A GUIDE TO DRUG DISCOVERY

Bayesian clinical trials

Donald A. Berry

Abstract | Bayesian statistical methods are being used increasingly in clinical research because the Bayesian approach is ideally suited to adapting to information that accrues during a trial, potentially allowing for smaller more informative trials and for patients to receive better treatment. Accumulating results can be assessed at any time, including continually, with the possibility of modifying the design of the trial, for example, by slowing (or stopping) or expanding accrual, imbalancing randomization to favour better-performing therapies, dropping or adding treatment arms, and changing the trial population to focus on patient subsets that are responding better to the experimental therapies. Bayesian analyses use available patient-outcome information, including biomarkers that accumulating data indicate might be related to clinical outcome. They also allow for the use of historical information and for synthesizing results of relevant trials. Here, I explain the rationale underlying Bayesian clinical trials, and discuss the potential of such trials to improve the effectiveness of drug development.

Frequentist
An approach to statistical inference that is an inverse of the Bayesian approach. The focus is on the probability of results of a trial — usually including the observed data — assuming that a particular hypothesis is true. For example, a frequentist P-value is the probability of observing results as extreme as or more extreme than the observed results assuming that the null hypothesis is true.

Department of Biostatistics and Applied Mathematics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, BOX 447, Houston, Texas 77030-4009, USA.
e-mail:
dberry@mdanderson.org
doi:10.1038/nrd1927

Statistical thinking has had a central role in raising the scientific standards of clinical research over the last two centuries, especially during the past 50 years. A major reason has been the appreciation of statistical inference by drug- and medical-device-regulatory agencies. Traditional frequentist statistics has had the dominant, and often exclusive, role in this scientific renaissance. The greatest virtue of the traditional approach may be its extreme rigour and narrowness of focus to the experiment at hand, but a side effect of this virtue is inflexibility, which in turn limits innovation in the design and analysis of clinical trials. Because of this, clinical trials tend to be overly large, which increases the cost of developing new therapeutic approaches, and some patients are unnecessarily exposed to inferior experimental therapies.

Owing to such issues, there is increasing interest in Bayesian methods in medical research. Advances in computational techniques and power are also facilitating the application of these methods (BOX 1). More than 100 ongoing clinical trials at the University of Texas M. D. Anderson alone have been designed or are being monitored from the Bayesian perspective. And of recent medical device approvals by the Center for Devices and Radiological Health of the US FDA, ~10% are based on Bayesian designs and analyses, as compared with none 10 years ago. Furthermore, at least one drug (Pravigard Pac; Bristol-Myers Squibb) was approved by the FDA on the basis of Bayesian

analyses of efficacy (BOX 2). And in May 2004, the FDA co-sponsored a workshop to address the role of Bayesian approaches in drug and medical device development, 'Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision-Making?' (The video/audio presentations are available as webcasts; see Further information).

After setting the context of the Bayesian approach by describing the frequentist perspective and relating the two approaches, this article discusses the Bayesian approach to the design and analysis of clinical trials, and to drug and medical device development more generally. The goal is to improve drug and medical device development, in terms of costs and the effective treatment of patients, both those in and those outside of clinical trials, and the Bayesian approach provides a better perspective, and a more efficient methodology, for accomplishing this goal. It should be emphasized though that I want to preserve the high scientific standards wrought by the hard and effective work of statisticians and other scientifically oriented clinical researchers during the past 50 years (indeed, the Bayesian approach is more closely in line with the scientific method[1]).

Statistical inference
Statistical inferences are based on mathematical models of experiments, including clinical trials. Each model corresponds to a 'state of nature', the underlying process that produces the experimental results. Candidate

Successes	0	1	2	3	4	5	6	7	8	9	10
$p = 0.35$	0.013	0.072	0.176	0.252	0.238	0.154	0.069	**0.021**	0.004	0.001	0.000
$p = 0.70$	0.000	0.000	0.001	0.009	0.037	0.103	0.200	**0.267**	0.233	0.121	0.028

Figure 1 | **Probabilities for a hypothetical clinical trial.**

Bayesian
An approach to statistical inference that uses Bayes rule. The focus is on the probability that a hypothesis is true given the available evidence.

Parameter
A population characteristic, such as the tumour-response rate when patients are treated with a particular therapy. Parameters serve to index the possible distributions of trial results. Parameters can never be known precisely because populations can never be fully assessed. Statistics are analogous to parameters but apply for samples from a population and so statistics are known when the sample becomes available. Statistics are commonly used to estimate parameters, such as when a response rate from a clinical trial is used to estimate the response rate of all patients having the disease in question.

Null hypothesis
An underlying state of nature in which there is 'no difference' among two or more treatments or interventions. A hypothesis is a specific value of a parameter. So a null hypothesis is when the parameter is an indicator of treatment effect and the value specified corresponds to no effect.

models are indexed by a number called a parameter. A natural parameter for an experiment that produces either successes (such as treatment responses) or failures is the proportion, p, of successes in some greater population of experimental units. The population is not fully observable and so the parameter can never be known with certainty. Observation is restricted to a sample from the population. This sample is generated by an experimental process. For example, a sample might be SSFSSFSSSF, consisting of seven successes and three failures. A statistical inference is a statement about the unknown value of p based on the sample, and in this example the sample proportion 7/10 might be adopted as an estimate of p.

Frequentist approach: the basics
When I use the term 'frequentist' I mean the Neyman–Pearson approach[2] that dominated biostatistics in the latter half of the twentieth century. In this approach, parameters are regarded as fixed and not subject to probability distributions. Probabilities are associated with experimental observations and can be calculated only by assuming fixed values of the various parameters. In our example, frequentists calculate the probability of observing data SSFSSFSSSF for values of p that are of interest. They also calculate the probabilities of observing other possible but unobserved results of the experiment for these same values of p. The set of 'other possible results' depends on the experimental design. For example, if the design called for making exactly 10 observations then there are 11 possible outcomes, one for each of 0, 1, 2, …, 10 successes, including 7, the outcome actually observed. Another possible experimental design is to make observations until obtaining the third failure. For such a design there are an infinite number of possible results, corresponding to 0, 1, 2, … experimental units, including 10, the actual observation.

Suppose that an experimental treatment is being used in a disease for which the historical success rate for standard treatment was 35%. Consider the first of the above two designs, the one that calls for treating exactly 10 patients. An important value of p is 0.35, called the null hypothesis. Probabilities of the possible results assuming $p = 0.35$ are shown in FIG. 1.

So when $p = 0.35$, the probability of the actual observation (7 successes) is 0.021. The conventional frequentist approach is to add in the probabilities of more extreme results — in this case 8, 9 or 10 successes — giving 0.026. This sum is called the significance level, or more briefly, the *P-value*. (P-values are usually made two-sided by including probabilities in the opposite tail of the distribution for observations with probabilities smaller than for the actual observation; including 0 successes as evidence against $p = 0.35$ would give a two-sided P-value of 0.039. Sometimes an approximate two-sided P-value is calculated by doubling the one-sided P-value.) The P-value is a frequentist measure of evidence against the null hypothesis that $p = 0.35$, with smaller P-values meaning stronger evidence against the null hypothesis. Conventionally, the results are called *statistically significant* if the P-value is less than 0.05, as in this example.

Had the design of the trial been other than taking exactly 10 observations, then the P-value for these data would be different as well. For example, if the design was to continue the trial until obtaining the third F and the results were SSFSSFSSSF (as before) then the P-value would have been 0.004, an order of magnitude smaller than the first P-value. So the evidence is now stronger that the success rate on the experimental treatment is greater than the historical rate, even though the results of the experiment are identical. This essential tie between trial design (and the intentions of the investigator) and consequent inferences characterizes the frequentist approach and exemplifies its inflexibility (as demonstrated below, Bayesian conclusions are the same for both designs). For example, investigators cannot change the design of the trial in mid-course for otherwise no frequentist inferences are possible. And if an investigator fails to specify a trial's stopping rule in advance (or fails to adhere to what is specified), then again no frequentist conclusion is possible, even if the trial is conducted with the utmost integrity and a conscientious desire to learn.

Bayesian approach: the basics
The defining characteristic of any statistical approach is how it deals with uncertainty. Unlike the frequentist approach, in the Bayesian approach all uncertainty is measured by probability. Anything that is unknown has a probability distribution. Everything that is known is taken as given and all probabilities are calculated conditionally on known values. In the example, because p is unknown, it has a probability distribution. This distribution can be used for calculating such quantities as the probability that p is equal to 0.35 or greater than 0.50 and so on. If experimental results are unknown — such as before the experiment — then they too have probabilities. However, once the results of an

Box 1 | Computational techniques for Bayesian analysis

One spur for the increased use of Bayesian methods in clinical research has been the improvement of computational techniques and the widespread availability of high-speed computers. Bayesian methods that have always seemed right and proper could not be carried out because of computational limitations, but this is no longer the case. With the availability of modern computational tools, essentially any Bayesian design or analysis can be constructed and validated. However, Bayesian software is not nearly as refined or as widely available as frequentist software. It is not difficult for statisticians to write their own Bayesian computer routines, but it is time-consuming. And the programs will require validation. An excellent set of programs called WinBUGS (Windows version of Bayesian inference Using Gibbs Sampling) is available online (see The BUGS Project in Further information). Moreover, SAS has some (mainly high-level) Bayesian macros and plans for incorporating additional Bayesian applications. However, available Bayesian software is limited.

## Box 2 | Bayesian analysis in regulatory decision-making: Pravigard Pac

In June 2003, the Center for Drugs and Experimental Research of the US FDA approved Pravigard Pac (Bristol-Myers Squibb) based on Bayesian analyses of efficacy that Scott Berry and I had carried out. Pravigard is a combination of pravastatin (Pravachol; Bristol-Myers Squibb), a cholesterol-lowering drug, with aspirin. The FDA had approved these two agents previously for the secondary prevention of cardiovascular events. In this review, I consider the occurrence of any myocardial infarction, whether fatal or not. We used the results of five secondary prevention trials in which pravastatin had been randomized and aspirin use had been recorded but not randomized. The Bayesian approach is ideally suited for synthesizing information from multiple heterogeneous sources. In addition, its focus on probabilities of hypotheses for existing data makes it ideal for retrospective analyses.

The FDA's approval was based on the posterior probability that the combination is more effective than either agent alone. We also provided the posterior probability that the combination is synergistic in the sense that the effect of the combination is greater than the sum of the effects of the separate agents.

Because aspirin use was not randomized it was important to adjust for baseline covariates. We adjusted for age, gender, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglyceride level, systolic and diastolic blood pressures, previous cardiac event and smoking status. We also considered the possibility of different treatment effects in the different trials (using a hierarchical model). An important concern was the attitude of some experts that aspirin was effective in the immediate post-event setting but that its effect might dissipate over time. Conversely, lowering cholesterol with pravastatin was thought by some to be less important in the acute setting but more important in the longer term. So the combination might not be necessary but instead aspirin could be taken early and later replaced by a statin. To address this possibility, we modelled the different treatment hazards over time, out to 5 years.

The accompanying figure shows the posterior means of the cumulative event rates for the four treatments and also for the 'Sum' of the two agents individually. The Bayesian approach allows for calculating the posterior probability of various treatment comparisons. The table gives the posterior probabilities that the combination is better than either agent separately ('Combo best') and that they are synergistic: 'Combo' better than 'Sum'.

For a detailed description of the study and statistical modelling see REFS 20,21.
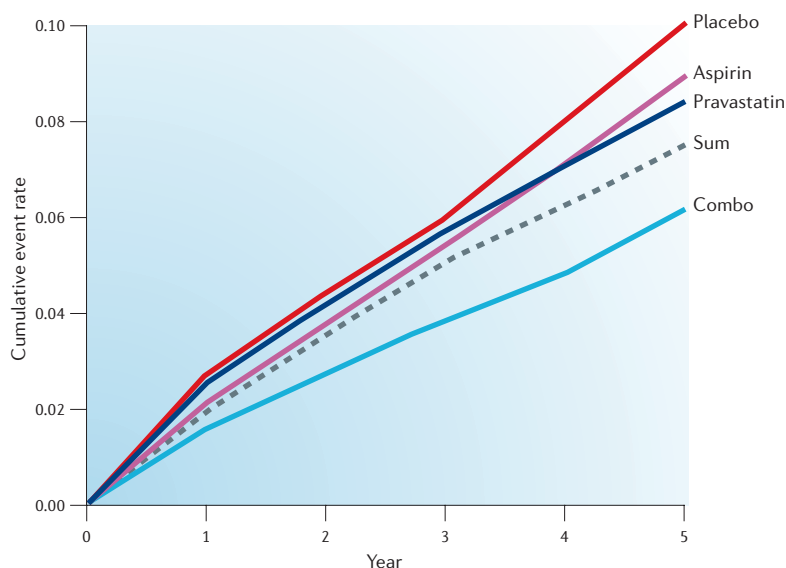
Probabilities that the combination is better than either individual drug and also of synergism

Comparison	Year 1	Year 2	Year 3	Year 4	Year 5
Combo best	0.968	0.993	0.999	0.999	0.999
Synergism	0.763	0.877	0.954	0.958	0.933

experiment become known, then they are taken as given and no longer subject to probabilities. In the example, probabilities regarding p are calculated conditionally on having observed 7 successes and 3 failures.

The mathematics of the Bayesian approach are quite simple and have their roots in Bayes rule, which I will describe next.

Bayes rule. An example use of Bayes rule that will be familiar to some readers is in the context of a diagnostic test for a disease. Suppose a test result is positive (+). Of interest is the probability that the patient has the disease in question (*dis*) given the results of the test, written Pr(*dis*|+). This is called the test's 'positive predictive value.' It is also called a posterior probability — 'after' the test. This probability cannot usually be found directly. It is related to standard characteristics of the test called sensitivity (*sens*), Pr(+|*dis*), and specificity (*spec*), Pr(–|*not_dis*). Both quantities have the test result on the opposite side of the vertical bar from the posterior probability. This inversion is a clear indication that Bayes rule applies. Indeed, Bayes rule (Equation 1) is sometimes called the rule of inverse probabilities[1]:

$$\Pr(dis|+) = \frac{sens \times prev}{\Pr(+)} \quad (1)$$

where *prev* is the prevalence of the disease. The denominator Pr(+) can be expanded as shown in Equation 2:

$$\Pr(+) = sens \times prev + (1 - spec) \times (1 - prev) \quad (2)$$

Suppose that the sensitivity of the test is 80% and its specificity is 90%. And suppose that on the basis of a patient's characteristics the probability of disease is *prev* = 15%. Then Pr(*dis*|+) is calculated as shown in Equation 3:

$$\Pr(dis|+) = \frac{0.80 \times 0.15}{(0.80 \times 0.15 + 0.10 \times 0.85)} = 0.585 \quad (3)$$

The same type of Bayesian calculation applies in any experimental setting. The results of experiments are used to update probabilities of parameters. Just as the diagnostic setting requires probabilities of results assuming both disease and no disease, no Bayesian calculation can be made on the basis of probabilities of observed results for a single value of a parameter. In addition, Bayesian calculations require analogues of disease prevalence: prior probabilities of parameters (see BOX 3 for discussion of prior distributions).

In the example trial, consider two possible values of *p*: 0.35 and 0.70. FIGURE 1 shows the probabilities of the various possible results for these two values of *p*. But now, in contrast to the frequentist approach, only the probabilities of the observed results matter. (The unused probabilities are shown in lighter type in FIG. 1.) The probabilities of 7 successes in FIG. 1 are the analogues of sensitivity and specificity. If, *a priori*, the two possible values of *p* are equally likely: Pr(*p*=0.35) = Pr(*p*=0.70) = 0.50, then the posterior probability of 0.35 is the ratio 0.021/(0.021+0.267) = 0.073, which compares with the prior probability of 0.50.
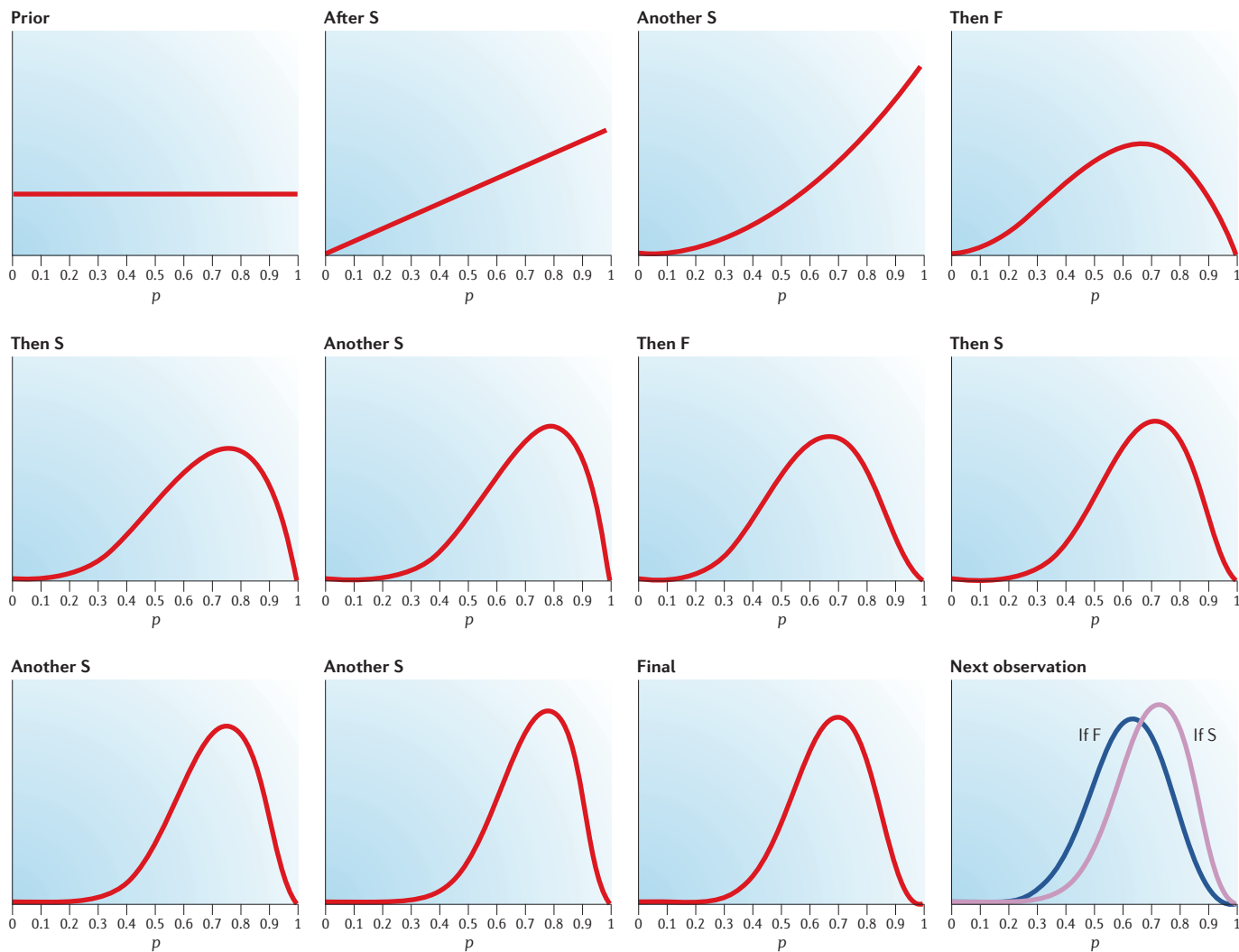
Figure 2 | **Sequence of probability distributions for success rate *p* corresponding to data SSFSSFSSSF.** The prior or initial distribution of *p* is in the upper-left corner. The distribution of *p* is updated with each new observation. The sequence in time proceeds from left to right and then down to the next row. Each S shifts the distribution to the right (larger values of *p* being therefore more likely) and each F shifts the distribution to the left. Shifts in distribution have a greater impact early in the trial and are less noticeable as the distribution becomes more concentrated. The panel labelled 'Final' shows the posterior (end of trial) distribution. The panel in the bottom-right corner shows the two possibilities for the distribution of *p* should a new patient, the eleventh one in all, enter the trial. That additional patient's result would be either success (purple curve) or failure (blue curve). As described in the text, the Bayesian (predictive) probability that the eleventh observation will be a success given the results of the first ten patients is (7+1)/(10+2) = 0.67 and the remaining 0.33 probability is for failure.

Stopping rule
Prespecified conditions that indicate what interim results would lead to stopping the trial. Possible decisions are to stop the trial's accrual of new patients or to announce the results of the trial. Even though a trial's accrual is complete and all patients have been treated, the trial's DSMB can keep the trial's results under wraps when so indicated by the stopping rule.

More realistically, *p* could be any value between 0 and 1. The distribution in the upper-left corner of FIG. 2 is an example of a distribution that is usually more realistic than one concentrated on just two possible values of *p*. For this flat (so-called 'non-informative') distribution, all intervals of *p* between 0 and 1 have probabilities equal to their widths. Bayes rule applies to update the distribution of *p* with each observation (although updating can take place as well for batches of observations, with the same result). Calculations are slightly more complicated than in the earlier setting[1] and are not shown. The subsequent panels in FIG. 2 show the updating process, moving from left to right. The relevant aspect of FIG. 2 is not the particular distributions, but that such updating is possible and that it is natural in the Bayesian approach. Continuous updating as information accrues distinguishes the Bayesian approach from frequentist approaches and leads to flexibility in conducting clinical trials and other experiments.

All Bayesian inferences are based on the current distribution of the unknown parameters and can be made at any time. After the tenth patient in the example, this distribution is the one labelled 'Final' in FIG. 2. An example calculation is the posterior probability that the experimental therapy does not improve the historical

success rate. This probability is 0.014, the area under the curve to the left of 0.35. After observing the first three patients' results, SSF, probabilities are calculated on the basis of the distribution in the upper-right corner. At that time the probability that the experimental therapy was not an improvement was 0.127.

Numerically, 0.014 is not very different from the P-values calculated earlier. But the two measures have very different interpretations. The frequentist P-value is difficult to understand (and even to say: 'the probability of observing a result as or more extreme than that observed assuming that the treatment is ineffective') whereas the Bayesian posterior probability is directly interpretable as the probability that the therapy is ineffective. Another difference is in the assumptions made. The P-value depends on the design of the trial and the intentions of the investigator, whereas the posterior probability depends on the prior distribution (BOX 3). For example, we saw that having equal probabilities on $p = 0.35$ and 0.70 gave a posterior probability of 0.073 that the therapy was not an improvement over the historical success rate. On the other hand, if the experimental therapy is a member of a class of drugs that all have a 35% success rate then the prior distribution may be concentrated near $p = 0.35$ and the posterior probability of $p \leq 0.35$ may be moderately large for the indicated data. In this case, it would take more than 10 observations with a 70% observed success rate to be persuasive that the drug is more effective than others in its class.

Had the trial been stopped before the tenth observation, or if an interim inference had been made at some time before the tenth observation, then the appropriate distribution is the corresponding one in FIG. 2. This aspect of the Bayesian approach is anathema in a frequentist perspective. For example, no P-value can be found after only three observations, say, because 'more extreme results' after three observations have not been identified and so their probability cannot be calculated. This distinction between the two approaches has received much attention in the literature[3–5].

Predictive probabilities and trial design
The Bayesian updating process has profound implications for trial design. Perhaps its most useful consequence is the ability to quantify what is going to happen in a trial from any point on (including from the start of the trial), given the currently available results. Future results cannot be predicted with certainty, of course, but the Bayesian approach allows for assessing the future with the appropriate amount of uncertainty.

The simplest continuation to consider is adding one more patient. In our example trial, will the treatment be successful or not for the next (eleventh) patient? There are two sources of uncertainty. First, even if p were precisely known, the observation cannot be predicted perfectly because there is a chance (p) that it will be a success and a complementary chance that it will be a failure. The other source of uncertainty is that p is itself not known precisely. It too has a probability

distribution, the one shown in the panel labelled 'Final' in FIG. 2. Calculating the probability that the next observation will be a success involves combining these two sources of uncertainty. The resulting probability and its calculation have a long history. The formula is known as 'Laplace's rule of succession': the number of successes plus 1 divided by the number of observations plus 2, or $(7+1)/(10+2) = 2/3$. This formula, which applies only for the flat prior distribution (BOX 3) shown in FIG. 2, also gives the current mean of the distribution of p[1].

The panel in the bottom-right corner of FIG. 2 shows the two candidate distributions for p after an eleventh patient's result is observed. The probability that it becomes the pink curve (success) is 2/3 and the probability it becomes the blue curve (failure) is 1/3.

More generally, the predictive probabilities of the final results of any trial can be found for any design. These probabilities are crucial in deciding which course a trial should take. They are also crucial at a trial's planning stage in assessing the value of any particular design. Each possible outcome of the trial has consequences and associated *utilities*, including costs[6]. These can be weighted by each outcome's probability to give an overall utility of the trial. Various candidate designs can be compared on the basis of their utilities.

Bayesian predictive probabilities are helpful in monitoring clinical trials. An example is a Phase II neoadjuvant HER2/*neu*-positive breast cancer trial conducted at M. D. Anderson Cancer Center[7]. Target accrual was 164 patients randomized to two treatment arms, chemotherapy with and without trastuzumab (Herceptin; Genentech). The primary endpoint was pathological complete response (pCR) of the tumour. Accrual was slower than expected, averaging about 1.5 patients per month. The trial was designed from a frequentist perspective and the protocol specified no interim analyses. The institution's Data and Safety Monitoring Board (DSMB) assessed the available results when 34 patients had data available for assessing pCR. The results were consistent with conclusions from much larger trials in metastatic breast cancer: the trastuzumab arm showed a dramatic improvement: 4 of 16 control patients (25%) and 12 of 18 trastuzumab patients (67%) experienced a pCR. The DSMB requested the Bayesian predictive probability of (standard frequentist) statistical significance when 164 patients had been treated. This probability is 95%. Armed with such compelling evidence regarding the trial's eventual conclusion, and in view of the questionable ethics of continuing to randomize patients in this Phase II trial, the DSMB overrode the protocol and stopped the trial. This override shows that Bayesian analysis can be legitimately used in conjunction with a frequentist design.

Choosing sample size using decision analysis
Some Bayesian and frequentist approaches to choosing the sample size of a clinical trial give the same answer[8]. But fully exploiting the Bayesian approach's ability to consider the trial's consequences can give very different

sample sizes. This is so even when the number of patients must be specified in advance and cannot depend on the results of the trial.

Suppose the goal is to treat patients as effectively as possible — all patients, those in the trial as well as those who come later and who will benefit from the results of the trial. The prevalence of the disease or condition in question should be considered explicitly. With a goal of treating as many patients as effectively as possible, clinical trials should be larger for conditions that affect millions of people than for those that affect a few hundred people per year. The conventional approach is the same. For example, Cheng et al.[9] consider a number of settings in which the optimal trial sample size has order of magnitude the square root of prevalence. So if sample size $n = 1,000$ is best for a trial involving therapies that will affect a million patients who have the disease in question, then $n = 32$ is best when the disease prevalence is 1,000. In cases for which prevalence is not precisely known, calculations can use rough estimates of its size. The availability of other therapies and uncertainties regarding the possible future availability of new therapies that will render the current therapies obsolete can be considered explicitly.

Current regulatory approval processes and journal publication policies ignore crucial issues such as prevalence of disease. For example, they apply the same standards of statistical significance level (using 0.05). Such an attitude is not consistent with delivering good medicine. Of course, regulators are aware of prevalence when making marketing approval decisions. But they are able to incorporate such information only informally, which is necessarily somewhat crude.

There is a Bayesian approach that — when possible — is better than having a fixed sample size. This is stopping the trial on the basis of the accumulating results when the answers to the scientific/clinical questions are known sufficiently well for applying the results to the broader patient population. This is one instance of the subject of the next section: adaptive trial designs.

Adaptive trial designs

The continuous learning that is possible in the Bayesian approach enables investigators to modify trials in midcourse. Modifications include stopping the trial, adaptively assigning patients to therapies that are performing better or that will give more information about the scientific question of interest, adding and dropping treatment arms, and extending accrual beyond that originally targeted when the answer to the question posed is not satisfactorily known.

I will consider two examples. One is a trial in acute myeloid leukaemia[10]. The experimental agent troxacitabine (T) was combined in turn with standard therapies idarubicin (I) and cytarabine (A) and compared with the two standard therapies in combination: TI versus TA versus IA. The maximal trial sample size was 75 patients, and the endpoint was complete remission (CR) within 50 days of initial treatment. Bayesian

probabilities of treatment comparisons were calculated continually. Patients entering the trial were assigned to therapy randomly, but imbalanced so as to favour therapies that had higher probabilities of being better. If a therapy's assignment probability was sufficiently low, then it was to be dropped from the trial, and the trial would stop before accruing 75 patients if only one therapy remained. In the actual trial, therapy TI dropped first, after the twenty-fourth patient had accrued, with no CRs out of its five patients. Therapy TA was dropped, and the trial ended, after the thirty-fourth patient had accrued, with three CRs out of its eleven patients. Standard therapy IA ended with 10 CRs (56%) out of 18 patients, which was consistent with its historical rate.

For these data and assuming flat priors for all three CR rates, the posterior probability that TA is an improvement over IA is 7.8% and that TI is better than IA is only 1.7%. Moreover, if either experimental combination is an improvement on IA, it is very unlikely to be much better. This example shows that, first, it is possible to learn from small samples, depending on the results, and second, that it is possible to adapt to what is learned to enable better treatment of patients.

The adaptive design used in this trial is not ideal, both from the perspective of treating patients effectively and of getting as much information as possible about therapy comparisons. It is a delicate compromise in both senses. And it makes the standard design of assigning the same number (such as 25) patients to each of the three therapies seem unethical in comparison.

For an adaptive design of a rather different type, consider a dose-finding trial. In a standard design, patients are allocated to a fixed number of doses in a grid. When the results become known, the investigators usually regret not having assigned patients in some other fashion. Perhaps the dose–response curve seems to be shifted to the left or right from that anticipated. If so, then assignment of patients on one end or the other of the dose range was wasted. Or perhaps the slope of the dose–response curve seems to be steeper than anticipated in a narrow interval. In this case the patients assigned to the flat regions of the curve would have been more informative had they been assigned doses in the region with steeper slope. Or perhaps early results made it clear that the dose–response curve was flat and that the trial could and should have been stopped earlier. Or the results of the trial could indicate that the standard deviation of response is greater or less than anticipated and so the trial should have been larger or smaller.

A better strategy, regardless of the results, is to proceed adaptively, analysing the data as it accumulates. In a trial investigating the dose–response of a neuroprotective agent for treating stroke[11–13], we used two stages, first dose-ranging and then confirmatory, if the latter was warranted. The dose-ranging stage (Phase II) continued until a decision was made that the drug was not sufficiently effective to pursue future development, or that the optimal dose for the confirmatory

Box 3 | **What prior?**

Some readers will be impressed by the elegance of the Bayesian approach because it embodies a generally accepted view of the scientific method: make an observation, update what is known, and decide what experiment is most informative and cost-effective to do next. Such readers might wonder why anyone takes a frequentist approach. Historically, a major reason was the Bayesian requirement for a prior distribution. Finding a distribution of a parameter posterior to an experiment without specifying a prior distribution for that parameter is not logically possible. The analogous statement in the diagnostic setting is that it is not possible to find the probability of having a disease based on test results without specifying the disease's prevalence.

Disease prevalence based on a patient's characteristics is not usually controversial. Assessing treatment effectiveness at the start of a clinical trial is another matter. For the prior distribution in FIG. 1 (upper-left panel), there is 65% probability that p is greater than 0.35. What is the basis for this assumption? There may not be a relevant historical database. And using animal data is problematic. A prior distribution reflects the information about the treatment in question separate from the experimental results. This information includes the investigator's understanding of the biology of the disease, historical results for the investigational and related treatments, and preclinical results for these treatments.

Whatever the form and substance of this information, a prior distribution can be assessed for any investigator. The rub is that prior distributions are specific to the investigator and might not be accepted by anyone else. So prior distributions are inherently subjective. Objectivity in science is generally elusive at best[1,4]. Indeed, the frequentist approach is itself subjective in a number of ways, including the models assumed, the parameters and hypotheses considered, and the experimental designs used. For example, as indicated above, conclusions from data SSFSSFSSSF depend on the investigator's intentions. 'Silent subjectivities' such as these are dangerous in that they are difficult or impossible to make explicit. By contrast, subjectivity in prior distributions is explicit and open to examination (and critique) by all.

There are several approaches for overcoming concerns about the subjective nature of prior distributions. One is to consider a variety of prior distributions in attempting to approximate the posterior distribution held by all types of readers. Another is to publish experimental results with instructions for readers to calculate their own posterior distributions. In a regulatory setting, it is important for sponsors and regulators to agree in advance as to the prior distribution(s) that will be used. (The same is true for assessing utilities when we progress to formally considering the consequences of clinical trials.)

A common approach is to assume a prior distribution that is 'non-informative' or 'open-minded' in the sense that it has little influence on the posterior distribution[1]. An example is the flat prior distribution in the upper-left panel of FIG. 2. For such a distribution, the results of the trial carry essentially all the influence in the posterior distribution. Moreover, when the trial is at least moderate in size, prior distributions that are not too different will give similar posterior distributions; in particular, all prior distributions that are reasonably flat near parameter values that are likely on the basis of the trial results will give nearly the same answers as the prior distribution shown in FIG. 1[22].

The approaches described above have been reasonably successful. Regulators are appropriately concerned about the choice of prior, but this no longer seems to be a stumbling block to using a Bayesian approach. And the same is true in medical research more generally.

---

stage (Phase III) was sufficiently well known. Switches from Phase II to III can be effected seamlessly, without stopping accrual.

Accrual in the stroke trial began in November 2000 and ended in October 2001. Information about the dose–response curve was updated continuously. Each entering patient was assigned a dose (one of 16, including placebo) that maximized information about the dose–response curve, given the results observed so far. Neither the number of patients assigned to any particular dose nor the total number of patients assigned was fixed in advance. The endpoint was improvement in stroke scale from baseline to 13 weeks. To gather information about dose–effect more rapidly, each patient's stroke scale was assessed weekly. We incorporated a longitudinal model of patient performance and carried out Bayesian predictions of the 13-week endpoint on the basis of available patient-specific information, and we updated probability distributions of dose–effect accordingly.

An adaptive design is more effective than standard designs at identifying 'the right dose'. And it usually identifies the right dose with a smaller sample size.

Another advantage is that many more doses can be considered in an adaptive design, even though some may be little used or even never used.

In the actual trial[14], the adaptive assignment algorithm was a great success, but the drug was not. The algorithm searched among the 15 positive doses and found nothing, finally focusing on assignments to the highest dose and placebo. The algorithm favoured stopping the trial for futility very early, but the sponsor had set a moderately large minimum sample size. The trial's DSMB accepted the algorithm's recommendation to stop the trial at this minimum, and there was no need for a confirmatory stage, seamless or otherwise.

A frequentist twist

The flexibility of the Bayesian approach can lead to complicated trial designs. Making many decisions during a trial's course can increase the rate of making an erroneous decision. Institutional review boards and others involved in clinical research, including regulators when the trial is for drug or medical device registration, require knowing the trial design's operating characteristics. These include false-positive rate and power (the

---

Placebo controls
Patients in a concurrent control group who are given a treatment that is indistinguishable from the experimental drug but which is inert — sometimes called a 'sugar pill'. Patients are 'blinded' or 'masked' as to the treatment they have been assigned. If clinicians are also masked as to the patients' assignments then the trial is 'double-blind.'

probability of concluding a benefit when there is actually a benefit), average total sample size, average proportion of patients assigned to the various treatment arms, probability of identifying the most effective dose and so on. Moreover, these bodies can request modifications in the design so as to ensure that the operating characteristics meet conventional benchmarks, such as having no greater than a 5% false-positive rate.

For complicated designs, these calculations would not have been possible before the availability of high-speed computers (BOX 1). In the modern era, they are straightforward using computer simulations. Treatment effects are specified in any particular computer run. For example, when assessing false-positive rate the experimental and control treatments are assigned the same values of the treatment efficacy parameters. 'Patients' are generated in accordance with the trial design to receive the therapy indicated by the design. These 'patients' respond according to the prespecified parameters and have the appropriate variability. When the trial stops, its result (advantage of experimental over control or not) is recorded. Other characteristics of interest — such as trial sample size — are also recorded. This process can be repeated many times. The proportion of the simulations in which the trial claims a benefit for the experimental therapy is the positivity rate. A histogram showing the simulated sample sizes is the distribution of sample size in the case assumed (such as the null case when there is no difference in treatments).

One can use the Bayesian approach to build a design and modify it to deliver predetermined frequentist characteristics, such as 5% false-positive rate and 90% power at a particular difference in treatment effects. The design is essentially frequentist, and the Bayesian has, in effect, become a frequentist. Though the process puts restrictions on the Bayesian's flexibility to update, the Bayesian approach served as a tool to build a frequentist design having good properties, such as small average sample size, fewer participants in the trial assigned to ineffective therapy and so on, with a consequent benefit for medical research.

Historical and other related information

In analysing the results of a clinical trial, the Bayesian attitude is to bring all available information to bear on the scientific question being addressed. Outside of a Bayesian perspective, such potentially important information is usually overlooked because the methodology used cannot incorporate it. Consider a randomized comparison of an experimental drug E and a control C, with survival as the primary endpoint. One type of information that is overlooked is patient-specific outcomes that might be correlated with survival. This is the subject of the next section. The present section considers relevant information that is available outside the trial.

Any particular trial is unlikely to be the first one conducted in the disease in question. Other trials might have considered therapy C, either as control or an experimental agent. In addition, databases of patients

with the disease are usually available. These sources of information should be exploited in analysing the results of the current trial, and the Bayesian approach provides a means for doing so. Patients in earlier trials might be different from those in the current trial. Therefore, patients in previous trials cannot be regarded as exchangeable with patients having the same treatment regimen in the present trial. This setting is ideally suited for a hierarchical Bayesian analysis[6,15–17].

In a hierarchical analysis, there are multiple levels of experimental units. When combining results of different trials there are two levels: patients within trials and the trials themselves. The population of trials has unknown characteristics, just as in a typical statistical problem. We have a sample from this population, numbering as few as two. The inferential problem is different from usual because the experimental units (trials) in the sample are not directly observable. Rather, we observe a sample nested within the sample. Patients provide partial information about the trials for which they represent and therefore they provide some information about the characteristics of the population of trials. This connection is a mechanism for borrowing information across trials.

The extent of borrowing is not dictated in advance, but instead is determined by the degree of concordance in the results of the various trials. Consider two trials. The first one consists of patients treated with C, which serves as a control for E in the second trial. If the results of control patients in the two trials differ greatly, then this suggests heterogeneity in the population of trials and so there will be little borrowing of historical information. However, if the results of control patients are similar in the two trials then this suggests homogeneity and enables greater borrowing. In a trial with a sample size fixed in advance, the number of concurrent controls can be reduced (but not to zero) to exploit this borrowing. It is better to have an adaptive design that enables the extent of borrowing to be assessed, and the proportion of patients assigned to control (and the overall trial size) determined, in an on-line fashion.

Bayesian hierarchical modelling has many applications in clinical trials. Consider cancer. Many drugs that are effective in breast cancer work in other solid tumours as well. The tradition of oncology drug development is one cancer at a time. But it would be better to include patients from a variety of cancers in a single trial to assess activity across diseases. One level of experimental unit in a hierarchical model is cancer type and another is patient within cancer type. And if more than one trial is involved, 'trial' can be included as still a third level in the hierarchy. It is also possible (and important) to model the potential roles of biomarkers that might be predictive of therapeutic benefit across diseases.

Still another level of hierarchy is especially important to regulators and drug developers: class of drug. Drugs in the same class may have similar effects, or not. Hierarchical Bayesian modelling allows for both possibilities. Borrowing results across trials of drugs in

the same class — to the extent determined by the data — can make for more informed decision making and smaller clinical trials. Other applications of hierarchical modelling are to drug safety and handling high-dimensional data, such as that from microarrays[18].

Biomarkers and auxiliary variables

Standard analyses of clinical trials compare the distributions of the primary endpoint in the treatment groups, perhaps adjusting for baseline differences in patient characteristics. For some endpoints, such as survival, not all patients experience the event in question during the trial. The ability to use accumulating results turns the Bayesian focus to information that is available on individual patients for help in comparing therapies. For example, in cancer trials, information is available about tumour response, disease progression, patient performance status and so on. Especially important (although usually ignored) is information about the relationships of these early variables with survival. The relationships might be different for different treatment groups and can be modelled as such. These are auxiliary variables or auxiliary endpoints, in that they enable more precise assessments of the primary endpoint.

Patients who are treated earlier in the trial contribute more information to understanding the relationships among auxiliary variables and long-term primary endpoints because these patients have longer follow-up times. Using information about these relationships in turn enhances the contribution of patients who are treated later because their early performance is utilized in drawing conclusions about their later (and unobserved) endpoints[16,19]. An additional and major benefit of modelling relationships between early and late endpoints is that it makes for stronger interim assessments of long-term endpoints and therefore improves the efficiency of adaptive designs.

Conclusions

The Bayesian approach has several advantages in drug development. One is the process of updating knowledge gradually rather than restricting revisions in study design to large, discrete steps measured in trials or phases. Another advantage of the Bayesian approach is that it is specifically tied to decision making, within a particular trial, within a drug development programme and within establishing a company's portfolio of drugs under development. Other advantages include the ability to use predictive probabilities and to build hierarchical models.

Bayesians can update at any time and without penalty. However, constructing a Bayesian design and then having to verify that its Type I error rate meets regulatory criteria exacts a penalty and loses part of the Bayesian advantage. However, it still has an advantage. Bayesian designs expand the frequentist envelope, even when accepting external constraints. Once Bayesian methods become more familiar to investigators and regulators, and once explicit decision-analytic criteria become commonplace in clinical trials, they will face fewer externally mandated restrictions.

This review has accentuated positive aspects of, and developments that result from, using Bayesian methods in clinical research. There have been disappointments and frustrations as well. Tradition has momentum and change is difficult. But the movement towards using Bayesian designs and analyses in clinical trials will continue, and at an accelerated pace. There is increasing demand from political bodies and consumer groups to make drug development more efficient, safer and yet faster. A danger is that we will abandon fundamental scientific principles. Using a Bayesian approach will lead to more rapid and more economical drug development without sacrificing good science.

Just as the Bayesian approach is used more in certain therapeutic areas of medical device development, the same will be true in drug development. To a large extent, this variability is due to personalities involved. But therapeutic areas in which the clinical endpoints are observed early obviously stand to benefit most. Diseases such as cancer in which there is a burgeoning number of biomarkers available for modelling the disease's progress will also benefit. These biomarkers will enable a patient's progress to be monitored more accurately and a more accurate assessment of the patient's outcome. The availability of early indicators of therapeutic benefit makes a therapeutic area ripe for Bayesian modelling.

In the immediate future, a barrier to incorporating Bayesian approaches more widely in drug development is the attitude of regulatory agencies such as the FDA. Even more important are pharmaceutical companies' frequently false perceptions of regulatory attitudes. However, regulators do not usually influence the designs of trials in Phases I or II and these are becoming increasingly Bayesian, especially in oncology. Moreover, strategic planning and portfolio management in some pharmaceutical companies is becoming increasingly Bayesian, including formal utility assessment and decision-making processes. Use leads to familiarity and to understanding. The advantages of the Bayesian approach in the various types of endeavours will become evident to policy makers and decision makers, and Bayesian methods will spill into other areas of drug development, from preclinical modelling to the design and analysis of Phase III clinical trials.

1. Berry, D. A. *Statistics: A Bayesian Perspective* (Duxbury, California, 1996). **Provides an elementary introduction to the Bayesian approach.**
2. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. *Am. Statistician* **59**, 121–126 (2005).
3. Berry, D. A. Interim analysis in clinical trials: the role of the likelihood principle. *Am. Statistician* **41**, 117–122 (1987).
4. Berger, J. O. & Berry, D. A. Statistical analysis and the illusion of objectivity. *Am. Scientist* **76**, 159–165 (1988).
5. Berger, J. O. & Wolpert, R. L. *The Likelihood Principle* (Institution of Mathematical Statistics, California, 1988).
6. Berry, D. A. & Stangl, D. K. *Bayesian Biostatistics* (Marcel Dekker, New York, 1996).
7. Buzdar, A. U. *et al*. Significantly higher pathological complete remission rate following neoadjuvant therapy

with trastuzumab, paclitaxel and epirubicin-containing chemotherapy: results of a randomized trial in human epidermal growth factor receptor 2-positive operable breast cancer. *J. Clin. Oncol.* **23**, 3676–3685 (2005).

8. Inoue, L. Y. T., Berry, D. A. & Parmigiani, G. Relationship between Bayesian and frequentist sample size determination. *Am. Statistician* **59**, 79–87 (2005).

9. Cheng, Y., Su, F. & Berry, D. A. Choosing sample size for a clinical trial using decision analysis. *Biometrika* **90**, 923–936 (2003).

10. Giles, F. J. *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.* **21**, 1722–1727 (2003).

11. Berry, D. A. *et al.* in *Case Studies in Bayesian Statistics V* (eds Gatsonis, C., Carlin, B. & Carriquiry, A.) 99–181 (Springer, New York, 2001).
    This chapter gives a detailed description of an adaptive dose-finding trial in stroke taking a Bayesian perspective. Doses are assigned based on accumulating data to maximize the efficiency of the trial. An additional innovation is allowing for proceeding seamlessly from Phase II to Phase III.

12. Malakoff, D. Statistics: Bayes offers a 'new' way to make sense of numbers. *Science* **286**, 1460–1464 (1999).

13. Farr-Jones, S. Better statistics. *BioCentury* **9**, 1–6 (2001).

14. Krams, M. *et al.* Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN). An adaptive dose–response study of UK-279, 276 in acute ischemic stroke. *Stroke* **34**, 2543–2548 (2003).

15. Berry, D. A. in *Cancer Medicine* 6th edn (eds Holland, J. *et al.*) 465–478 (Decker, London, 2003).

16. Berry, D. A. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Sci.* **19**, 175–187 (2004).

17. Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* (Wiley, Chichester, 2004).
    This is a comprehensive reference for Bayesian statistical methodology. It is an ideal guide for drug developers and medical researchers generally.

18. Berry, S. M. & Berry, D. A. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixed model. *Biometrics* **60,** 418–426 (2004).
    This is a methodology paper that applies Bayesian borrowing in an important aspect of drug development. The methodology is useful more generally, including incorporating biological knowledge about genetic pathways in the analysis of microarray data.

19. Inoue, L. Y. T., Thall, P. & Berry, D. A. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* **58**, 264–272 (2002).
    This paper demonstrates important advantages of the Bayesian approach in drug-trial design. It uses frequent interim analyses based on predictive probabilities and models relationships between the primary endpoint and early measures of a patient's performance.

20. Hennekens, C. H. *et al.* Additive benefits of pravastatin and aspirin to decrease risks of cardiovascular disease: Randomized and observational comparisons of secondary prevention trials and their meta-analysis. *Arch. Intern. Med.* **164**, 40–44 (2004).

21. Berry, S. M. *et al.* Bayesian survival analysis with nonproportional hazards: Metanalysis of pravastatin-aspirin. *J. Am. Statistical Assoc.* **99**, 36–44 (2004).

22. Edwards, W., Lindman, H. & Savage, L. J. Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242 (1963).

Competing intersts statement

The author declares competing financial interests: see Web version for details.