

# Clinical Trials

<http://ctj.sagepub.com>

---

## Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis

Donald A Berry

*Clin Trials* 2005; 2; 295

DOI: 10.1191/1740774505cn100oa

The online version of this article can be found at:  
<http://ctj.sagepub.com/cgi/content/abstract/2/4/295>

---

Published by:

 SAGE Publications

<http://www.sagepublications.com>

On behalf of:



The Society for Clinical Trials

Additional services and information for *Clinical Trials* can be found at:

**Email Alerts:** <http://ctj.sagepub.com/cgi/alerts>

**Subscriptions:** <http://ctj.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

**Citations** (this article cites 4 articles hosted on the  
SAGE Journals Online and HighWire Press platforms):  
<http://ctj.sagepub.com/cgi/content/abstract/2/4/295#BIBL>

# Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis

Donald A Berry

The Bayesian approach and several of its advantages in drug and medical device development are described. One advantage from the perspective of analysis is that it provides a methodology for synthesizing information. However, taking a Bayesian approach to designing clinical trials is potentially more valuable than using this approach in analyzing trial results. Bayesian methodology provides a mechanism for updating what is known as results accumulate during a trial. Such updating can be incorporated completely explicitly and prospectively. An important way in which the Bayesian approach can be used is in calculating the predictive probability distribution of future results on the basis of current results. I show how to exploit predictive distributions in adapting to results that accumulate during the course of a trial. Possible adaptations including decreasing or increasing sample size, dropping treatment arms, and modifying the randomization proportions to the various arms depending on the interim results. Consequences of taking a Bayesian approach to clinical trial design are efficiency, better treatment of patients in the trial, and greater precision regarding the primary endpoints. An example of the last of these is Bayesian modeling of the relationship between early and longer term endpoints. Such modeling also enables earlier decision making. Case studies 2 and 3 deal with trials that were shorter and smaller, respectively, because of such modeling. *Clinical Trials* 2005; 2: 295–300. [www.SCTjournal.com](http://www.SCTjournal.com)

## Introduction

Researchers at M. D. Anderson Cancer Center are increasingly applying Bayesian statistical methods in laboratory experiments and clinical trials. Over 50 current trials at M. D. Anderson have been designed from the Bayesian perspective. In addition, the pharmaceutical and medical device industries are becoming more interested in and are using the Bayesian approach. Many applications in both venues use adaptive methods, which is a primary focus of this presentation.

## The fully Bayesian approach

There are two approaches to implementing Bayesian statistics in drug and medical device development:

a fully Bayesian approach, and using Bayes as a tool to expand the frequentist envelope. Choosing the appropriate approach depends on the context in which it will be used. Is the context that of company decision making, or does it involve the design and analysis of registration studies? Pharmaceutical company decisions involve questions such as whether to move on to phase III, and if so, how many doses and which doses to include, whether to incorporate a pilot aspect of phase III, how many phase III trials should be conducted, and how many centers should be involved. These questions beg for a decision analysis of what can be called a fully Bayesian approach, using the likelihood function, the posterior distribution and a utility structure to arrive at a decision.

---

Frank T. McGraw Memorial Chair for Cancer Research, Professor and Chair, Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center

**Author for correspondence:** Donald A Berry, Frank T. McGraw Memorial Chair for Cancer Research, Professor and Chair, Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 447, Houston, TX 77030-4009, USA. E-mail: [dberry@mdanderson.org](mailto:dberry@mdanderson.org)

## Bayes as a tool

In the context of designing and analyzing registration studies, the Bayesian approach can be a tool to build good frequentist designs. For example, we can use the Bayesian paradigm to build a clinical trial that requires a smaller average sample size. The design may be complicated, but we can find the frequentist operating characteristics using simulation. In particular, we can ensure that the false-positive rate is within the range acceptable to regulatory agencies.

## Why Bayes?

Bayesian methods support sequential learning, allow for finding predictive distributions of future results and enable borrowing strength across studies. Regarding the first of these, we make an observation, update the probability distributions of the various parameters, make another observation, update the distributions again and so on. At any point we can ask which observation we want to make to make next – which dose we want to use for the next patient, for example. Finding predictive distributions – the probabilities that the next set of observations will be of a specific type – is uniquely Bayesian. Frequentist methods allow for calculations that are conditional on particular values of parameters, so they are able to address the question of prediction only in a limited sense. In particular, frequentist predictive probabilities that depend only on the available data are not possible.

The Bayesian paradigm allows for using historical information and results of other trials, whether they involve the same drug, similar drugs or possibly the same drug but with different patient populations. The Bayesian approach is ideal for borrowing strength across patient and disease groups within the same trial and across trials.

Some trials that are proposed by pharmaceutical companies to be conducted in whole or in part at my home institution are deficient in ways that can be improved by taking a Bayesian approach. For example, a company may regard its drug to be most appropriate for a particular disease but is unsure just which subtypes of the disease will be most responsive. So they propose separate trials in the different subtypes. To be specific, consider advanced ovarian cancer, a particularly difficult disease to achieve tumor responses. In exploring the possible effects of its drug a company proposed to treat 35 patients in one group and 35 patients in a second group, but to run two separate trials. All 35 patients would be accrued to each trial with the goal of achieving at least one tumor response. Suppose there were 0 responses out of the 35 patients accrued

in trial 1 and 0 responses out of 25 patients accrued so far in trial 2. By design, they would still add 10 more patients in trial 2. It makes no sense. We are learning two things: first, the drug is not very active, and second, the two patient subgroups respond similarly to the drug. It makes sense to incorporate what has been learned from trial 1 into trial 2. Bayesian hierarchical analysis would enable this. And reasonably such analysis would show that with high probability it is futile (and ethically questionable) to add the remaining 10 patients in trial 2.

Bayesian designs: 1) incorporate sequential learning whenever possible; 2) use predictive probabilities of future results; and 3) borrow strength across studies and patient subgroups. These three Bayesian characteristics have implications for analysis as well as for design. All three involve modeling in building likelihood functions.

## Goals of Bayesian trials

Bayesian goals are to learn faster by using more efficient designs of trials and more efficient drug and medical device development, while at the same time providing better treatment of patients who participate in clinical trials. Physician researchers and patient advocates are particularly interested in the potential for Bayesian trial designs to provide the most effective care possible, while not sacrificing scientific integrity.

## Limitations of traditional trials

Traditional drug development is slow because clinical trials are usually too large, have inflexible designs, focus on single therapeutic strategies, are partitioned into discrete phases, restrict to early endpoints in phase II but employ different, long-term endpoints in phase III and restrict statistical inferences to information in the current trial. The rigidity of the traditional approach inhibits progress. For example, a pharmaceutical company runs a trial with a predetermined sample size and balanced randomization to several doses to learn the appropriate dose for its drug. That is like saying to a student, "Study statistics for N [specify] hours and you will be a statistician". Perhaps the student will become a statistician long before N. Or there may be no N for which this particular student could become a statistician. Drug companies pretend that they know the right dose for an experimental drug after they have carried out the canonical clinical trial(s) designed to answer that question. In fact, they never "know" the right dose.

## Better methods

A clinical trial should be like life: Experiment until you achieve your objective or until you learn that your objective is not worth pursuing. Better methods for drug development are based on decision analyses, flexible designs, assessing multiple experimental drugs, using seamless trial phases, modeling the relationships among clinical and early endpoints and synthesizing the available information. Flexible designs allow the data that are accruing to guide the trial, including determining when to stop or extend accrual. We should broaden the range of possibilities for learning in the early phases of drug development. For example, we should use multiple experimental oncology drugs in a single trial. If we are going to defeat cancer with drugs, it is going to be with selections from lists of many drugs and not with any single drug. So we need to be able to study many drugs in clinical trials. We might use, say, 100 drugs in a partial factorial fashion, while running longitudinal genomic and proteomic experiments. The goal would be to determine the characteristics of the patients who respond to the various combinations of drugs – perhaps an average of 10 drugs per patient. Then, and in the same trial, validating these observations. We cannot learn about the potential benefits of combinations of therapies unless we use them in clinical trials. Considering only one experimental drug at a time in clinical trials is an inefficient way to make therapeutic advances.

Regarding the process of learning, we should do away with the notion of discrete phases of drug development. An approach that is consistent with the Bayesian paradigm is to view that drug development as a continuous process. For example, seamless trials allow for moving from one phase of development to the next without stopping patient accrual. Another approach is allowing for the possibility of ramping up accrual if the accumulating data warrant it. Modeling relationships among clinical and early endpoints will enable early decision making in trials, increasing their efficiency. Synthesizing the available information involves using data from related trials, from historical databases and from other related diseases, such as other types of cancer.

## Examples of the Bayesian approach to drug and medical device development

Some case examples are given to illustrate the Bayesian design characteristics of predictive probabilities, adaptive randomization, seamless phase II/III trials, extraim analyses, hierarchical analysis and safety assessment.

www.SCTjournal.com

## Predictive probability

Predictive probability plays a critical role in the design of a trial and also in monitoring trials. For example, conditioning on what is known about patient covariates and outcomes at any time during a trial allows for finding the probability of achieving statistical significance at the end of the trial. If that probability is sufficiently small, the researchers may deem that continuing is futile and decide to end the trial. Assessing such predictive probabilities is especially appropriate for data safety monitoring boards (DSMBs) quite apart from the protocol, but it is something that can be and should be explicitly incorporated into the design of a trial.

A drug trial at M. D. Anderson for patients with HER2-positive neoadjuvant breast cancer serves as an example of using predictive probability while monitoring a trial [1]. The original design called for balanced randomization of 164 patients to receive standard chemotherapy either in combination with trastuzumab or not (controls). The endpoint was pathologic complete tumor response (pCR). The protocol specified no interim analyses. At one of its regular meetings, the institution's DSMB considered the results after the outcomes of 34 patients were available. Of 16 control patients there were four (25%) pCRs. Of 18 patients receiving trastuzumab, there were 12 (67%) pCRs. The DSMB calculated the predictive probability of statistical significance if the trial were to continue to randomize and treat the targeted sample size of 164 patients, and this turned out to be 95%. They also considered that the trial's accrual rate had dropped to less than two patients per month. They stopped the trial and made the results available to the research and clinical communities. This was many years sooner than if the trial had continued to the targeted sample size of 164. The researchers presented the trial results at the subsequent annual meeting of the American Society of Clinical Oncology.

## Adaptive randomization and early stopping for futility

An M. D. Anderson trial in the treatment of acute myeloid leukemia (AML) serves as an example of adaptive randomization [2]. We wanted to compare the experimental drug troxacitabine to the institution's standard therapy for AML, which was idarubicin in combination with cytarabine, also known as ara-C. Three treatment strategies were compared: idarubicin plus ara-C (IA), troxacitabine plus ara-C (TA), and troxacitabine plus idarubicin (TI). The maximum trial size was set in advance at 75. The endpoint was complete response (CR). Early CR was of special interest because it is an important clinical endpoint in AML, and we modeled time to

*Clinical Trials* 2005; 2: 295–300

CR within the first 50 days. The study design called for randomizing based on the currently available trial results. In particular, when a patient entered the trial we calculated the probabilities that TI and TA were better than IA, and the probability that TA was better than IA, and used those current probabilities to assign the patient's therapy. If one of the treatment arms performed sufficiently poorly, its assignment probability would drop, with better performing therapies getting higher probabilities. An arm doing sufficiently poorly would be dropped. See Giles *et al.* [2] for design specifics.

In the actual trial, the TI arm was dropped after 24 patients. Arm TA was dropped (and the trial ended) after 34 patients, with these final results for CR within 50 days: 10 of 18 patients receiving IA (56%, a rate consistent with historical results); three of 11 patients on TA (27%) and 0 of 5 patients on TI (0%).

These results and the design used have been controversial. Some cancer researchers feel that having 0 successes out of 5 is not enough to abandon a treatment. This would be the case in some settings, but not when there is an alternative that produces on the order of 56% CRs. In view of the trial results, the Bayesian probability that either TA or TI is better than IA is small. Moreover, if either has a CR rate that is greater than that of IA, it is not much greater.

#### Adaptive randomization: screening phase II agents

The traditional approach in drug development is to study one drug at a time. Direct comparisons of experimental drugs with either standard therapies or other experimental drugs are unusual in early phases. Combinations of experimental drugs are anathema. Focusing on one drug means that hundreds of others are waiting their turns in the research queue. Simply because of its size, the queue is likely to contain better drugs than the one now being studied. A better approach is to investigate many drugs and their combinations at the same time. One might screen drugs in phase II in a fashion similar to screening in a preclinical setting. The goal is to learn about toxicity and efficacy of the candidate drugs as rapidly as possible. Another goal is to treat patients effectively, promising them in the informed consent process that if one therapy is performing better then they are more likely to receive it.

Consider a one-drug-at-a-time example in cancer. Suppose the historical tumor response rate is 20%. A standard design for a clinical trial has two stages. The first stage consists of 20 patients. The trial ends after the first stage if four or fewer tumor responses are observed, and also if nine or more tumor responses are observed. Otherwise proceed to the second stage of another 20 patients. A positive result moves the drug into phase III, or to some other type of further investigation. Progress is slow.

Consider an alternative adaptive design with many drugs and drug combinations. Assign patients to a treatment in proportion to the probability that its response rate is greater than 20%:  $r = P(\text{rate} \geq 20\% \mid \text{current\_data})$ . Add drugs as they become available and drop them if their probability of having a response rate greater than 20% is not very high. Drugs that have sufficiently large  $r$  move on to phase III.

As an illustration, consider 10 experimental drugs with a sample size of 200 patients: nine of the drugs have a mix of response rates (20% and 40%), and one is a "nugget," a drug with a 60% response rate. The standard trial design finds the nugget with probability less than 0.70. That is because if the nugget may not be among the first seven or so drugs in the queue. On the other hand, the adaptive design has better than a 0.99 probability of finding the nugget. This is because all drugs have some chance of being used early in the trial. Randomizing according to the results means that the high probability of observing a response when using the nugget boosts its probability of being assigned to later patients. So we identify the nugget with very high probability and, we find the nugget much sooner: after 50 patients of the 200 for being adaptive as opposed to 110 of the 200 in the standard design (conditioning on finding it at all). Adaptive randomization is also a better method for finding the drugs that have response rates of 40%.

If we have many more drugs, such as 100, and proportionally more patients (2000), then the relative comparisons are unchanged from the earlier case. We find the 1/100-drug essentially with certainty and we find it much more quickly using adaptive randomization. The consequences of using adaptive randomization are that we treat patients in the trial more effectively; we learn quickly; and we also are able to identify the better drug sooner, which allows it to move through the process more rapidly. Benefits accrue to both the patient and the drug developer.

#### Seamless phase II and III trial designs

In Dr Temple's presentation at this workshop he included an example of a setting where there is pharmacologic or pathophysiologic information about a patient's outcomes. He suggested that biologic justification of an early endpoint is required. If the early endpoint is to serve as a surrogate for the clinical endpoint in the sense that it replaces the clinical endpoint then I agree. But early endpoints can and should be used whether or not the biology is understood. All that is required is that it be correlated with the clinical endpoint. The possibility of such correlation can be modeled.

If the data in the trial point to the existence of correlation – which may depend on treatment – then the early endpoint is exploited. If the data suggest a lack of correlation then the early endpoint plays no role.

An example is in case study no. 2 of this workshop where we modeled the possible correlation between success of a spinal implant at 12 months and at 24 months. We do not *assume* that those endpoints are correlated, but instead let the data dictate the extent to which the 12-month result supports comparisons of the 24-month endpoint. The focus of the study is the success of the spinal implant at 24 months. The earlier endpoint at 12 months is not a surrogate but rather an auxiliary endpoint.

Case study no. 3 in this workshop provides a second example – for details see [3]. We modeled the possible relationship between stroke scale at early time points, weeks 1 through 12, but the primary endpoint remained week 13 stroke scale. We did not employ anything so crude as “last observation carried forward” but built a longitudinal model and updated the model as evidence about relationships accumulated in the trial.

I will discuss another type of early endpoint here: tumor response. Early information from tumor response is used to construct a seamless phase II/III trial. In conventional cancer drug development, phase II addresses tumor response. Sufficient activity in phase II leads to phase III, which is designed to determine if the drug provides a survival advantage. A conventional phase II process generally requires more than 18 months, after which phase III generally requires at least another two years. In contrast, the seamless phase II/III trial with modeling for the relationship between tumor response and survival can take less than a total of two years.

In a seamless trial [4] we start out with only one or two centers. We accrue a small number of patients per month, randomizing to experimental and control. If the predictive probability of eventual success is sufficiently promising, expand into phase III. All the while the initial centers continue to accrue patients. This is important because these early patients have longer periods of exposure and thus provide the best information about survival.

The seamless design involves frequent analyses and uses early stopping determinations based on predictive probabilities of eventually achieving statistical significance. Specifically, we look at the data every month (for a total of 18 looks in the example) and we use predictive probabilities to determine when to switch to phase III, to stop accrual for futility if the drug's performance is sufficiently bad and to stop for efficacy if the drug is performing sufficiently well.

Inoue *et al.* [4] compare the seamless design with more conventional designs having the same

operating characteristics (Type I error rate and power) and find reductions in average sample size ranging from 30% to 50%, in both the null and alternative hypothesis cases. And the total time of the trial is similarly reduced.

The trial presented in case study no. 3 of this workshop was designed to be seamless phase II/III, but the switch is of a sort that is more dramatic than the one presented above. Namely, it is based on dropping all but two doses, 0 and the one identified from phase II as the ED95.

### Extraim analysis

Extraim analysis is an analog of the interim analysis, addressing the question of continuing beyond the target sample size rather than early stopping. In many drug development trials, the ultimate question remains unanswered at the trial's conclusion. I do not mean that the answer is not the one you like, I mean that you are on the cusp between positive and negative conclusions. From a Bayesian perspective, if the answer is not known, consider expanding accrual. That is the basis of extraim analysis.

As an illustration consider a drug trial proposed to have 800 patients. The power from the trial was not great (80%). As always, there was a possibility of getting to the end of the trial without an answer. I suggested prospectively building in the possibility of extending the trial, depending on the data available at the targeted closure of the trial. Because there was a delay in response the DSMB would not have information on the endpoint for all the patients in the trial at the time for the decision to extend accrual or not. However, under the assumed accrual rate and delay in response evaluation they would have such information for about 450 of the patients. This information would be used to determine the predictive probability of reaching statistical significance once the full information on the patients accrued had become available. If this probability was moderate in size, neither too big nor too small, patient accrual would continue for up to 300 additional patients. A second extensions of accrual for up to another 300 patients increased the maximum sample size to 1400 patients. There are implications of such extensions on the Type I error rate. Appropriate adjustments were made to keep overall one-sided Type I error rate to less than 0.025.

The trial's statistical power increased from 80% to more than 95%. This benefit is effected with only a modest increase in average sample size. The trial is most unlikely to reach the maximum sample size of 1400. When the sample size is larger than the target of 800 patients, it is larger precisely when it has to be larger, that is, only when the answer is not clear otherwise. The alternative is running a second trial,

which is both more time consuming and more expensive.

### Assessing synergy between drugs

In 2003, the FDA approved a combination drug that combines pravastatin, a cholesterol-lowering agent, with aspirin [5–8]. An article in *Science* [9] described this as the only drug (actually, drug combination) that had been approved by the FDA based on the use of an exclusively Bayesian analysis of efficacy. The Bayesian approach was used to synthesize information through a meta-analysis of data from five previous pravastatin secondary prevention trials. Hierarchical modeling allowed for diverse sets of patients within the various trial.

Some experts had suggested that aspirin's beneficial effects might be in the immediate period after an event, such as a myocardial infarction (MI). Pravastatin, on the other hand was thought by some experts to be less important early but to be beneficial in the longer term. So perhaps a patient could as well take aspirin first and follow it by pravastatin, with no benefit of copackaging. To address this possibility and to have a more robust model we considered the hazards over time independently within each year after experiencing an event. We found that, contrary to the above scenario, the benefit of the combination was present in each of the five years following an event.

As indicated earlier in this workshop, Bayesian methods allow for calculating probabilities of variables of most interest. We calculated the probability of synergy (or superadditivity) of the two drugs. We found that the probability of synergy was in excess of 90% for all endpoints considered.

### Safety assessment of Bayesian methods and hierarchical modeling

In discussing an article of mine [10], Chi *et al.* suggested the following: "Safety assessment is one area where frequentist strategies have been less applicable. Perhaps Bayesian approaches in this area have more promise." Safety assessment of drugs involves multiplicities (of various possible adverse effects) and is challenging for all statistical approaches. How to determine whether any of the hundreds of observed differences between drug group and controls is real, and which ones are real?

To address this issue we developed a three-level hierarchical model that borrows strength across the adverse effects, borrows strength within the adverse effects in the same body systems, so there are three levels of experimental units [11]: the patient, the patient within the body system and body system within the set of body systems. So, the model allows

for borrowing strength in several different ways to decide whether an elevated rate of particular adverse effect is caused by the experimental drug.

### Summary

In summary, Bayesian trial designs use predictive probabilities to adaptively randomize, address whether to proceed to the next stage of drug development, possibly seamlessly, extram analysis, and hierarchical modeling in order to borrow strength across studies, study groups, and types of adverse events. Using the Bayesian approach represents a fundamental change in the conduct of medical research. The consequence is more rapid progress in drug development and at lower cost. Moreover, we will be more likely to get the dose right, and we will provide better treatment of patients in clinical trials as well as of those outside and those who benefit from the results of clinical trials.

### References

1. **Buzdar AU, Ibrahim NK, Francis D et al.** Significantly higher pathological complete remission rate following neoadjuvant therapy with trastuzumab, paclitaxel and epirubicin-containing chemotherapy: results of a randomized trial in HER-2-positive operable breast cancer. *J Clin Oncol* 2005; **23**: 3676–85.
2. **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**: 1722–27.
3. **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, 88–181.
4. **Inoue LYT, Thall P, Berry DA.** Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 2002; **58**: 264–72.
5. Retrieved 7 June 2005 from: [http://www.fda.gov/ohrms/dockets/ac/02/slides/3829s2\\_03\\_Bristol-Meyers-meta-analysis.ppt](http://www.fda.gov/ohrms/dockets/ac/02/slides/3829s2_03_Bristol-Meyers-meta-analysis.ppt)
6. **Berry SM, Berry DA, Natarajan K, Lin C-S, Hennekens CH, Belder R.** Bayesian survival analysis with nonproportional hazards: Metaanalysis of pravastatin-aspirin. *J Am Stat Assoc* 2004; **99**: 36–44.
7. **Berry DA.** Statistical innovations in cancer research. In Kufe DW, Pollock RE, Weichselbaum RR, Bast RC, Gansler TS, Holland JF, Frei E III, eds. *Cancer medicine, sixth edition.* London: BC Decker, 2003, 465–78.
8. **Berry DA.** Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science* 2004; **19**: 175–87.
9. **Couzin J.** The new math of clinical trials. *Science* 2004; **303**: 784–86.
10. **Berry DA.** Adaptive clinical trials and Bayesian statistics (with discussion). *Pharmaceutical Report* 2002; **9**: 1–11.
11. **Berry SM, Berry DA.** Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixed model. *Biometrics* 2004; **60**: 418–26.