed dose of radiation. Traditionally, phase I studies in oncology consider a binary end point as the primary outcome. This binary end point is an indicator of whether or not each patient experienced a dose-limiting toxicity, as defined in the protocol. This single end point reduces all toxicity information across grade or severity of the toxicity and across organ systems into a single yes-or-no outcome. (Berry et al. discuss the use of a hierarchical model to borrow strength across types of toxicities within organ systems in the context of drug safety monitoring [35]). In most phase I oncology settings, however, the patient is at risk of several qualitatively different toxicities, each occurring at several possible levels of severity. Moreover, the different toxicities often are not of equal clinical importance.

The design of this soft-tissue sarcoma phase I study represented a radical departure from conventional phase I study design in oncology. It was based on an underlying probability model that characterized the relationship between dose and the severity of each type of toxicity. The model included a set of correlated normally distributed latent variables to induce associations among the risks associated with the different toxicities. Additionally, there were weights or numerical scores to characterize the importance of each level of each type of toxicity. The statistician met with the physicians prior to initiation of the trial to elicit from them these scores. An algorithm combined the scores associated with each type and level of toxicity with the probability of observing each particular type and level of toxicity. This algorithm produced a weighted average toxicity score. This weighted average toxicity score informed decisions about doses for successive cohorts of patients in this phase I study.

Concerns expressed by the oncologists motivated the development of this design. The clinicians wanted a dose-finding method that would account for the fact

that, clinically, the toxicities that they had identified are not equally important. Additionally, the different toxicities do not occur independently. The investigators also requested that the dose-finding method utilize the information contained in the grade or severity of an observed toxicity. That is, if patients experience a low-grade toxicity at a given dose, while not dose limiting, this event suggests that higher doses may be more likely to lead to a higher grade of that toxicity. The Bayesian framework of this study's design was capable of addressing all of the investigators' concerns regarding characterization of toxicity while also incorporating key design aspects required for institutional approval of the protocol, such as early trial termination for excessive toxicity at the lowest dose. At the end of the study, the model recommended a dose to take forward into phase II, and the investigators were in complete agreement with this choice as the appropriate dose.

Joint Modeling Toxicity and Biomarker Expression in a PhaseI/II Dose-Finding Trial

In this example, the investigators used a Bayesian framework to model jointly a binary toxicity outcome and a continuous biomarker expression outcome in a phase I/II dose-finding study of an intravesical gene therapy for treating superficial bladder cancer (36). Since the toxicity and efficacy profiles of the gene therapy were unknown, the investigators proposed a phase I/II dose-finding study with four possible doses.

This trial's motivation was partially attributable to the increasing use of biomarkers as indicators of risk or as surrogate outcomes for activity and efficacy. In many contexts, the biomarker is observable immediately after treatment, allowing the investigators to learn about the therapeutic potential of the compound without having to wait months or even years as survival data mature. Unlike conventional phase I studies, this study's objective was to determine the *best* dose based on both biomarker expression and toxicity. This dual outcome required a joint model for the two end points. For ethical reasons, the study escalated doses between patients sequentially. An algorithm based on the joint model chose the dose for each successive patient using both toxicity and activity data from patients previously treated in the trial. The modeling framework incorporated a correlation between the binary toxicity end point and the continuous activity outcome via a latent Gaussian random variable. The dose-escalation/de-escalation decision rules were based on the posterior distributions of model parameters relating to toxicity and to activity. The study's stopping rule called for it to stop if the estimated risk of toxicity appeared excessive or if there was clear evidence that the treatment was not modulating the biologic marker.

The Bayesian framework used in this study allowed for flexible modeling of some rather complicated outcomes. In addition, this framework provided a coherent mechanism for incorporating prior information into the modeling process. The study ended, in fact, when it became evident that the drug was not modulating the biologic marker.

Adaptive Randomization

Investigators wished to evaluate the effectiveness of combinations of three drugs (an immunosuppressive agent, a purine analog anti-metabolite, and an anti-folate) to prevent graft-versus-host disease (GVHD) after transplantation (37). The study used adaptive randomization and was to enroll a maximum of 150 patients. A success was defined in this study as "alive with successful engraftment, without relapse, and without a GVHD 100 days after the transplant." The design called for comparing each treatment to the control arm (i.e., the combination treatment with the immunosuppressive agent and anti-folate) in terms of the probability of success in the following manner. Let p_0 be the success probability in the control arm. Similarly, let p_1, p_2, p_3, and p_4 be the success rates in the 4 other treatment arms (three-arm combination treatments with varying doses of the purine analog anti-metabolite). As information accrued about the treatments, the investigators altered the randomization probabilities from equal randomization to biased randomization based on the posterior probability that each treatment-specific success probability exceeded that of the control arm. That is, the randomization would adapt to favor treatments associated with success probabilities that were greater than that of the control via $P(p_k > p_0 \mid \text{data})$ (for $k = 1, 2, 3, 4$) after appropriate scaling.

In addition, the study's design allowed for early stopping based on predictive probabilities. Specifically, the investigators dropped a treatment arm if the predictive probability that its success probability will be greater than p_0 was less than 0.05, given the data at hand and the data yet to accrue. The design was successful in that it limited the number of patients who received the inferior treatments to 18.2% of all of the 110 patients randomized to one of the four experimental arms. By contrast, a design that randomized patients equally to the treatments and did not allow for early stopping would have exposed 50% of patients to these ineffective therapies.

What Did Not Work and Why

When designing clinical studies, the collaborating statistician should be aware of potential pitfalls associated with the design or designs of choice. This is true of Bayesian designs, which may have some unique issues to consider. The most common difficulties include problems with the computer code, such as bugs that lead to incorrect posterior probability calculations; human error in data entry and management; and reconciling differences in how statisticians (or statistical models) define adequate evidence of treatment effects and how physicians define these effects. Below we give examples of three of these potential problems and discuss steps one can take to avoid them.

Over time, Bayesian designs have found more application and become more complicated. While most of the designs developed in the early 1990s focused on binary end points, current implementations include models for time-to-event end points that include parameter effects for treatment, patient-specific covariates (e.g., patient's risk of death) and covariate-by-treatment interactions (e.g., Xian et al. [38]). For very simple designs based on a binary end point, the data management requirements for posterior updating were relatively straightforward. These types of models only require keeping track of the number of patients in the trial and the total number of patients who have experienced the event of interest. In contrast, as the models have become increasingly complex, more data (and more data management) are required for calculation of posterior probabilities. As a consequence, an increase in data management can lead to data entry errors.

For example, Maki et al. (39) describe a two-arm open-label phase II clinical study in sarcoma with tumor response as the primary end point. The study employed a Bayesian adaptive randomization procedure that accounted for treatment-by-sarcoma-subgroup interactions. Specifically, the adaptive randomization scheme incorporated information on the type of sarcoma. After randomizing the first 30 patients equally to the two treatment regimens, the design called for adapting the randomization probabilities for subsequent patients to favor the better performing treatment, according to the accrued data. The investigators subsequently found that the initial recorded sarcoma subtypes for some patients were incorrect. The consequence of this incorrect labeling was that, for one sarcoma subtype, the probability of randomization to the top performing arm was less than it should have been, relative to the other treatment arms. While in this example all patients continued to have higher probability of randomization to the better performing treatment arm, it is conceivable that if such an error were not discovered early, patients could have been randomized to inferior treatments. Therefore, it is extremely important that the statistician be involved with data-management oversight to ensure that such errors do not occur.

One of the key considerations in designing Bayesian clinical trials involves navigating the relationship between the proposed Bayesian model and the realities of medical research. A model may indicate that one treatment confers benefit over another (calculated via posterior probabilities), but if one is claiming this benefit on a very small number of patients, one is going to have a hard time convincing a medical audience that the results are *robust* (robust in an English and not statistical sense). For example, Giles et al. (40) reported a phase II trial that randomized patients to receive one of three treatment regimens: idarubicin and ara-C (IA); troxacitabine and ara-C (TA); and troxacitabine and idarubicin (TI). The study's Bayesian design adaptively randomized patients to the treatments. Initially, there was an equal chance for randomization to IA, TA, or TI, but treatment arms with higher success proportions progressively received a larger fraction of patients. The adaptive randomization led to a total of 18 patients randomized to the IA arm; 11 patients randomized to the TA arm; and just 5 patients randomized to the TI arm. The small sample size associated with the TI arm left this trial open to concerns that the results were not conclusive.

This story is reminiscent of the controversy surrounding the early randomized trials of extracorporeal membrane oxygenation (ECMO) for neonates in respiratory failure. Two early ECMO trials (41, 42) included adaptive randomization algorithms that led to very few babies receiving the non-ECMO treatment. In the end, a vocal part of the medical community seemed to think that these trials included too few patients treated conventionally (i.e., without ECMO) to justify making ECMO the standard treatment for neonates in respiratory distress. (See Ware and related discussion for more information about the ECMO trials [43].) Eventually, a randomized clinical trial without adaptive randomization in the United Kingdom demonstrated the benefit of ECMO (44). The lesson to learn is that one should ensure that the trial will include some minimum number of patients in all treatments (subject to safety assurances) before it begins to adapt the randomization in light of the accruing evidence.

A common criticism voiced by some investigators with whom we have collaborated relates to recommendations based on Bayesian models that do not match the investigators' expectations based on experiences with other designs. This tension is

exemplified in the context of dose escalation decisions in phase I studies in oncology. Although we described and illustrated Bayesian phase I oncology studies earlier in this chapter, most of these phase I studies use non-Bayesian algorithms for dose-finding, such as the 3 + 3 design (45). Their popularity is driven by the fact that clinicians can easily understand these trial designs, and the decision rules employed make intuitive sense. Yet, much is left unspecified in the implementation of these methods. For example, algorithmic designs *implicitly* target toxicity risks smaller than 33% (1 in 3) as being acceptable. In contrast, while Bayesian phase I designs may seem (to some clinicians) to be black boxes, these models make explicit the outcomes being targeted. In particular, all Bayesian designs *explicitly* specify a target probability of toxicity (usually between 25% and 33%). We believe that one of the main reasons this criticism occurs is a lack of communication between the statistician and the clinical investigator. This lack of communication may result, in part, from difficulty explaining these methods to non-statisticians (46). One way to overcome these difficulties is by making the underlying assumptions of the Bayesian model clear to the investigator. One can illustrate these assumptions by providing the investigator with sample trajectories of virtual trials simulated under different scenarios, in addition to providing the operating characteristics of the trial's average behavior (as discussed earlier in this chapter). While potentially time consuming, this type of upfront examination and assessment before the study begins will help the clinician understand both the merits and limitations of the design and underlying model contained in the protocol.

SUMMARY OF RECOMMENDATIONS

In this chapter, we have illustrated the use of Bayesian methods in the design of clinical studies. Although we work with investigators interested in treating cancer, the examples illustrate ideas that are applicable in all disease areas. The main advantages of Bayesian ideas in the design of clinical trials are the inherent flexibility of Bayesian inference; the ease with which one can incorporate information from outside of the study, including measured outcomes of mixed types (e.g., continuous and discrete); the natural notion of evolving knowledge evinced by the transformation from prior uncertainty to posterior uncertainty based on observations; and the way the Bayesian methodology allows one to make decisions and maximize utility, taking into account all uncertainty captured in the basic probability model. Although our examples concerned novel designs and new methodology, Bayesian ideas are applicable when designing any clinical study.

References

1. Carlin BP, Louis TA. Bayesian Methods for Data Analysis, 3rd ed. Boca Raton: Chapman & Hall/CRC; 2008.
2. Little RJ. Calibrated Bayes: A Bayes/Frequentist roadmap. *Am Stat.* 2006;60(3):213–223.
3. Berry DA. A case for Bayesianism in clinical trials. *Stat Med.* 1993;12(15–16):1377–1393; discussion 95–404.
4. Berry DA. Decision analysis and Bayesian methods in clinical trials. In: Thall PF, ed. *Recent Advances in Clinical Trial Design and Analysis*. Boston: Kluwer Academic Publishers; 1995:125–154.
5. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton, FL: Chapman & Hall/CRC; 1999.
6. Berger JO, Wolpert RL. *The Likelihood Principle.* Hayward, California: Institute of Mathematical Statistics; 1984.
7. Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. *Eur J Cancer.* 2007;43(5):859–866.
8. Anscombe FJ. Sequential medical trials (Com: p384–387). *J Am Stat Assoc.* 1963;58:365–383.
9. Armitage P. Sequential medical trials: some comments on F. J. Anscombe's paper. *J Am Stat Assoc.* 1963;58(302):384–387.
10. Armitage P. The search for optimality in clinical-trials. *Int Stat Rev.* 1985;53(1):15–24.
11. Bather JA. On the Allocation of Treatments in Sequential Medical Trials. *Int Stat Rev.* 1985;53(1):1–13.
12. Royall RM. Ethics and statistics in randomized clinical trials. *Stat Sci.* 1991;6(1):52–62.
13. Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med.* 1998;17(14):1563–1580.
14. Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clin Oncol* .1996;14(1):296–303.
15. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med.* 1995;14(4):357–379.
16. Thall PF, Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics.* 2004;60(3):684–693.
17. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics.* 2002;58(4):823–831.
18. Thall PF. A review of phase II/III clinical trial designs. *Lifetime Data Anal.* 2008;14(1):37–53.
19. Berry DA, Müller P, Grieve AP, et al. Adaptive Bayesian Designs for Dose-Ranging Drug Trials. In Gatsonis C, Carlin B, Carriquiry A, eds. *Case Studies in Bayesian Statistics V.* New York: Springer-Verlag; 2001:99–181.
20. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, UK: Wiley & Sons; 2004.
21. DeGroot MH. Optimal Statistical Decisions. New York: McGraw-Hill; 1970.
22. Kadane JB, ed. *Bayesian Methods and Ethics in a Clinical Trial Design.* New York: Wiley & Sons; 1996.
23. Berry DA, Wolff MC, Sack D. Decision making during a phase III randomized controlled trial. *Cont Clin Trials.* 1994;15(5):360–378.
24. Carlin BP, Kadane JB, Gelfand AE. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics.* 1998;54(3):964–975.
25. Stallard N, Thall PF. Decision-theoretic designs for pre-phase II screening trials in oncology. *Biometrics.* 2001;57(4):1089–1095.

26. Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*. 1999;55(3):971–977.
27. Rossell D, Müller P, Rosner GL. Screening designs for drug development. *Biostatistics*. 2007;8(3):595–608.
28. Ding M, Rosner GL, Müller P. Bayesian Optimal Design for Phase II Screening Trials. *Biometrics*. 2008;64(3):886–894.
29. Chen M-H, Ibrahim JG. The relationship between the power prior and hierarchical models. *Bayesian Anal*. 2006;1(3): 551–574.
30. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46–60.
31. Carlin BP, Chaloner K, Church T, Louis TA, Matts JP. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *Statistician*. 1993;42(4):355–367.
32. Carlin BP, Chaloner KM, Louis TA, Rhame FS. Elicitation, monitoring, and analysis for an AIDS clinical trial (with discussion). In Gatsonis C, Hodges JS, Kass RE, Singpurwalla ND, eds. *Case Studies in Bayesian Statistics, Vol. II*. New York: Springer-Verlag; 1995:48–89.
33. Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. *Statistician*. 1993;42(4):341–353.
34. Bekele BN, Thall PF. Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J Am Stat Assoc*. 2004;99(465): 26–35.
35. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60(2):418–426.
36. Bekele BN, Shen Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics*. 2005;61(2):343–354.
37. de Lima M, Couriel D, Munsell M, et al. Pentostatin, tacrolimus, and "mini"-methotrexate for graft-versus-host disease (GVHD) prophylaxis: A phase I/II controlled, randomized study. *Blood (ASH Annual Meeting Abstracts)* 2004;104:727.
38. Xian Z, Suyu L, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clin Trials*. 2008;5(3):181–193.
39. Maki RG, Wathen JK, Patel SR, et al. Randomized phase II study of gemcitabine and docetaxel compared with gemcitabine alone in patients with metastatic soft tissue sarcomas: results of sarcoma alliance for research through collaboration study 002 [corrected]. *J Clin Oncol*. 2007;25(19):2755–2763.
40. Giles FJ, Kantarjian HM, Cortes JE, et al. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *J Clin Oncol*. 2003;21(9):1722–1727.
41. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics*. 1985;76(4):479–487.
42. O'Rourke PP, Crone RK, Vacanti JP, et al. Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized study. *Pediatrics*. 1989;84(6): 957–963.
43. Ware JH. Investigating therapies of potentially great benefit: ECMO (with discussion). *Stat Sci*. 1989;4(4):298–340.
44. UK Collaborative ECMO Trial Group. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. *Lancet*. 1996;348(9020):75–82.
45. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon R. A comparison of two phase I trial designs. *Stat Med*. 1994;13:1799–1806.
46. Rosenberger WF, Haines LM. Competing designs for phase I clinical trials: a review. *Stat Med*. 2002;21(18):2757–2770.