An Introduction to the Bayesian Analysis of Clinical Trials

Roger J Lewis, MD, PhD*
Robert L Wears, MD, FACEP†

From the UCLA School of Medicine, Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, California; and Division of Emergency Medicine, University of Florida Health Science Center, Jacksonville.†*

Received for publication April 17, 1992. Revision received September 20, 1992. Accepted for publication December 7, 1992.

This study was supported in part by a methodology grant from the Emergency Medicine Foundation and the Society for Academic Emergency Medicine.

Although most clinical trials comparing therapies are analyzed using classical hypothesis testing and *P* values, such methods do not yield the information most useful to the clinician, that is, the probability that one treatment is more efficacious than another. Bayesian inference can yield this probability but only if we quantify our prior beliefs about the possible efficacies of the treatments studied. This article gives a brief introduction to Bayesian methods and contrasts them with classical hypothesis testing. It shows that the quantification of prior beliefs is a common and necessary part of the interpretation of clinical information, whether from a laboratory test or published clinical trial. Advantages of Bayesian analysis over classical analysis of clinical trials include the ability to incorporate prior information regarding treatment efficacies into the analysis; the ability to make multiple unscheduled inspections of accumulating data without increasing the error rate of the study; and the ability to calculate the probability that one treatment is more effective than another. Because it is likely that Bayesian methods will be used more often in the analysis of future clinical trials, investigators and readers should be aware of the two schools of statistical thought and the strengths and weaknesses of each.

[Lewis RJ, Wears RL: An introduction to the Bayesian analysis of clinical trials. *Ann Emerg Med* August 1993;22:1328-1336.]

INTRODUCTION

Clinical trials usually are performed with the goal of determining which of two treatments is more effective. When clinicians read the results from such a trial, their goal should be to estimate the probability that the conclusions drawn from the data are truly correct, based on the plausibility of the results and on the quality of the study design and data analysis. If their estimate of this probability is very high, then they should incorporate the findings of the study into their clinical practice. If their estimate is low, then no change in clinical practice is warranted.

Most conclusions drawn from clinical studies are supported by classical statistical inference. Classical inference, or "classical hypothesis testing," uses a probability value (*P* value) as a measure of the strength of the evidence against a particular conclusion that might be drawn from the data. Classical hypothesis testing is ubiquitous in the analysis of clinical trials, even though the *P* value is, at best, only an indirect measure of the probability that one treatment actually is more efficacious than another.[1] Unlike medical researchers, theoretical statisticians have not embraced classical statistical theory to the point of forsaking other methods of statistical analysis.

Bayesian statistical inference is an alternate method of statistical analysis that is distinct from classical hypothesis testing. It has its roots in an article by Thomas Bayes published posthumously in 1763 that contains what we recognize today as Bayes' theorem.[2] Bayesian methods allow the estimation of the probability that one treatment is more effective than another but only if we quantify prior information regarding the efficacies of the treatments.

Bayesian thinking differs fundamentally from classical thinking in two areas: the nature of the probabilities that we are trying to estimate from the data and the way in which we use the data to modify our estimates of those probabilities. In the classical school of thought, the probability of an event represents the rate or frequency at which the event would occur if the situation in which it might occur was reproduced an infinite number of times. For this reason, the classical school commonly is referred to as the "frequentist" school. Because most medical situations cannot be reproduced exactly a large number of times, most clinically important probabilities are not strictly interpretable in the frequentist context. For example, one cannot create a large number of identical patients with a fever of 38.5 C, right lower quadrant abdominal pain of a given severity, and a WBC count of 16,000 mm^3 to determine the probability that such a patient actually has appendicitis.

Bayesians, on the other hand, view probabilities as estimates of the certainty of an event. This interpretation of probabilities is useful in the clinical context. The behavior of the consulting surgeon, for example, should be determined by his or her subjective estimate of the certainty of the diagnosis of appendicitis.

Bayesian and classical methods differ in the way data are used to reach conclusions. Bayesian analysis is conditional on the observed data; it is concerned with the probability that a conclusion or hypothesis is true given the available data. Classical inference is not conditional on the observed data; rather, it is concerned with the behavior of a statistical procedure over an infinite number of repetitions considering all data that might have been observed, given a hypothesis. Bayesians deal with the probabilities of hypotheses, given a data set, whereas frequentists deal with the probabilities of data sets, given a hypothesis.[3] Because, for example, clinicians are interested in the probability that one treatment is superior to another (rather than the probability of obtaining certain data assuming the treatments are equal), Bayesian thinking is in closer alignment with our natural mode of clinical reasoning.

In this article, we provide a brief review of classical statistical inference and point out some problems that occur with the frequentist point of view. We then provide a nontechnical exposition of Bayesian inference and point out the advantages it provides over the more familiar classical methods. We close by summarizing what we believe are compelling reasons to adopt an informal Bayesian viewpoint in the interpretation of clinical investigations.

CLASSICAL HYPOTHESIS TESTING

The term "classical statistics" refers to that body of theory and technique stemming from work done in the 1920s and 1930s by Fisher, Neyman, and Pearson.[4] Interestingly, the Neyman-Pearson-Fisher theories were considered departures from earlier methods used by Laplace and Gauss, whose thinking was much more along Bayesian lines.[5]

Classical hypothesis testing hangs on a double negative and entails five possible steps: definition of null and alternate hypotheses, calculating a *P* value, and accepting or rejecting the null hypothesis. A classical analysis proceeds as follows:

First, define a null hypothesis. This usually is the opposite of what we believe to be true. For example, in a clinical trial, the null hypothesis might be that the response rate in the treatment group is equal to that in the control group.

Second, define an alternate hypothesis. This typically is an expression of what we think might be true and is what we would like to "prove" statistically. The alternate hypothesis might be that the response rate in the treatment group is greater than that in the control group.

Third, perform a "test of significance" on the null hypothesis. The test of significance is a calculation that first assumes that the null hypothesis is true and then determines the probability of obtaining the observations found in the data, or other observations even more

extreme. This probability is the *P* value. Note that the alternate hypothesis, which is what we are really interested in and generally believe to be true, is not tested.

Fourth, accept or reject the null hypothesis, based on a previously determined standard. If the probability of observing the actual data under the null hypothesis is small (a small *P* value), then we should doubt that hypothesis. The general idea is that if the probability under the null hypothesis of observing the actual results is very small, then there is a relative conflict between the null hypothesis and the observed data, and we should assume that the null hypothesis is not true.

How small is very small? Most investigators and editors simply accept an offhand remark by Fisher that if the probability were less than one in 20 (.05), we would be justified in rejecting the null hypothesis. Thus, the significance level, or α, has traditionally been fixed at $P = .05$, despite good arguments for raising or lowering it in certain circumstances. If the calculated *P* value is less than this standard, then we conclude that the data are not compatible with the null hypothesis, and we reject it. If, on the other hand, the *P* value is not small enough to reject the null hypothesis, we can only conclude that the evidence is not strong enough to contradict the null hypothesis. Note that this is not equal to saying that the null hypothesis is true; it merely says the data are insufficient to refute the null hypothesis. Absence of proof is not proof of absence.

Fifth, accept the alternate hypothesis. If we reject the null hypothesis, we accept the alternate hypothesis by default. Unfortunately, there may be many alternate hypotheses different from the original one that might have been accepted based on this evidence had they been proposed. For example, in a study comparing survival under two treatments, the null hypothesis might be that survival with A and B are equal, whereas one explicit alternative hypothesis might be that survival with A is at least 10% more than that with B. A second alternative hypothesis might be that survival with A is only 5% more, and a third might be that survival with A is only 1% more. If we reject the null hypothesis, we accept by default the alternative hypothesis that was proposed originally; it is not possible using classical hypothesis testing to determine which of the possible alternative hypotheses that might have been proposed is the closest to the truth. Although sample size and power calculations may help in the choice of a realistic alternative hypothesis, they cannot provide an indication of the relative merit of two such alternatives.

Although classical statistical analysis has achieved a high degree of practical success, it has some chronic problems that it seems unable to overcome. The most fundamental is that classical hypothesis testing does not provide the information that the clinician or investigator desires, namely, the probability that the alternate hypothesis, or any other hypothesis, is true.

In addition, to many the double negation is confusing. The most obvious symptom of this confusion is the common misinterpretation of the *P* value. The *P* value actually is the probability that we would obtain the data that we have observed, or data even more extreme, *if the null hypothesis were true*. Clinicians realize that the *P* value is a probability. The probability that interests a clinician is the probability that the alternate hypothesis is true, but classical statistics give us little information about that. Many clinicians then leap to the seemingly natural but erroneous conclusion that the *P* value must be the probability that the null hypothesis is true. Others do not even make it that far, regarding the *P* value as somehow indicating the clinical importance of the results.

A major problem with classical hypothesis testing is that the interpretation of the *P* value obtained at the end of a clinical trial depends on the reason the study was stopped.[6] For example, consider a condition with a well-known mortality of 30%. Suppose investigators have tried a new treatment on a randomly selected sample of ten patients with this condition, of whom only one died, for a sample mortality of 10%. In analyzing this trial classically, we first would define the null hypothesis (mortality is 30%) and the alternate hypothesis (mortality is 10% or less). The *P* value of the classical exact binomial test of the null hypothesis is the probability of obtaining one or fewer deaths in a sample of ten patients with this condition, assuming the true mortality rate to be 30%. This probability is .149, which would not be considered statistically significant, and the null hypothesis that the new treatment had no effect would not be rejected.

Now suppose that the investigators admit that they did not originally have a ten-patient sample in mind. Although they were careful to select patients randomly, they wanted to be sure they had at least one death, so they continued enrolling patients until one patient died, and it just happened that it took a total of ten patients to do this. The sample mortality remains 10%, but the random variable now is the denominator (the final sample size), not the numerator as assumed previously. The classical exact *P* value now becomes the probability of having to enroll ten or more patients before experiencing a death, assuming the null hypothesis is true. This *P* value is .040. Based on the additional knowledge of the investigators' plan, we now would reject the null and embrace the alternate

hypothesis. This leaves us with the uncomfortable observation that, in classical analysis, the intent of the investigator can change the conclusion to be drawn from a single set of data. Even worse, in practice, some mixture of the two extremes given here determines many sample sizes; we cannot say which method of analysis should be used.

A related problem in classical hypothesis testing is that peeking at the data as they accumulate affects the analysis. Consider a study that compares two drugs, A and B, and finds that 24 of 60 patients (40%) have a positive response with drug A and 13 of 61 patients (21%) have a positive response with drug B. A classical analysis of these results by the two-tailed Fisher's exact test yields $P = .031$, leading to the conclusion that drug A is associated with a significantly greater positive response.

Suppose the investigators had analyzed the results at an earlier point, when 12 of 30 patients (40%) responded to drug A and seven of 30 patients (23%) responded to drug B. The two-tailed P value here was .267, so they continued to enroll new patients in the trial because they believed the trend favoring drug A was real, although it was not statistically significant. (If they had obtained a significant P value, they would have stopped and published at this point.) Because the final analysis is a second look, the investigators must adjust the α level for significance to .029 (by the Pocock method) to preserve an overall α of .05 for the entire trial.[7] Now the final data are not significant, and the null hypothesis of no association between drug and response is not rejected. In classical statistics, looking at the data as they accumulate, with the goal of stopping the trial if the results are convincing enough, can change the interpretation of the final set of data. This fact is counterintuitive to most researchers; to a Bayesian, it is nonsensical.

Despite these problems, classical statistical analysis has been almost universally accepted as the standard method for analyzing biomedical data. This acceptance, however, is primarily historical; practical methods for using classical hypothesis testing were developed early and were computationally (if not logically) easier to use in the precomputer era. Few, if any, would assert that the dominance of classical methods over Bayesian ones has resulted from a reasoned, objective comparison of the two approaches.

BAYESIAN INFERENCE

Given the conceptual and practical problems associated with the use of classical hypothesis testing in the analysis of clinical trials, it is not surprising that alternate methods have been explored. The most influential proponent of the Bayesian analysis of clinical trials was Jerome Cornfield, whose work at the National Institutes of Health in the 1960s resulted in a number of landmark papers.[8-13] Subsequently, a number of researchers have continued to work to expand the role of Bayesian methods in the analysis of clinical trials.[3,14-21]

Bayes' theorem allows one to calculate the probability that a particular hypothesis regarding treatment efficacies is true, based on an observed set of data and our estimates (before we knew the data) of the probabilities that various hypotheses were true. These probability estimates, made before observing the data, are called the prior probabilities. In determining the prior probabilities, we should incorporate all available information regarding the efficacies of the control and test treatments. If the information is vague, unreliable, or biased, as in the case in which only anecdotal reports of the efficacy of the test treatment are available, then this uncertainty regarding the accuracy of the prior information can be incorporated quantitatively into the analysis.

A Bayesian analysis entails four basic steps: definition of knowledge prior to the study, acquisition of the data, revision of the prior information to form posterior estimates, and interpretation of the resulting posterior estimates. A Bayesian analysis proceeds as follows:

First, define prior knowledge. Information regarding the likely efficacy of the treatments and the uncertainty in this prior information may be obtained from the medical literature, pilot studies, or recognized experts in the clinical area in which the trial is to be conducted. It is important that the experts represent a range of points of view, so that the uncertainty in their estimates is appreciated. This information is expressed mathematically as a probability distribution. Frequently, it is useful to assume that very little prior knowledge exists to avoid biasing the results if the available information is unreliable.

In most practical situations, the particular form of the prior information has little influence on the final conclusion because it is overwhelmed by the weight of experimental evidence. This practical point prevents one from unduly influencing the trial result by using an overly optimistic or pessimistic set of prior estimates.

Second, acquire the data. Unlike a classical trial, the number of patients to be enrolled in the trial or the timing of the interim analyses do not need to be predetermined. Other considerations, such as the rate of patient recruitment or funding constraints, can be used to determine the number and timing of the data analyses.

Third, revise the prior estimates. The data obtained from the trial are used with Bayes' theorem to revise the

prior estimates of treatment efficacy and create "posterior" estimates. These posterior estimates will have a narrower range of uncertainty, reflecting the additional information now available.

Fourth, interpret the posterior estimates. The posterior estimates contain the information from both the prior estimates and the experimental observations. They allow the calculation of the probability that the efficacy of the control or test treatments, or the difference in efficacy, falls into a given range. There is no arbitrary cutoff point between a negative and a positive trial result. This contrasts with classical hypothesis testing, in which a *P* value of .049 may be considered positive, whereas a *P* value of .051 may be considered negative.

Unlike classical analysis, Bayesian analysis gives the probability that a hypothesis is true. For example, a Bayesian analysis can support a statement such as, "Based on our prior information regarding the treatment's efficacy and on the data observed in this trial, there is an 87% probability that the treatment increases survival by 10% or more over the placebo." This is much more clinically relevant than the classical statement that "the treatment was shown to be statistically significantly associated with increased survival with a *P* value of less than .05."

Because one may calculate from the posterior estimates the probability that the variable of interest lies in any particular range, it is straightforward to find the smallest range for which the varible has a given probability (usually 95% or 99%) of occurring. Such an interval may be interpreted as a 95% or 99% "confidence interval." Although confidence intervals also may be calculated in a classical analysis, classical confidence intervals have a slightly different interpretation. For this reason, Bayesian confidence intervals are called probability intervals, whereas the classical intervals are simply called "confidence intervals." The Bayesian probability intervals actually have the stated probability of containing the true value of the variable of interest, whereas the classical confidence intervals may not. In general, the classical confidence intervals have the property that if one were to construct a large number of 95% confidence intervals, approximately 95% of them would contain the true value of the variable of interest.

When first exposed to Bayesian analysis, many investigators are bothered by the requirement of seemingly subjective estimates of prior probabilities. They may be reassured on two grounds. First, the effect of prior probabilities is large only when the data are weak and unconvincing; in such situations, clinicians feel free to act on their intuition, which actually is their prior probability. Second, physicians actually are quite used to using prior estimates in interpreting data every day, but their incorporation of prior probabilities is not quantitative. Several clinical examples follow.

Consider the decision to admit a 34-year-old man with a fever of 38.8 C who has no identifiable source of infection, looks clinically well, and has had an extensive emergency department evaluation. If the patient has no pertinent medical history, he can be sent home with appropriate instructions. If, on the other hand, the patient has had a recent splenectomy for trauma, then he should be admitted. Most physicians are comfortable with such distinctions and never stop to think that all of the clinical data (eg, physical examination, laboratory and radiographic results) might be the same in two patients, yet the appropriate interpretation of those data might lead to diametrically opposite conclusions based on the physician's subjective prior estimate of the probability that the patient has a clinically important occult bacterial infection. In the patient with no medical history, the prior estimate is almost zero, and the observed data do not increase the estimated probability to a degree that would warrant admission. For the recently splenectomized patient, the physician's prior estimate of the probability of occult bacteremia is low to moderate, and when this estimate is modified by the observed data (fever and the lack of an identifiable source of infection), the resulting estimate is quite high, warranting admission and empiric antibiotic therapy pending culture results.

Clinicians use similar prior probability estimates in the interpretation of all test results. Consider the interpretation of an ECG with nonspecific T-wave changes. If it were obtained from a 54-year-old man with a history of two hours of pressure-like chest pain, long-standing untreated hypertension, and smoking, then the estimated prior probability of ischemic heart disease would be quite high, and after seeing the ECG, the estimate of this probability would remain high. Suppose, however, that after the physician reads the ECG and admits the patient to the CCU, it is discovered that the ECG was obtained in error from an 18-year-old woman with an apparent ankle sprain who was in an adjacent bed. Suddenly, the appropriate interpretation of the ECG is different, and the T-wave changes should be attributed to normal variation. This is because the physician's estimate of the prior probability that the young woman has occult cardiac ischemia is vanishingly small. What clinicians call the "clinical setting" usually is an estimate of prior probabilities.

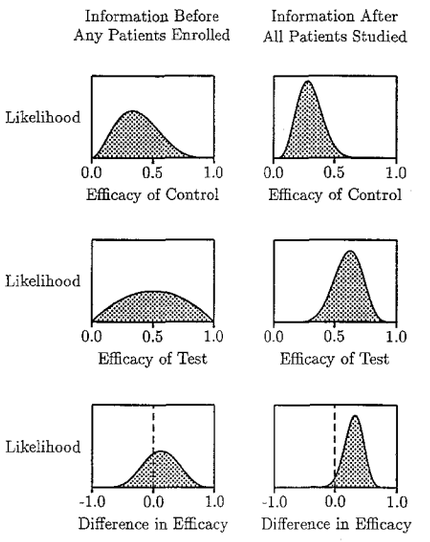
Using the example of the ECG obtained on the wrong patient, it is useful to consider the meaning and interpretation of a classical *P* value. The *P* value represents the

probability, under the null hypothesis, of obtaining the observed data or data more inconsistent with the null hypothesis. For the ECG example, this would be the probability of obtaining an ECG in a normal subject as or more inconsistent with the null hypothesis that there is no cardiac ischemia. Suppose the proportion of ECGs from normal subjects as or more suggestive for ischemia than the one obtained is one in 25, or .04. Then, using classical hypothesis testing, the null hypothesis that there is no cardiac ischemia should be rejected, regardless of the characteristics of the patient. This is an inappropriate conclusion for the young woman. If, on the other hand, the proportion as abnormal or more abnormal was one in ten, or .10, then the null hypothesis that there is no ischemia would be accepted. This is an inappropriate conclusion for the 54-year-old man. This example illustrates that physicians can interpret clinical tests properly only if they behave as Bayesians, incorporating their estimates of the prior probabilities of significant disease into their interpretation of all results.

If we accept the necessity of incorporating prior information into the analysis of clinical information, then conceptually it is straightforward to extend this idea to the Bayesian analysis of a clinical trial. Assume that we are comparing the efficacy of a new treatment, the "test" treatment, with that of a placebo, the "control" treatment, for a disease with only two outcomes: "success" and "failure." The symbols P_c and P_t denote the probability of a successful outcome with the control and test treatments, respectively. We will denote the true difference in efficacy between the two treatments with Δ where $\Delta = P_t - P_c$. We begin by determining what prior information is available for either treatment. As often is the case, more information will be available regarding the control therapy than the test therapy. The three left panels of Figure 1 graphically represent the prior information. Despite the available information, there remains significant uncertainty about which therapy is superior; a substantial fraction of the area under the curve, in the bottom of the left panels, falls on either side of the line at $\Delta = 0$.

Because of uncertainty about which therapy is superior, patients are enrolled in the trial. Suppose that 12 patients receive the control therapy and three have successful outcomes, whereas 14 receive the test therapy and nine have successful outcomes. Unlike a classical analysis, when using a Bayesian analysis we need not concern ourselves with the reason for stopping at this point, whether this is an interim or final analysis, or the number of analyses before this one. Given these data, Bayes' theorem can be used to update our estimates of the treatment efficacies.

Figure 1.
Information before and after a small clinical trial. The prior information is shown by the three left panels, and the information after the trial (posterior information) is shown by the three right panels. The upper left panel shows the probability estimates for the efficacy of the control treatment, prior to enrolling any patients in the trial. The height of the curve at any value of control efficacy shows the likelihood of that value. The control is thought to lead to a successful outcome in approximately 20% to 60% of patients, although values of a few percent to 80% are considered possible. Thus, the width of the curve shows the degree of uncertainty regarding the true efficacy of the control. The middle left panel shows the prior probability estimates for the efficacy of the test treatment. This curve is wider than the curve for the control treatment, showing greater uncertainty about the true efficacy of the test therapy. The prior estimates for the control and test therapies can be used to calculate a prior estimate for the difference in efficacy, and this is shown in the bottom left panel. A positive difference in efficacy signifies that the test therapy is more efficacious than the control. The prior estimates reflect some weak evidence that the test therapy is better than the control, although both positive and negative values for the difference in efficacy are quite likely. The three right panels show the updated (posterior) efficacy estimates, based on the results of the trial. All of the curves are narrower than their respective prior estimates, reflecting the greater precision now possible in our estimates of the control and test efficacy and in our estimate of the difference in efficacy. It can be seen from the bottom right panel that it is nearly certain that the difference in efficacy is positive, that is, that the test therapy is more effective than the control.



The right panels of Figure 1 show the results of incorporating the new information from these data. The bottom of the right panels shows the posterior estimates for the difference in efficacy. Clearly, the overwhelming probability is that $\Delta > 0$. Thus, we conclude that the test therapy is more effective than the control therapy.

Bayesian analysis is inherently sequential. The prior information used for the interpretation of later data in a clinical trial actually is the posterior distribution obtained from analysis of early data. By using the posterior distribution at any analysis point as the prior distribution for the next bit of data to be incorporated, a Bayesian analysis automatically "adjusts" for previously obtained information, combining information from the new data with information from earlier data and from the medical literature or expert opinion.

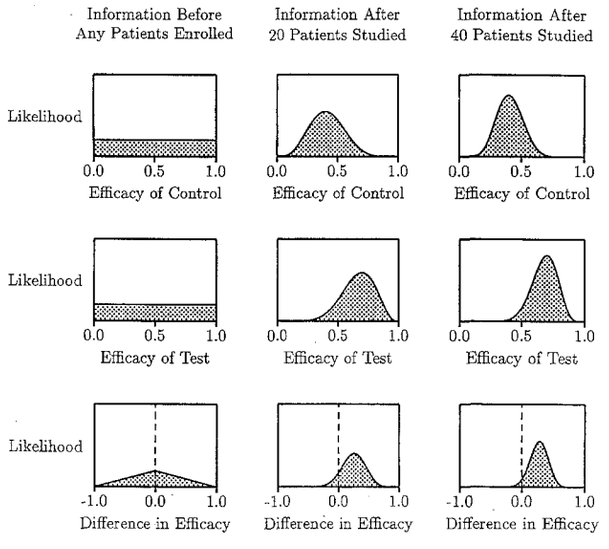
To clarify the differences between the classical and Bayesian methods of data analysis, especially with respect to interim analyses of data, we now compare a quantitative Bayesian analysis and the classical analysis of a clinical trial with one interim data analysis.

Table.
Comparison of Bayesian and classical analyses

Bayesian Analysis	Classical Analysis
1. Define prior knowledge. Although we may have some "hunches" about which treatment is superior, we decide that a conservative approach would be to assume that almost nothing is known about the success rates with the two treatments. Another way of expressing this lack of knowledge is to consider equally likely any success rate from 0% to 100%. This is illustrated in the top and middle left panels of Figure 2. The curves for successes for both treatments are flat, indicating that all success rates are equally likely. These two curves (called probability densities) can be combined to produce a prior estimate of the probability density for the difference in the success rates. This is shown in the bottom left panel of Figure 2. Although the curve is very wide, reflecting great uncertainty in our knowledge about the true difference in efficacy, values close to zero are considered most likely.	1. Define the null hypothesis. We choose a null hypothesis that the success rate for the treatments are equal.
2. Acquire the data. We then enroll 20 patients, randomly assign them to treatment groups, and observe their outcomes.	2. Define an alternate hypothesis. We decide that if the success rate with B were more than that with A by 20% or more (eg, 40% compared with 60%), then clinicians should abandon A and adopt treatment B. This is the smallest treatment effect we would like to detect. With .05 as the maximum significant *P* value, power calculations show that we have only a 34% chance of detecting such a difference if it is present, but we would have a 70% chance of detecting a difference of 30%. Accordingly, we select a difference in efficacy of 30% as our alternative hypothesis.
3. Revise prior estimates. Using Bayes' theorem to combine the prior distributions in Figure 2 with the observations available (four successes of ten with treatment A and seven successes of ten with treatment B), we obtain new posterior probability densities for the efficacies of treatments A and B. These new probability densities are shown by the top and middle panels in the middle column of Figure 2. These densities then are used to obtain a new posterior density for the difference in efficacy, as shown by the bottom of the middle panels in Figure 2.	3. Acquire the data. We now begin to enroll the first 20 patients, randomizing them to treatments A and B, and observing their outcomes.
4. Interpret posterior estimates. Notice that the posterior density curve peaks near a difference of .30, which is the observed difference in efficacy (40% versus 70%). The probability that treatment B is superior to treatment A is the area under the curve to the right of the dashed line marking zero difference. This probability is .901. The probability that the absolute difference in efficacy is more than 20% is .613, and the 95% probability interval for the true difference in efficacy runs from −.12 to .61. Thus, at this point we are 90% sure that B is better than A and more than 60% sure that the benefit is at least a difference of 20% in raw success rates.	4. Perform a test of significance. We calculate a test of significance at the interim analysis point, adjusting our maximum significant *P* value to .029.[7] Fisher's exact test on these results yields a two-tailed *P* value of .370.
5. Acquire the data. Because the 95% probability interval for the difference in efficacy is still quite wide and includes both positive and negative differences, we decide to continue the trial with the enrollment of 20 additional patients.	5. Accept the null hypothesis, or reject it and accept the alternative hypothesis. Since our interim *P* value was not smaller than the adjusted maximum significant *P* value of .029, we cannot reject the null hypothesis and stop the trial at this point. We also cannot assume that the null hypothesis is true, only that the evidence from these 20 patients is not sufficient to doubt its truth.
6. Revise prior estimates. The information obtained on the second group of 20 patients enrolled now can be used to revise our previous estimates. Put another way, the posterior estimates from the first phase of the trial now are used as prior estimates for the interpretation of the data from the second group of 20 patients. The resulting updated probability densities for the treatment efficacies and the difference in efficacy are shown in the three right panels of Figure 2.	6. Acquire the data. We continue to enroll another 20 patients, for a total of 40 patients, to reach the final termination point for the trial.
7. Interpret posterior estimates. The updated posterior distribution has a much sharper peak around a difference of .30. The area to the right of the dashed line is .969, so we are approximately 97% sure that treatment B is better than treatment A. The 95% probability interval extends from .00 to .55, so we are 95% sure that the true difference in efficacy lies in this range.	7. Perform a test of significance. A second significance test is performed after all 40 patients have been evaluated. Fisher's exact test on the final results yields a *P* value of .111, which is still more than the adjusted maximum significant *P* value of .029.
	8. Accept the null hypothesis, or reject it and accept the alternative hypothesis. There still is insufficient evidence to reject the null hypothesis. We conclude that there is not a statistically significant difference between the two treatments but with the caveat that given our small sample size, our chance of detecting a difference of 20% in success rate was less than 35%.

The sample study will compare the effects of two treatments, A and B, on a dichotomous outcome, "success" and "failure." Thus, the study result may be summarized by the difference in the proportion of successes after treatments A and B. We will assume that the budget will support enrollment of as many as 40 patients. We also have a secondary goal of exposing as few patients as possible to an inferior treatment, if there is one. The design and the resulting data are as follows. After the first 20 patients have been evaluated, an interim data analysis will be performed to see if the trial may be stopped early and a reliable conclusion drawn from the data. At that time, four of ten patients (40%) will have had a successful outcome with treatment A, and seven of ten patients (70%) will have had a successful outcome with treatment B. If the analysis of these data shows that the trial cannot be stopped at this point, then an additional 20 patients will be enrolled, and then the trial will be stopped again. At this point, eight of 20 patients (40%) will have had a successful outcome with treatment A, and 14 of 20 patients (70%) will have had a successful outcome with treatment B. The Bayesian and classical approaches to the analysis of these data are compared in the Table.

Figure 2.
This figure illustrates the information available about the efficacy of two treatments during a clinical trial with one interim analysis. The left panels show the prior information available before the trial, the middle three panels show the information available at the interim analysis, and the right panels show the information at the conclusion of the trial.



Despite having many advantages over classical statistical methods, the Bayesian approach is not without its difficulties. Often, the quantification of prior information is difficult or, even worse, different equally valid sources of prior information may contradict each other. Often, it is prudent to show that a given Bayesian analysis will yield a conclusion that is independent of the choice of prior probabilities. Such an analysis is called "robust" because it can stand up to different interpretations of the prior information. In addition, the Bayesian approach can be computationally complex. Note that we give no equations for the calculation of Bayesian probabilities or confidence intervals. With the advent of inexpensive and powerful personal computers, however, the numeric complexity of Bayesian analysis is no longer a real barrier to its use. Detailed information about the mechanics of performing Bayesian analyses can be found in any of several general references on Bayesian methods.[23-27]

SUMMARY

Bayesian methods offer an alternative to classical hypothesis testing for the analysis of clinical trials. The major advantage of Bayesian analysis is ease of interpretation of the results. Only a Bayesian analysis can provide a direct estimate of the probability that one treatment is superior to another. The Bayesian approach is more consistent with clinicians' common sense method of reasoning and emphasizes the estimation of effect magnitude rather than the artificiality of hypothesis testing. Although we have mentioned only simple analyses of dichotomous outcomes, Bayesian methods have been developed for more complex analytic problems such as regression, analysis of variance, and adjustment for confounding variables.

In addition, Bayesian inference is not subject to some of the internal inconsistencies and practical problems that plague classical hypothesis testing. Bayesian inference permits unlimited inspection of the data as they accumulate. This is a strong argument for its use in clinical trials because it may be possible to terminate trials earlier, thus exposing fewer patients to ineffective or harmful therapy. Because the posterior estimates from one trial can be used as the prior estimates for another, Bayesian methods can be used to combine results from different trials in a quantitative and coherent way.

Clinical investigators and readers of clinical trials should be aware of the two statistical schools of thought and, more important, of the inherent weaknesses of

classical methods. In the future, it is likely that Bayesian methods will be used more often in the analysis of clinical trials and that this use will stimulate the development of increasingly practical methods for such analyses.

REFERENCES

1. Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 1983;98:385-394.

2. Bayes T: An essay towards solving a problem in the doctrine of chances. *Phil Trans Roy Soc* 1763;53:370-418. Reprinted with a biographical note by Barnard GA in *Biometrika* 1958;45:293-315.

3. Berry DA: Interim analysis in clinical trials: The role of the likelihood principle. *Am Stat* 1987;41:117-122.

4. Neyman J, Pearson ES: The testing of statistical hypotheses in relation to probabilities a priori. *Proc Cambridge Phil Soc* 1933;29:492-510.

5. Efron B: Why isn't everyone a Bayesian? *Am Stat* 1986;40:1-11.

6. Iverson GR: *Bayesian Statistical Inference.* Newbury Park, California, Sage Publications, 1984.

7. Geller NL, Pocock SJ: Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics* 1987;43:213-223.

8. Halperin M, DeMets DL, Ware JH: Early methodological developments for clinical trials at the National Heart, Lung and Blood Institute. *Stat Med* 1990;9:881-892.

9. Meier P: Jerome Cornfield and the methodology of clinical trials. *Control Clin Trials* 1981;1:339-345.

10. Ederer F: Jerome Cornfield's contributions to the conduct of clinical trials. *Biometrics* 1982;38(suppl: *Curr Topics Biostat Epidemiol*):25-32.

11. Cornfield J: Sequential trials, sequential analysis and the likelihood principle. *Am Stat* 1966;20:18-23.

12. Cornfield J: A Bayesian test of some classical hypotheses—With applications to sequential clinical trials. *J Am Stat Assoc* 1966;61:577-594.

13. Cornfield J: The Bayesian outlook and its application. *Biometrics* 1969;25:617-657.

14. Berry DA: Interim analysis in clinical trials: Classical vs Bayesian approaches. *Stat Med* 1985;4:521-526.

15. Berry DA, Ho CH: One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* 1988;44:219-227.

16. Berry DA: Monitoring accumulating data in a clinical trial. *Biometrics* 1989;45:1197-1211.

17. Spiegelhalter DJ, Freedman LS: Bayesian approaches to clinical trials, in Bernardo JM, DeGroot MH, Lindley DV, et al (eds): *Bayesian Statistics 3.* Oxford, Oxford University Press, 1988, p 453-477.

18. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986;5:421-433.

19. Freedman LS, Spiegelhalter DJ: Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Control Clin Trials* 1989;10:357-367.

20. Lewis RJ: Sequential Bayesian analysis of clinical trials (abstract). *Ann Emerg Med* 1990;19:493.

21. Lewis RJ, Berry DA: A comparison of Bayesian and classical group-sequential clinical trial designs (abstract). *Ann Emerg Med* 1992;21:641.

22. Fleiss JL: *Statistical Methods for Rates and Proportions.* New York, John Wiley & Sons, 1973.

23. Polard WE: *Bayesian Statistics for Evaluation Research.* Beverly Hills, California, Sage Publications, 1986.

24. Lee PM: *Bayesian Statistics: An Introduction.* New York, Oxford University Press, 1989.

25. Phillips LD: *Bayesian Statistics for Social Scientists.* London, Thomas Nelson & Sons, 1973.

26. Box GEP, Tiao GC: *Bayesian Inference in Statistical Analysis.* Reading, Massachusets, Addison-Wesley, 1973.

27. Press SJ: *Bayesian Statistics: Principles, Models, and Applications.* New York, John Wiley & Sons, 1989.

Dr Lewis thanks Donald Berry for many useful discussions regarding the Bayesian analysis of clinical trials.

Address for reprints:

Roger J Lewis, MD, PhD

Department of Emergency Medicine, D9

Harbor-UCLA Medical Center

1000 West Carson Street

Torrance, California 90509