In Cancer Medicine, ed.6. pp 465-478. London: BC Decker. (Ed: Holland J, Frei T et al.)

33

Statistical Innovations in Cancer Research

Donald A. Berry, PhD

This chapter supplements and complements Chapter 32, “Theory and Practice of Clinical Trials,” by Marvin Zelen. In particular, understanding the basic principles espoused in Chapter 32 is a prerequisite for the present chapter. The two chapters support each other to a substantial degree, but are different in attitude. The purpose of the present chapter is to describe statistical approaches to cancer research that allow for building new designs and incorporating new analyses. Some of the methods described here have been introduced into research practice and others are still being developed. The goals of the innovations presented here are (1) to more effectively use patient resources while treating patients in clinical trials more effectively, and (2) to identify better drugs and other therapies more rapidly, moving drugs more quickly through the development process. The methods exploit available evidence and place information gleaned from an ongoing clinical trial into the context of what is already known. These new methods tend to be intuitively appealing. But like most innovations, some are controversial. Although they presage the future of clinical trial research, in oncology and more generally, not every method described here is destined to become standard in medical research.

This chapter addresses two overall categories of innovations. One represents a fine-tuning of the traditional practice of statistics. The other is based on an alternative view of the foundations of statistics. Separating the two categories is not possible. I will set each method in the context of statistical approach, but I will present the various methods in an integrated fashion.

The “alternative view” of the foundations of statistics is the Bayesian approach. Since not all readers will be familiar with this approach, I will describe it and relate it to the more traditional frequentist approach. Readers who are familiar with Bayesian ideas may wish to skip “Bayesian Updating” below. An important distinction between the two approaches is one of attitude. The Bayesian approach is ideal for on-line learning (as data accrue), and the frequentist approach is tied to a particular experimental design. But the two approaches support each other. For example, much of this chapter’s development of clinical trial design employs the Bayesian approach as a tool for finding designs that tend to treat patients in the clinical trial more effectively and that identify better drugs more rapidly. But the design thus derived is checked for its frequentist properties (such as false-positive rate and power). Ensuring that a design has pre-specified frequentist properties means that the design is frequentist and that the Bayesian approach is a tool for finding good frequentist designs.

Expanding the horizons of statistical designs and analyses to the extent described here relies on the availability of high-speed computers and sophisticated computational methods. In the past ten years there has been an explosion of Bayesian computational procedures that can be used to derive efficient designs. In addition, high-speed computers can be used to simulate trials having these designs to evaluate and compare their properties, such as power and false positive rate.

BAYESIAN UPDATING

The purpose of this section is to describe the Bayesian approach and to relate it to the more traditional frequentist approach. Any such introduction is necessarily cursory. Suggestions for further reading include a comprehensive but elementary introduction to Bayesian ideas and methods,[1] a discussion of their role in medical research,[2] and a text describing more advanced Bayesian methods.[3]

The defining characteristic of any statistical approach is how it deals with uncertainty. In the Bayesian approach, uncertainty is measured by probability. Any event that is unknown has a probability. The frequentist approach uses probabilities as well, but in a more restricted fashion, as will be seen in the next section. Examples of probabilities in the Bayesian but not in the frequentist approach: The probability that the drug is effective, and the probability that Ms Smith will respond to a particular chemotherapy.

The Bayesian paradigm is one of learning. As information becomes available one updates what one knows. The fundamental tool for learning under uncertainty is Bayes’ rule. Bayes’ rule relates inverse probabilities. A familiar example is finding the *positive predictive value* (PPV) for a diagnostic test: In view of a positive test result, what is the probability that the individual has the disease in question? The inverse probability is that of a positive test given the presence of the disease, which is called the test’s *sensitivity*. PPV also depends on the test’s *specificity*, which is the probability of a negative test, given that the individual does not have the disease. And it depends on the prevalence of disease in the population. The analog of PPV in the application of Bayes’ rule to statistical inference is the “posterior probability” that a hypothesis is true given experimental results. The analog of disease prevalence is the “prior probability” that the hypothesis is true.

For a simple numerical example, consider two very specific hypotheses: the alternative hypothesis $H1$ in which a population success rate r is 0.75 and a null hypothesis $H0$ in which r is 0.5. Viewing “success” as a positive test result, $H1$ is analogous to “has disease” and $H0$ is “does not have disease.” The “sensitivity of the test” is 0.75—the rate of a success under $H1$—and the “test’s specificity” is 0.5—the rate of failure under $H0$. Suppose the prior probabilities of $H1$ and $H0$ are both 50%: $P(H1) = P(H0) = 0.5$.

Convention is to write conditional probabilities using a vertical bar. The event following the bar is given—that is, taken as known with certainty—in calculating the probability of the event that appears before the bar. For example, the probability of success assuming hypothesis $H1$ is written $P(\text{success}|H1)$.

After observing a success, according to Bayes’ rule the updated (posterior) probability of $H1$ is

$$P(H1|\text{success}) = P(\text{success}|H1)\mathrm{P}(H1)/P(\text{success})$$

where the denominator follows from the law of total probability:

$$P(\text{success}) = P(\text{success}|H1)P(H1) + P(\text{success}|H0)P(H0).$$

For further explanation, see chapter 5 of *Statistics: A Bayesian Perspective*.[1] The success rate r is 0.75 under $H1$, and $r = 0.5$ under $H0$. Therefore

$$P(\text{success}) = (0.75)(0.5) + (0.5)(0.5) = 5/8.$$

This is the average of the two rates of success, 0.75 and 0.5, calculated with respect to the corresponding prior probabilities. So the posterior probability of $H1$ is

$$P(H1|\text{success}) = (0.75)(0.5)/(5/8) = 60\%.$$

The new evidence boosts the probability of $H1$ from 50% up to 60% (and since total probability must be 100%, it lowers the probability of $H0$ from 50% down to 40%).

Consider a second independent observation. The prior probability for this observation is that which is posterior to the previous observation. If this new observation is also a success, then a second use of Bayes’ rule gives $P(H1|\text{success, success}) = 9/13 = 69\%$. On the other hand, had the second observation been a failure then $P(H1|\text{success, failure}) = 3/7 = 43\%$. This process can go on indefinitely, updating either continually as observations are made or all at once. The current probabilities of the various possible values of success rate r can be found at any time. These probabilities depend on the original prior probability and on the intervening data. This process of updating and on-line learning is the advantage of using the Bayesian approach in clinical trials.

Bayes' rule is generally accepted as appropriate for finding the positive predictive value of a diagnostic test. What is controversial is whether Bayes' rule should be used to find probabilities of hypotheses—such as the hypothesis that a therapy is effective—on the basis of evidence from an experiment. Such probabilities have wide appeal. The sticking point for some is that they cannot be found without explicitly considering the corresponding prior probabilities. One must assess or otherwise come up with the probability that a therapy is effective in the absence of (or prior to) the evidence from the current experiment. Inevitably, therefore, prior probabilities have a subjective component.[1] Explicit subjectivity in science is objectionable to some, but others view the learning process as inescapably subjective.[4, 5]

The ability to include prior information in making inferences is a major benefit of the Bayesian approach. But collecting and assessing information is work, and probability assessment requires care. I will address assessing prior probabilities. But in much of the remainder of this chapter I will take prior probabilities to be given. Typically, I assume prior probabilities that are non-informative or open-minded [1] in the sense that all competing hypotheses are assigned the same prior probabilities. The assignment in the example above, $P(H1) = P(H0) = 0.5$, is non-informative in this sense. In general, it is a good idea to consider different prior probabilities and examine how the posterior probabilities change—a sensitivity analysis. Also, when the Bayesian approach is used to find designs with good frequentist properties, the prior distribution can be viewed as an aspect of the design that can be varied to effect more desirable frequentist properties.

The Bayesian approach is subjective, with probabilities depending on the individual assessor. However, many probabilities vary little if at all from one individual to the next. For example, suppose you plan to toss the coin and observe whether it lands as "heads." Most people's probability of heads will be 1/2. That is to say, given the choice between receiving some valued prize should heads or tails obtain, most people would be indifferent. Put another way, they would prefer either over the other if its prize was slightly more valuable.

However, there is an aspect of this assessment in which people will differ. Suppose everyone agrees on 1/2 and the coin is tossed and results in heads. The probability of heads on the next toss will vary from one individual to another (although it will not be less than 1/2 for anyone). One person may accept that the coin is fair and no amount of evidence would change that opinion. Another may suspect that you have a coin that is weighted to one side or the other. That person may have no particular reason to choose heads over tails, and so has probability 1/2. However, the latter person learns while the former does not. Learning occurs in the Bayesian approach by formulating uncertainty about the parameters in question, in this case the probability of heads, call it r.

Figure 33-1 shows four candidate prior distributions in the left-hand panels. Even though these distributions reflect different types of information, in all four cases the mean of r is 1/2. Therefore (again by the law of total probability) the probability of heads is 1/2 in each case.

The person in case "a" of Figure 33-1 starts with an open mind concerning r, in that each possible value of r has the same probability (height of the curve). Such a distribution is non-informative in the sense defined above. After observing heads on a toss of the coin, and using Bayes rule, this person's probability distribution becomes the one in the right-hand panel of "a." The updated probability of heads is 2/3. Such an increase over 1/2 reflects person a's open mind. (Had the coin toss resulted in tails, then the new probability of heads would have been 1/3.)

The left-hand panels of cases b through d in Figure 33-1 indicate successively more information available *a priori*. Therefore, the observation (heads) changes the updated probability of heads (mean of the distribution in the right-hand panel) less in proceeding from top to bottom of the figure.

To make the above point about differential learning, I could have chosen any types of distributions for the left-hand panels in Figure 33-1. I chose these particular distributions to illustrate a separate point. Consider person a's prior distribution (left-hand panel of Figure 33-1a). Modify it by observing heads (right-hand panel of Figure 33-1a). Now suppose a second toss of the coin results in tails. The new distribution of r is that of person b (left-hand panel of Figure 33-1b). The same is true moving down the figure. So a person who starts with the distribution in the left-hand panel of Figure 33-1a and observes heads, tails, heads, tails, heads, tails, heads would move through the distributions in Figure 33-1 from left to right and top to bottom and end up in the right-hand panel of Figure 33-1d. (An implication is that the functional forms of the eight curves shown in Figure 33-1, moving left to right and top to bottom, are proportional to the following: 1, r, $r(1–r)$, $r^2(1–r)$, $r^2(1–r)^2$, $r^3(1–r)^2$, $r^3(1–r)^3$, $r^4(1–r)^3$.)

Assessing Degrees of Belief In *A Brief History of Time*,[6] physicist Stephen Hawking makes clear the view that science is subjective. He addresses the "climate of thought" in various eras. Regarding scientific theories he uses language such as: "It was generally accepted," "we now believe," "They believed" and "if you believe." Subjective probabilities quantify *degrees of belief*, with probability 1 meaning complete acceptance and no chance for anything else to be true, and probability 0 meaning complete rejection. Degree of belief between these extremes indicates the extent of one's uncertainty. Degree of belief depends on the person who has the belief (as well as on the event in question). This person could be any investigator or observer. The event in question is arbitrary and could be any of those referred to above, such as that one treatment is more effective than another.

To measure degree of belief requires a scale, just like any other measurement. For degrees of belief the scale is a *calibration experiment*. The assessor must be able to imagine an experiment with equally likely outcomes. Candidates are coin tosses, die rolls, selecting a chip from a bowl, etc. To decide whether outcomes are equally likely, suppose the assessor gets to choose any one of the possible outcomes. I promise to give the assessor a valuable reward should the experimental result be the outcome

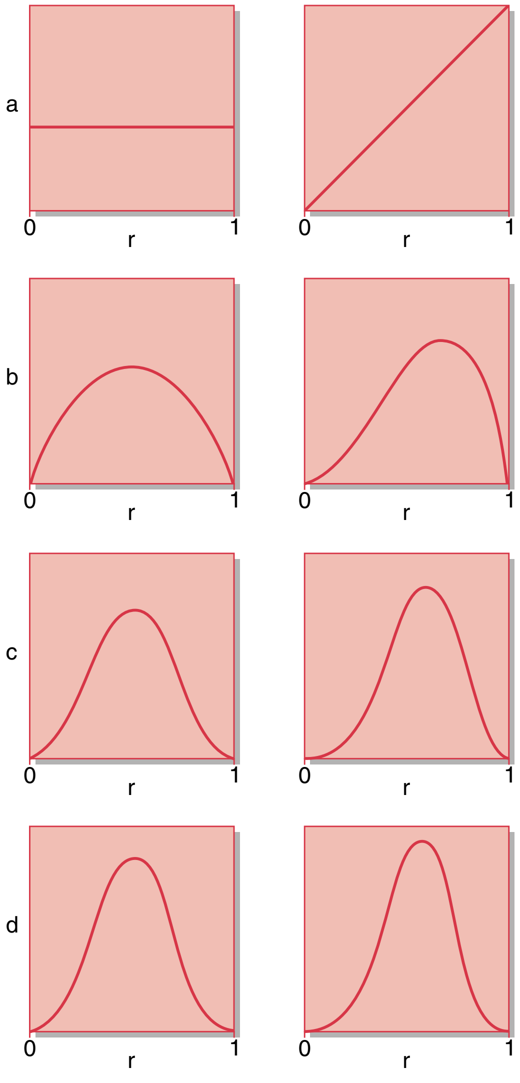

Figure 33-1 Left-hand panel in each pair is the prior distribution of rate r of heads in coin tossing. Right-hand panel is the posterior distribution of r after having observed heads on one toss of the coin. The probability of heads for each left-hand panel is 0.500, increasing to 0.667, 0.600, 0.571, 0.556 in cases "a" through "d", respectively, in right-hand panels. Changes are greater and learning more rapid when the prior distribution reflects greater uncertainty.

chosen. Outcomes in the set are equally likely for the assessor if the assessor is indifferent among them. In particular, the assessor would strictly prefer any one outcome over all others if I were to increase the reward on that one by an arbitrarily small amount.

Consider the problem of eliciting someone's probability that a drug has a response rate (in a particular population of patients) of at least 60%. Present that person with a choice between a reward should that event obtain and various outcomes of a calibration experiment. For example, observing a red chip in a selection from a bowl of chips containing at least two colors. Varying the proportion of red chips allows for homing in on the individual's probability of the event. For details, see sections 4.4 and 7.4 of *Statistics: A Bayesian Perspective*.[1] Considering alternatives to the response rate of 60% allows for finding the assessor's full distribution. Suppose it turns out to be that in Figure 33-1a, left-hand panel.

There are a variety of ways to check an assessor's probabilities. One is via prediction. Each assessment implies a learning rate. For example, the prior distribution in the left-hand panel of Figure 33-1a implies a probability of response for the first patient of 50%. It also implies that if the first patient responds then the probability of a response in the second patient is 0.667. The assessor should be told this and asked whether it corresponds to his or her opinion. It may not. If the updated probability of a response would be closer to 0.6, then the distribution in the left-hand panel of Figure 33-1b may more accurately reflect the assessor's opinions.

Prior distributions may be—and usually are—based on historical data. Suppose that a similar drug (or the same drug in a different patient population) gave a response rate of 50% in 20 patients: 10 responders and 10 non-responders. The corresponding likelihood of response rate r (see "Frequentist/Bayesian Comparison" below) is $r^{10}(1–r)^{10}$. It would not be reasonable to use this as a prior distribution for r, but it would be reasonable to exploit this information in some fashion. One possibility (another will be described in "Hazards over Time") is to discount the historical evidence. For example, counting the historical data at a proportion of 20% would mean using a prior distribution proportional to $r^2(1–r)^2$. This happens to be the distribution shown in Figure 33-1c, left-hand panel.

An advantage of basing prior distributions on historical information is that more observers are likely to have similar prior views, and therefore similar posterior views.

Robustness Principle An important *robustness principle* is that in the presence of at least a moderate amount of data, essentially all observers will have the same posterior distribution. That is, the particular prior distribution assumed does not matter much when the sample size is moderate to large. As an example, consider the eight distributions shown in Figure 33-1 and think of them as being the prior distributions of eight different people. Parameter r is response rate to a particular drug. Suppose that 40 patients were treated in a trial and there were 20 responders and 20 non-responders. Applying the robustness principle, the eight individuals in question will come to very nearly the same conclusion about response rate r. The eight posterior distributions are shown in Figure 33-2. These are very similar, as is evident from the figure. And the corresponding 95% probability intervals will also be very similar.

There are circumstances in which the robustness principle does not apply. All have to do with individuals who assign small probabilities with possible values of the parameter in question. As a special case, an assessor who dogmatically assigns prior probability 1 to a particular interval will also assign posterior probability 1 to that interval. Such an individual does not learn, except when restricted to that interval. In the example with 20 responders of 40 patients, someone who assigns 0 prior probability to $r < 0.5$ would have a posterior distribution similar to that in Figure 33-2, except that it would be 0 to the left of $r = 0.5$.

Frequentist/Bayesian Comparison In contrast to the Bayesian approach, the frequentist approach does not allow hypotheses to have probabilities. Rather, the approach restricts probability assignments to data, assuming particular values of the unknown parameters (or hypotheses) in calculating these probabilities. For example, a p-value is the probability of the observed data or more extreme data, assuming that the null hypothesis is true. In symbols:

$$\text{Frequentist } p\text{-value: } P(\text{observed or more extreme data}|H0)$$
$$\text{Bayesian posterior probability: } P(H0|\text{observed data})$$

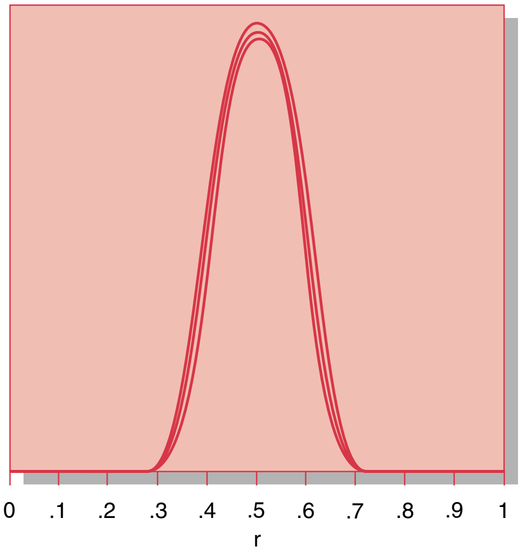

Figure 33-2 Posterior distributions for response rate r based on an experiment with 20 successes and 20 failures. The eight prior distributions considered are the eight distributions shown in Figure 33-1. Except for proportionality constants these are 1, r, $r(1–r)$, $r^2(1–r)$, $r^2(1–r)^2$, $r^3(1–r)^2$, $r^3(1–r)^3$, and $r^4(1–r)^3$. The corresponding posterior distributions are proportional to $r^{20}(1–r)^{20}$, $r^{21}(1–r)^{20}$, $r^{21}(1–r)^{21}$, $r^{22}(1–r)^{21}$, $r^{22}(1–r)^{22}$, $r^{23}(1–r)^{22}$, $r^{23}(1–r)^{23}$, and $r^{24}(1–r)^{23}$. These are very similar, demonstrating the robustness principle.

As an example, consider a single-arm Phase II trial for testing $H0$: $r = 0.5$ versus $H1$: $r = 0.75$. Assuming a type I error rate a = 5%, a sample size of $n = 33$ gives 90% power. Suppose there are 22 successes and 11 failures. The (frequentist) one-sided p-value is the probability of 22 or more successes of the 33 patients assuming the null hypothesis, $H0$: $r = 0.5$. Under this assumption the probability of observing 22, 23, 24, . . . successes is 0.0225+0.0108+0.0045+ . . . = 0.0401. Since this p-value is less than 5%, observing 22 successes is said to be "statistically significant."

The Bayesian measure is the posterior probability of the hypothesis that $r = 0.75$ (which is one minus the probability of $r = 0.5$) given 22 successes out of 33 trials. (As indicated above, the Bayesian calculation depends only on the probability of the data actually observed, 22 successes of 33, while the frequentist calculation also includes probabilities of 23, 24, etc. successes.) Using Bayes' rule:

$$P(H1|22\text{ of }33) = P(22\text{ of }33|H1)P(H1)/P(22\text{ of }33)$$

As above, the denominator follows from the law of total probability:

$$\begin{aligned} P(22\text{ of }33) &= P(22\text{ of }33|H1)P(H1) + P(22\text{ of }33|H0)P(H0) \\ &= (0.0823)(0.5) + (0.0225)(0.5) \\ &= 0.0524 \end{aligned}$$

So

$$P(H1|22\text{ of }33) = (0.0823)(0.5)/0.0524 = 0.785$$
$$P(H0|22\text{ of }33) = (0.0225)(0.5)/0.0524 = 0.215$$

The above calculation considers just two hypotheses, $r = 0.5$ and $r = 0.75$. In considering other values of r, Bayes' rule weighs them by $P(22\text{ of }33|r)$, which is called the *likelihood function* of r. The likelihood function is pictured in Figure 33-3. It indicates the degree of support for success rate r provided by the observed data. Values of r having the same likelihood are equally supported by the data. Only relative likelihoods matter. For example, conclusions about $r = 0.5$ versus 0.75 depend only on the ratio of their likelihoods 0.0823 and 0.0225, values that are highlighted in Figure 33-3. Since 0.0823/0.0225 = 3.66, the data lend 3.66 times as much support to $r = 0.75$ as compared with $r = 0.5$.

The conclusions of the two approaches are fundamentally different conceptually, and they are also different numerically. In the frequentist approach the results are statistically significant, with p-value = 0.0401. Such a small p-value is interpreted by some as being sufficient to conclude that $H0$ is not true. On the other hand, the Bayesian posterior probability of $H0$ is 0.215. This is decreased from the prior probability of 0.50, but it is more than 5 times larger than the p-value.

Interval estimates also have different interpretations in the two approaches. In the Bayesian

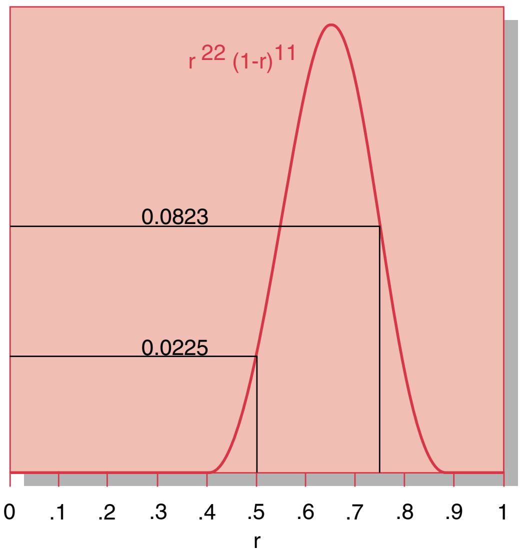

Figure 33-3 Likelihood of r for 22 successes out of 33 observations, $P(22 \text{ of } 33|r)$, which is proportional to $r^{22}(1–r)^{11}$. The likelihoods at $r = 0.5$ and 0.75 are highlighted. These values are used in the calculational example in the text.

approach, one can find the probability that a parameter lies in any given interval. In the frequentist approach, confidence intervals have a long-run frequency interpretation for fixed parameters. However, despite such very different interpretations, there is a point of agreement between the two approaches. Namely, if the prior distribution is non-informative then the Bayesian posterior probability of a confidence interval is essentially the same as the frequentist level of confidence. For example, if the prior distribution is non-informative then the Bayesian posterior probability that a parameter lies in its 95% confidence interval is in fact 95%. For other prior distributions the posterior probability of a 95% confidence interval may be greater than or less than 95%.

Predictive Probabilities The Bayesian approach allows for calculating the probability of data without having to condition on a particular parameter value; namely, one averages the conditional probabilities of the data over the various possible parameter values. This is advantageous for both monitoring trials and designing trials from the outset. Predictive probabilities will be exploited extensively in this chapter.

Consider an example based on the trial described above. Suppose that the first 16 patients have responded, with 13 successes and 3 failures. What will be the results after the full complement of 33 patients? It is impossible to say for certain, of course, but there is some information available for predicting this number. Conditioning on the information that is currently available allows for calculating probabilities for the future results in the Bayesian approach.

It might seem reasonable to estimate r to be $13/16 = 0.81$, the currently observed success proportion, and to calculate the probability of the results of the next 17 patients assuming this value of r. But this would be wrong. For one thing, we have restricted consideration to only two values of r, 0.5 and 0.75. But even if r were unrestricted and could be any value between 0 and 1, the information in a finite sample is not sufficient to say for certain that r has a particular value. The uncertainty in r is considered explicitly in a Bayesian approach.

Bayesian predictive probabilities incorporate two types of variability. One is the usual sampling variability that applies even if success rate r were perfectly known. (You don't always get the same result when you toss a fair coin.) The other uncertainty is in the success rate r. (You don't really know that a coin is fair even if you've tossed it many times.) Assuming 13 successes in the first 16 patients, the possible numbers (S) of successes after 33 patients are shown in the first column of Table 33-1. The second and third columns show the probabilities of the possible values of S for the two r's assumed in this example. The corresponding probabilities without conditioning on r are shown in the fourth column. This is a weighted average of the second and third columns. The weights are the respective probabilities of the two values of r conditional on having observed 13 successes in the first 16 patients: 0.039 for $r = 0.5$ and 0.961 for $r = 0.75$. The fourth column evinces greater variability (greater standard deviation) than either of the previous two columns. Typically, including when all values of r between 0 and 1 are considered (that is, all values have positive probability), predictive probabilities reflect greater uncertainty about future results than when conditioning on a particular value of r. (The last column of Table 33-1 will be discussed in the next section)

For convenience, in this example I assumed equal prior probabilities: $P(H1) = P(H0) = 0.5$. Although there is no vertical bar "|" in these expressions, these probabilities can depend on other available evidence, such as results of earlier clinical and pre-clinical trials. There may be additional information from biological assessments, such as when considering targeted therapies. These overall conditions are taken to be understood in setting down $P(H0)$ and $P(H1)$.

Bayesian versus Frequentist Interim Analyses There are numerous commonalties and a few differences between the Bayesian and frequentist approaches. This section addresses a principal difference. In the Bayesian approach, one makes an observation and updates the probabilities of the various hypotheses. This simple process implies a degree of flexibility that is difficult to mimic in the frequentist approach.

Consider the trial design described above, with $n = 33$ patients and testing $H0$: $r = 0.5$ versus $H1$: $r = 0.75$. Observing 22 or more successes will be sufficient to reject $H0$ in favor of $H1$. However, assigning 33 patients to an experimental therapy without assessing interim results is ethically problematic and would likely be questioned by institutional review boards. If the results are conclusive (either positive, strongly suggesting $r > 0.5$, or negative, suggesting $r \leq 0.5$) part of the way through the trial, then it should be stopped. Suppose, for example, that after 16 patients, 13 are successes and 3 are failures. From a Bayesian perspective, the updated probability of $H1$ is 96.1% (assuming prior probability $P(H0) = 0.5$).

Whether this probability is "conclusive" is not clear. The decision as to whether to continue a trial is complicated. It depends on the consequences of the trial given the current results and also given future results. In the Bayesian approach, consequences of future results can be weighed by their

Table 33-1 Predictive Probabilities of Number S of Successes after 33 Patients given 13 Successes in the First 16 Patients

S (of 33)	P(S|r = 0.5)	P(S|r = 0.75)	P(S|13/16)	P(H1|S/33)
13	0.0000	0.0000	0.0000	0.0002
14	0.0001	0.0000	0.0000	0.0006
15	0.0010	0.0000	0.0000	0.0017
16	0.0052	0.0000	0.0002	0.0050
17	0.0182	0.0000	0.0007	0.0148
18	0.0472	0.0001	0.0019	0.0432
19	0.0944	0.0005	0.0042	0.1192
20	0.1484	0.0025	0.0082	0.2887
21	0.1855	0.0093	0.0162	0.5491
22	0.1855	0.0279	0.0341	0.7851
23	0.1484	0.0668	0.0701	0.9164
24	0.0944	0.1276	0.1263	0.9705
25	0.0472	0.1914	0.1857	0.9900
26	0.0181	0.2209	0.2129	0.9966
27	0.0052	0.1893	0.1820	0.9989
28	0.0010	0.1136	0.1091	0.9996
29	0.0001	0.0426	0.0409	0.9999
30	0.0000	0.0075	0.0072	1.0000

Columns P(S|r = 0.5) and P(S|r = 0.75) assume the indicated value of r in calculating the probability. Column P(S|13/16) is the weighted average of the two previous columns, where the respective weights are 0.039 and 0.961. The last column gives P(H1|S/33), the probability of H1: r = 0.75, given S successes after 33 patients. The shaded cells are described in the text.

predictive probabilities. (see “Decision Analysis and Choosing Sample Size” below) For example, if the impact of the trial depends on whether the posterior probability of $H1$ is > 95% when the data from the full complement of 33 patients becomes available, then one can calculate the predictive probability of this event. The last column in Table 33-1 shows the posterior probability of $H1$ assuming S successes of 33 patients. The shaded values are those having $P(H1|S/16) > 95\%$. To achieve > 95% posterior probability requires at least 24 successes in the 33 patients. The predictive probability of this event is the sum of the predictive probabilities for $S \geq 24$ (the fourth column in Table 33-1), which is 0.8642. Although the current probability of $H1$ is > 95%, this characteristic will be lost with probability 1 – 0.8642 = 0.1358. That this has moderate probability indicates the tentative nature of the current conclusion. The possibility that the current conclusion is moderately likely to change can be factored into the decision to continue the trial.

Alternatively, and mixing Bayesian and frequentist concepts, if the impact of the trial depends on achieving (one-sided) statistical significance then the Bayesian predictive probability of this event means adding the probabilities of 22 and 23 successes to 0.8642, the total being 0.9684.

If the predictive probability that the current conclusion will be maintained is sufficiently high then one may reasonably decide to stop a trial. This is true for both claims of futility and superiority. The possibility of stopping a trial early on the basis of predictive probability should be stated explicitly in the trial’s protocol.

The focus of the frequentist perspective is the type I error rate, a. This is the probability of rejecting $H0$ when $H0$ is true, which depends on the trial design. For a fixed sample size of 33 patients, the calculation is straightforward. Rejecting $H0$ for ? 22 successes means a = 0.0401 (see previous section). The calculation becomes more complicated when there is a possibility of stopping the trial early. In the example, if the trial is stopped and $H0$ is rejected if there are 13 or more successes in the first 16 patients then a is increased because there is additional opportunity for rejecting $H0$. Assuming $r = 0.5$, the probability of rejecting $H0$ is now 0.0712. (The possibility that r is different from its null value plays no role in calculating the type I error rate.) Since this is greater than 0.05, the convention is to modify the stopping and rejection criteria to reduce a to about 0.05. For example, rejecting only if there are 14 successes or more out of 16 patients, or if there are 23 or more successes after 33 patients, gives an overall type I error rate of 0.0326.

It follows that it is more difficult to draw a conclusion of statistical significance when there are interim analyses. The reason is that the type I error rates are calculated assuming that a particular hypothesis (the null hypothesis of no effect) is true. In a sense an investigator is penalized for interim analyses in the frequentist approach. There are no such penalties for interim analyses in a Bayesian perspective. The reason is that Bayesian probabilities do not condition on any particular hypothesis.

Although it is not a Bayesian quantity, the type I error rate of any Bayesian design—however complicated—can be evaluated. If the design has interim analyses, then such a calculation incorporates appropriate penalties. This calculation is straightforward in a simple example such as that given above. In more complicated settings it can require Monte Carlo simulations. To find a via simulation in the above example, toss a fair coin 16 times. Make a tick mark if you get 13 or more heads and stop tossing. Otherwise toss the coin an additional 17 times and make a tick mark if the total number of heads is 22 or more. Repeat the process thousands of times. (Program a computer to do the tossing!) Estimate a to be the number of tick marks divided by the number of times you simulated the process. Assuming that your random number generator is working properly, you’ll find that the proportion with tick marks is about 7%.

ANALYSIS ISSUES

The purpose of this section is to consider two rather special analysis issues. The first is a natural extension of the previous section. The second is unrelated to the first and deals with a particular aspect of survival analysis.

Hierarchical Modeling: Synthesizing Information When analyzing data from a clinical trial, other information is usually available about the treatment under consideration. This section deals with a method for synthesizing information from a variety of sources. The method applies for incorporating historical information and for meta-analysis.

Suppose the Phase II trial discussed above gave 21 successes in 33 patients. The one-sided p-value is 0.08, and so the results are not statistically significant at the 5% level. The posterior probability of $H1$: $r = 0.75$ (in comparison to $H0$: $r = 0.5$) is only 54.9%, barely changed from its prior probability of 50%. However, in another trial using the same treatment there were 15 successes among 20 patients. This other trial may have been conducted at another institution or at an earlier time (perhaps Phase I, in which response to therapy was also assessed) at the same institution. In either case, it seems wrong to use a statistical procedure that ignores the information.

Bayes’ rule allows for combining the results of many trials, but there are pitfalls. From a frequentist perspective, one can assume that the data resulted from a single trial with a fixed sample size of 53 patients. The p-value for 36 successes among 53 patients is 0.0063, highly statistically significant. An analogous Bayesian assumption is that the same success rate r applied in both trials. The corresponding likelihood is shown in Figure 33-4. The ratio of likelihoods at $r = 0.75$ and $r = 0.5$ is 16.7, and so the former nets 16.7 times as much support as the latter. Assuming $P(H0) = P(H1) = 0.5$ as before, the corresponding updated probability of $H1$: $r = 0.75$ is 94.3%.

The naive frequentist analysis in the above paragraph is wrong because there were two trials and not one. And it is not clear how to repair that analysis. A variety of approaches have been suggested, but none is very satisfactory. The frequentist approach is experimental-specific and modifications of the experiment make conclusions difficult.

The above Bayesian analysis is also wrong because the success rate r may reasonably vary from one trial to another. The two success rates may be different even if the eligibility criteria in the two trials are the same. Because of differences in concomitant therapy and limitations in assessing prognostic variables, the two rates may even be different if the patients admitted have the same distribution of clinical characteristics. A way to repair the analysis is to explicitly consider two values of r, say r_1 for the first trial and r_2 for the second. These two r-values may be the same or different.

A hierarchical model is a *random-effects model*. One level of experimental unit is patient (within trial) and the second level is trial itself. Of interest is the population of trials and its distribution. The data give information about whether this population has little variability (homogeneity) or much variability (heterogeneity). In the former case the precision of parameter estimates will be greater than in the latter case because there will be greater “borrowing of strength” across the trials. Should it happen that the results of the trials vary greatly from one to the next then there will be little borrowing and the information from an individual trial will not apply much beyond that trial.

To understand the concept of a hierarchical model, think of the two r’s as coming from a population, one indexed by the possible trials. If this population is homogeneous then the two r’s are likely to be similar, and if it is heterogeneous then the two may well be quite different. Whether it is homogeneous or heterogeneous is not known. As

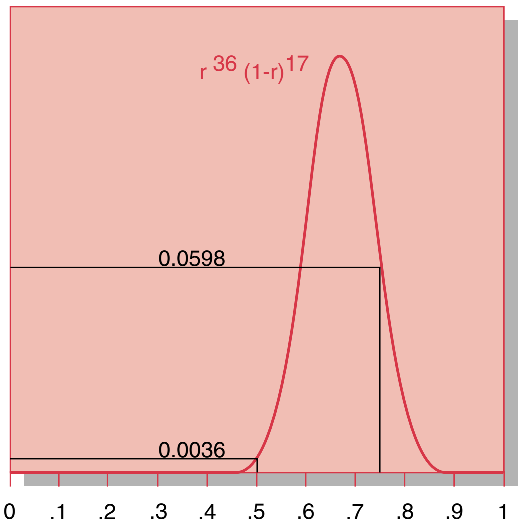

Figure 33-4 Likelihood of r for 36 successes out of 53 observations, $P(36 \text{ of } 53|r)$, which is proportional to $r^{36}(1–r)^{17}$.

usual in a Bayesian approach, unknowns have probability distributions. So the distribution of r-values itself has a distribution, one about which we have information from the two trials. Namely, we have 15 successes for 20 patients on r_1 and 21 successes for 33 patients on r_2. In a typical statistics problem, one makes observations from a population of interest. The observations are usually numerical. But in the present circumstance, instead of observing the number r_1 for trial 1, the observation is the likelihood function for r_1: in a sense, each possible value of r_1 is observed with weight given by its likelihood, with the total weight being 1. Similarly for trial 2 and r_2. The two likelihood functions are shown in Figure 33-5. The tightness in these curves conveys the uncertainty that is present in these two observations.

Although we want to incorporate the information from the first trial into our inferences, our principal focus is r_2. Bayesian calculations require a prior distribution. For illustrative purposes I will assume one that is especially simple. Continue to assume that the r's can be only 0.5 or 0.75. Further assume that either (1) all the r's in the population of trials are equal (prior probability 0.5) or (2) half of them equal 0.5 and the other half equal 0.75 (with the remaining prior probability of 0.5). In case (1), $r_2 = r_1$ and so the data can be simply pooled. In case (2) the data for each trial stand on their own. The prior weights of cases (1) and (2) are both 0.5, but these probabilities are to be updated in view of the results of the two trials.

Calculating the posterior probabilities of cases (1) and (2) is straightforward, but somewhat tedious. It turns out to be 0.864 for 1) and 0.136 for (2). Now the posterior probability of $H1$: r_2 is 0.75 is easy find. This calculation uses the above calculations that the probability of $r_2 = 0.75$ is 94.3% in the pooled analysis (1) and 54.9% in the separate analysis (2). Since we know the posterior probabilities of these two cases:

$$P(H1|\text{data}) = (0.864)(0.943) + (0.136)(0.549) = 0.889.$$

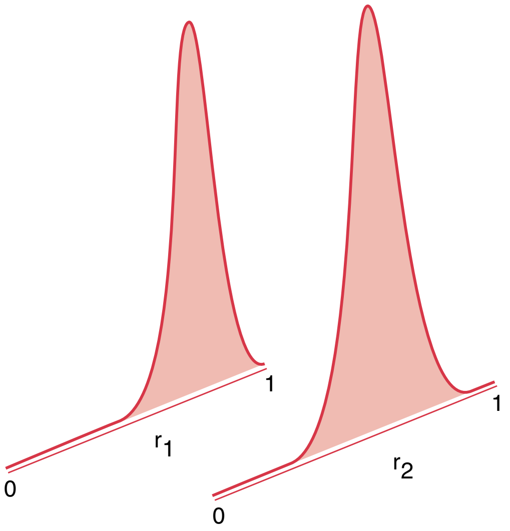

Figure 33-5 Likelihood of r_1 is $r_1^{15}(1–r_1)^5$ for trial 1. Likelihood of r_2 is $r_2^{21}(1–r_2)^{12}$ for trial 2. These two likelihoods represent a sample of size 2 from the population of likelihood functions, one for r_1 and the other for r_2. The perspective in the drawing is meant to suggest the r-dimension in this population. (For convenience, the two likelihoods are shown having the same area under the curve. However, these areas are irrelevant since inferences depend only on ratios of likelihoods within the same curve.) These two likelihoods contrast with the single likelihood shown in Figure 33-4, wherein r_1 and r_2 are assumed to be equal and r is their common value.

The probability of $H1$: r_2 = 0.75 is 88.9%, increased from that for the separate analysis but not as compelling as the pooled analysis.

More generally, there may be any number of related studies or databases that provide supportive information regarding a particular therapeutic effect. The studies may be heterogeneous and may consider different patient populations. The next example is generic but it is more complicated than the previous example because it includes nine studies.[7] The only commonalty in the studies is that all addressed the efficacy of the same therapy.

This setting is more realistic than that of the previous example because the success rates can take on any value between 0 and 1. The number S of successes and sample size n is shown for each study in Table 33-2 and in Figure 33-6. There are nine true success rates, one for each study (of which the sample success proportions S/n are estimates). Assuming the same success rate r applies in all nine studies, and pooling the data accordingly, there were 106 successes among 150 patients. The posterior distribution of success rate r (assuming a non-informative prior distribution) is labeled "pooled analysis" in Figure 33-6.

It is questionable whether one should ever assume that different studies have the same success rate. In this example such an assumption is especially suspect. The nine observed success proportions evince more variability than is consistent with the possibility of equal success rates. Therefore, assuming homogeneity and pooling the results of the nine studies seem inappropriate. The "hierarchical analysis" curve in Figure 33-6 is a Bayesian estimate of the distribution of success rates in the population of studies. (This curve is the mean posterior distribution assuming a non-informative prior on a particular class of distributions, called *beta distributions*.) As is typical of hierarchical analyses, this curve shows greater variability than does the analog assuming homogeneity.

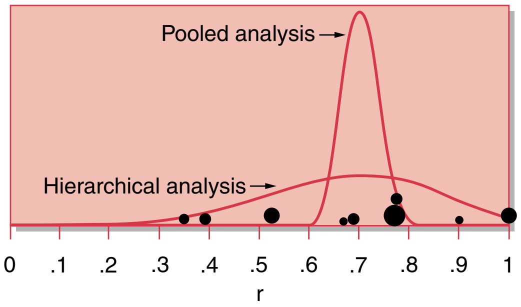

Figure 33-6 The dot plot on the r-axis shows the observed success proportions given in Table 33-2. The areas of the dots are approximately proportional to sample sizes n. The pooled analysis curve shows the distribution of success rate r assuming no study effect. The hierarchical analysis curve shows the Bayesian estimate of the distribution of success rates allowing for heterogeneity across the various studies.

In a hierarchical analysis, an individual study's success rate has a distribution that depends on the data from that study, but it also depends on the data from the other studies. The last column of Table 33-2 shows the mean of the distribution of each study's true success rate. This is also the predictive probability of success for a future patient in that study. The individual study probabilities are shrunk toward the overall mean. This shrinkage is greater for smaller studies, and for studies with observed proportions further from the overall mean.

Figure 33-7 provides a pictorial comparison of the rightmost two columns in Table 33-2, demonstrating shrinkage. The Bayesian estimates are intermediate between simple pooling (complete shrinkage) and each trial standing alone. The amount of shrinkage—including the above two extremes—depends on the prior distribution of the population of trials. This aspect of the prior distribution should be set in advance, or varied to allow for assessing the sensitivity of the overall conclusion.

Table 33-2 Numbers of Successes S, Sample Size n, Observed Success Proportions (including its standard error) and Adjusted Estimates of Success Rates by Study

Study	Successes, S	Sample size, n	Success prop. (standard error)	Bayes estimate (standard dev)
1	11	16	0.69 (0.116)	0.69 (0.094)
2	20	20	1.00 (0.000)	0.90 (0.064)
3	4	10	0.40 (0.155)	0.53 (0.121)
4	10	19	0.53 (0.115)	0.57 (0.094)
5	5	14	0.36 (0.128)	0.48 (0.109)
6	36	46	0.78 (0.061)	0.77 (0.058)
7	9	10	0.90 (0.095)	0.80 (0.097)
8	7	9	0.78 (0.139)	0.73 (0.110)
9	4	6	0.67 (0.192)	0.68 (0.125)
Totals	106	150	0.71 (0.037)	0.68 (0.064)

The Bayes estimate column is described in the text.

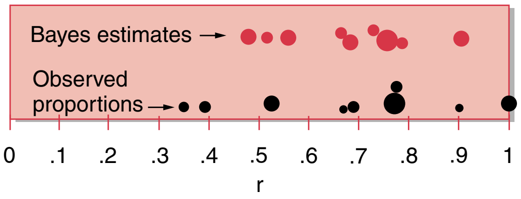

Figure 33-7 Comparison of the two rightmost columns in Table 33-2. The dot plot on the *r*-axis shows the observed success proportions, just as in Figure 33-6. The Bayes estimates assume a hierarchical model and show shrinkage toward the overall mean.

Shrinkage is a consequence of hierarchical modeling. The motivation for such modeling is to utilize the available information appropriately in improving precision or in decreasing sample size required. Consider study 1 in Table 33-2. Simply pooling the data from the other eight studies would greatly increase the precision of its estimated success rate. Namely, the standard error would be reduced from 0.116 to 0.037. But in view of the heterogeneity in the studies, such pooling would not be justified.

Borrowing hierarchically also strengthens the conclusion, with the standard deviation of the Bayes estimate being about 20% smaller, from 0.116 to 0.094. Although not nearly as great as the reduction with unabashed pooling, hierarchical borrowing is defensible because it does not make the assumption that all studies had the same true success rate, and because the extent of borrowing is determined by the data. This reduction implies more than 50% savings in sample size necessary to carry out a clinical trial (in the setting of study 1) with the same precision: (0.116/0.094)^2 – 1 = 52%. For example, to achieve the same standard error in a stand-alone study would require 25 as opposed to 16 patients.

Patient covariates can be incorporated into a hierarchical analysis, thus adjusting for known differences in the studies but still accounting for unknown effects. In this example and in more complicated hierarchical settings as well,[8] modeling allows for borrowing from other studies and databases. If the results are consistent across studies then the amount of borrowing will be greater. If the results are sufficiently different (after accounting for covariates) then this suggests heterogeneity among the studies and there is little borrowing.

Hazards over Time

Time-to-event analyses are ubiquitous in cancer research: the word "survival" appears about 100 times in Chapter 32. There are Bayesian analogues of survival analyses as described in that chapter. And there are hierarchical Bayesian analogues in which survival curves are allowed to depend on category of patient or to vary with the study in meta-analyses. However, my purpose in this section is not to extend the more traditional survival models and analyses to the Bayesian setting. Rather, I will focus on a narrow and simple aspect of survival analysis, but one that opens up understanding not possible otherwise. I will do this using data from a clinical trial, number 8541 of the Cancer and Leukemia Group B (CALGB).[9]

This trial considered three different dose-schedules of cyclophosphamide, doxorubicin, and 5-fluorouracil (CAF) in node-positive breast cancer: high, moderate (mod) and low. These are respectively, four cycles of CAF at 600, 60 and 600 mg/m^2, six cycles at 400, 40, 400, and four cycles at 300, 30, 300. The primary end-point was disease-free survival, which is shown in Figure 33-8 for the three dose-groups using Kaplan-Meier plots. I am not providing *p*-values for the various comparisons (high vs mod, high vs low) because whether these are statistically significant is not material to my purpose.

Time-to-event curves such as those in Figure 33-8 do not tell the whole story regarding any benefit of increasing dose and dose-intensity. A clearer picture is contained in hazard plots over time.

Hazards are the proportions of events from one time period to the next for those patients who are at risk at the beginning of the period. For example, if there are 100 patients in a group and 10 of these recur in the first year, then the first-year hazard is 10%. Going into the second year there are 90 patients at risk. If another 10 recur in the second year, then the second-year hazard is 10/90 = 11%. When calculating hazards from survival plots such as those in Figure 33-8 (which incorporate censored observations), subtract the current year's survival proportion from last year's survival proportion and divide by the last year's survival proportion. The resulting yearly values are shown in Figure 33-9.

A striking observation from Figure 33-9 is that all three hazards decrease over time (after the first year). This is a reflection of the heterogeneity of breast cancer. The most aggressive tumors recur early, giving the high hazards evident in the first few years. Once their tumors have recurred, patients are removed from the at-risk population. The remaining tumors are much less aggressive and so they recur at a lower rate.

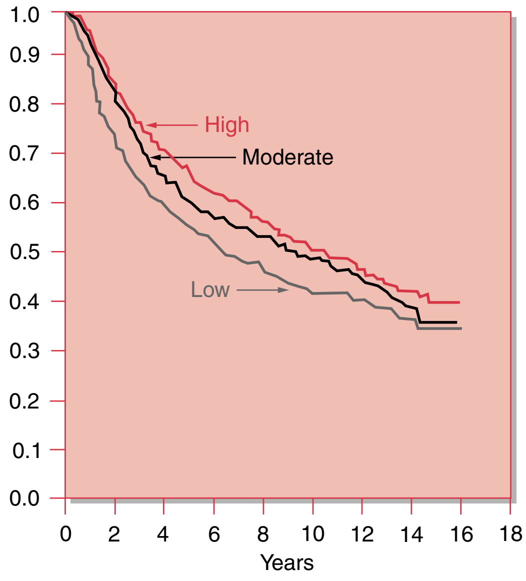

Figure 33-8 Disease-free survival proportion for the three CAF dose-groups of CALGB 8541.

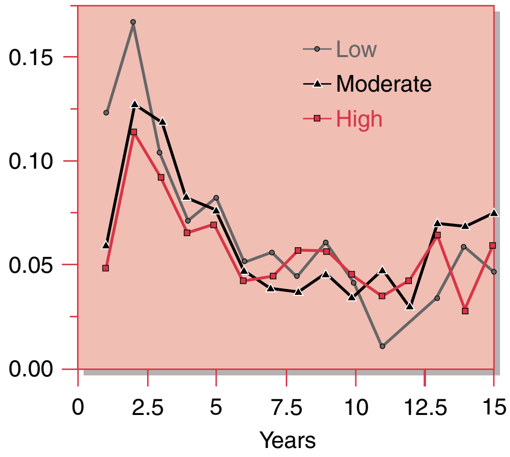

Figure 33-9 Hazards for the three CAF dose-groups of CALGB 8541, derived from Figure 33-8.

Regarding a treatment-arm effect, the apparent benefit of the high-dose schedule is restricted to the first five years or so. Actually, the hazard for patients on the high-dose schedule is lower than those of the other two arms in each of the first six years. (Although it is not much lower in the last few of these six years and it is not much lower than the moderate-dose schedule at any time.) This observation is impressive because each year is like a new study, with previous recurrences not counted when starting a new year.

Another observation from Figure 33-9 is that after 5 years the risks of all three groups converge, with the annual risk of recurrence being approximately 5% in all three groups.

The reduction in hazard of recurrence for high versus low is 14% over the 18 years of follow-up (95% confidence interval: 6-22%). This is an average over these years (weighted over time because of differences in at-risk sample sizes over time). But since there is no reduction at all in the later years, the overall reduction is being carried by the early years. Restricting to the first 3 years, the reduction is 24% (13-33%). A benefit of chemotherapy that is restricted to the first few years is typical in breast cancer trials. An implication is that a hazard reduction seen early in a trial, say one with a median of three years of follow-up, will deteriorate over time. This is because the comparison will eventually involve averaging over periods where there is no longer a treatment benefit.

In the later years, the hazards of about 5% are very similar to the annual hazard for node-negative breast cancer patients. Interestingly, convergence to about 5% applies irrespective of the number of positive lymph nodes. Figure 33-10 shows this effect. It gives hazard plots for three categories of positive nodes: 1-3, 4-9 and 10 or more (for the three dose groups combined). Early in the trial, patients with 10+ positive nodes have a very high annual recurrence rate of 20-30%. However, after five years or so, the annual hazard is about 5% in all three groups. So

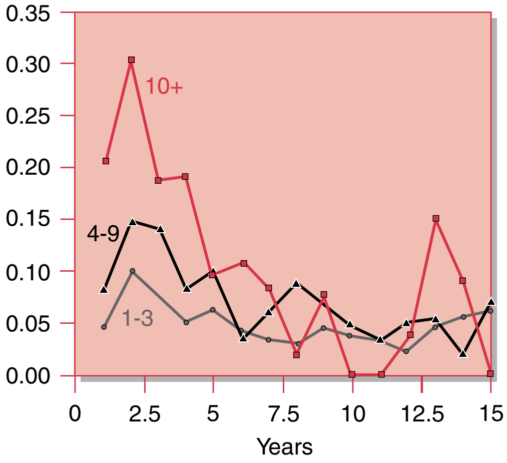

Figure 33-10 Hazards for the three categories of positive lymph nodes (1-3, 4-9, and 10 or more) for CALGB 8541. There are few patients at risk in the later years, especially in the 10+ group, and for two reasons. One is that this was the smallest group to start, with 174 of the 1550 patients in the trial, and the other is that most recurred early. The asterisk at 13 years indicates a point where there were only 24 patients at risk, and 3 of these recurred in the 13th year.

a woman with a large number of positive nodes who has not experienced disease recurrence in the first five years or so has the same updated prognosis as a woman with a small number of positive nodes, including no positive nodes. The effects of both the number of positive nodes and dose of CAF have worn off after 5 years.

An important aspect of CALGB 8541 is the role of tumor HER-2/neu expression, and in particular its interaction with dose of CAF.[10] HER-2/neu assessment was carried out for a subset of 992 patients from the original study. Its interaction with dose was shown to be significant in a multivariate proportional hazards model. But the manner of interaction is easiest to understand using hazards. Figure 33-11 shows the effect of dose of CAF separately for patients with HER-2/neu–negative ($n = 720$) and –positive ($n = 272$) tumors. HER-2/neu negatives show no dose effect. The entire benefit of the high- over the moderate-dose treatment schedule, and the high- over the low-dose treatment schedule that is observed in these patients is concentrated in HER-2/neu positives. Moreover, this benefit occurs through a reduction in hazard in each of the first 3-4 years. Again, each year is a separate study and so each of these years provides a separate confirmation of the overall conclusion. The hazard reduction in the first three years for patients receiving the high-dose treatment schedule as compared with the other two groups combined, was 65% among HER-2/neu–positives. HER-2/neu overexpression apparently conveys a poor prognosis for lower doses but not for higher doses—it might even provide a favorable prognosis for patients receiving higher doses.

Many of the conclusions in this section would have been difficult or impossible to make without considering hazards over time.

A final comment regarding hazards relates to the common problem of predicting survival results into the future for patients already accrued to a trial. This differs from the general problem of prediction discussed in "Predictive Probabilities" earlier. Consider Figure 33-8. Some of the patients have as little as 10 years of follow-up information. As more follow-up information becomes available, there will be no change in these curves prior to the 10-year time point. But the curves are subject to change beyond 10 years. Because the focus is on patients for whom the tumor has not yet recurred, the way the curves will change depends on the hazards beyond 10 years. The information available about these hazards is shown in Figure 33-9. For predicting when and whether a patient's disease will recur, consider hazards one year at a time, always building on her current year of follow-up. Each incremental hazard prediction depends on the data for the corresponding year.

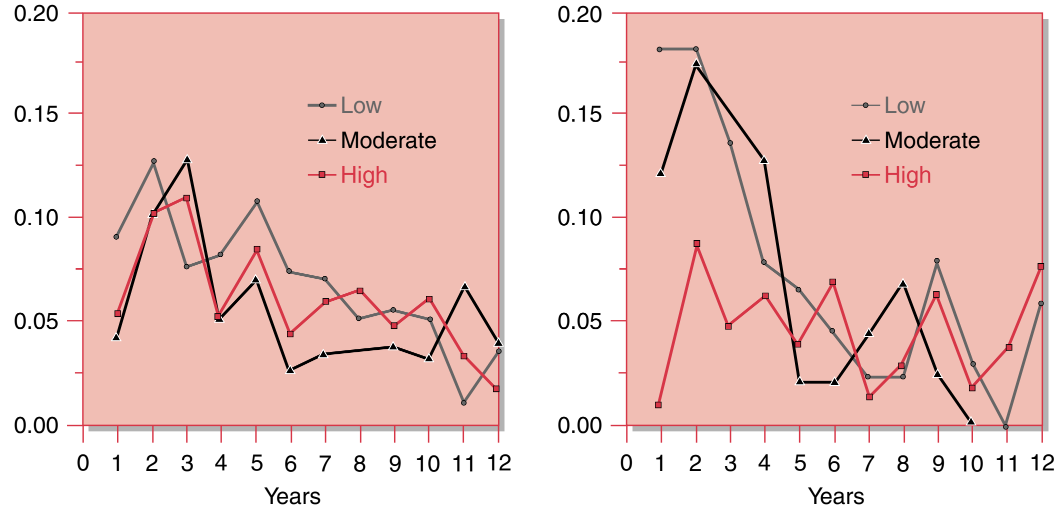

Figure 33-11 Annual disease-free survival hazards for a subset of patients ($n = 992$) in CALGB for whom expression of HER-2/neu in the patient's tumor was assessed. Patients in the left-hand panel were HER-2/neu–negative and those in the right-hand panel were HER-2/neu–positive.

DECISION ANALYSIS AND CHOOSING SAMPLE SIZE

Clinical practice and clinical research involve making decisions. An example of the latter is choosing the sample size of a clinical trial. It is impossible to precisely predict the result of making a particular decision. But one can list the possible results and associate (predictive) probabilities with each. Also associated with each possible result are the consequences of that result. A list of results, probabilities and consequences characterizes each decision, and allows for choosing one decision over another.

The consequences of a particular decision are many faceted. I consider the case in which consequences are unidimensional, and numerical. A particular "number" can always be assigned to a consequence, at least in theory. Numerical assignments that indicate the overall worth or benefit of a consequence is called its *utility*. Given a list of results, their probabilities and their utilities, it is still not clear how to weigh the various utilities. The convention in decision analysis is to average the utilities with respect to the associated predictive probabilities. The resulting average is the utility of the decision in question, and the various possible decisions can be compared on the basis of their utilities. The central role of predictive probabilities in this process makes the Bayesian approach ideally suited for decision making.

The terms decision making, decision analysis, and decision theory are used more or less interchangeably. Many references develop this subject more deeply than is possible here.[11–14]

A simple example may help fix the concept. You are offered a chance to win a prize worth $10 to you—that is, its utility is 10. It costs $1 to play. There are two decisions: play and not play. If you play then you will end up with a utility of either 9 or –1. Suppose that the probability of the former is p. (A class of decisions in which one can get information about p is interesting and revealing, but it is not considered here.) The utility of playing is then $9p–(1–p) = 10p–1$. The utility of not playing is 0. The former is greater than the latter if $p > 0.1$ and so the playing is better (according to decision-analytic convention) if $p > 0.1$.

Averaging utilities to assess taking gambles is well and good. But it is less clear that it is appropriate when outcomes are health states or results of clinical trials. Moreover, it may be difficult to assess the utilities of such outcomes. However, one must make a decision. And when faced with a list of outcomes and their associated probabilities for each available decision, reducing the lists to a single dimension greatly facilitates choosing among them. In addition, varying the probabilities and utilities assumed allows for assessing sensitivity of the various aspects of the decision process.

Consider a decision-analytic approach for choosing a sample size in a two-armed clinical trial. The purpose of clinical trials is to learn about competing therapies. The reason for wanting to learn about competing therapies is to affect the future treatment of patients having the disease in question. In a decision analysis one can consider good medicine to all patients who have the disease. The utility of any particular design of a clinical trial is its consequent impact on patients who have the disease, including those patients in the trial and those outside the trial. Let *N* be the size of the "patient horizon," those patients who will benefit from the conclusions of the clinical trials being planned.

The value of patient horizon *N* varies depending on the disease and the available treatments. The population of patients with primary breast cancer who might benefit from an advance in therapy is very large. But few would benefit from an advance in a rare type of children's cancer. These two extremes are addressed in the same way when choosing sample sizes via power calculations. Taking the same tack in both cases cannot be right from the perspective of treating as many patients with the disease in question as effectively as possible. In the case of small *N*, a substantial portion of the patients—perhaps all—may be in the clinical trial and so few if any patients will get to take advantage of results obtained in the trial. In the case of large *N*, the trial may be too small to enable an informed choice between the two treatments and so the very large number of patients outside the trial may be treated with an inferior therapy. A conclusion of this section is that when the goal is treating as many patients as effectively as possible, the sample sizes of the clinical trials in these two extremes should be very different.

The size of *N* is not usually precisely known. In particular, *N* depends on the effectiveness and side effects of both treatments, which are also unknown. However, a consequence of this section is that only the order of magnitude of *N* should be considered in choosing sample sizes of clinical trials. Precision in fixing *N* is not very important. Considering the extremes, diseases or conditions that are very common (large *N*) call for larger trials than do rare diseases (small *N*). Moreover, when *N* is unknown, the results of this section rely upon replacing *N* with its mean. So experts could assess the annual size of the patient population and the potential availability of other therapies over the next several years. Patients presenting in the future could be discounted by the probability they will be treated using one of the treatments involved in the trial. This gives an expected value of *N* that can be used in designing the trial.

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

The optimal trial sample size has order of magnitude square root of *N*.[15] If there are two clinical trials (followed by clinical practice with the better performing therapy) then the first of the two should have sample size with order of magnitude cube root of *N*. Table 33-3 considers a setting in which the optimal trial sample size for a common disease with *N* of about one million turns out to be 1000. The point of the table is to compare this with the corresponding optimal sample sizes for rarer diseases. The table also shows the optimal sample size for the first of two clinical trials. The sample sizes are strikingly different for common versus uncommon diseases.

In a decision analysis one can explicitly consider asymmetry in information concerning the treatment arms under consideration. The allocation proportions should also be asymmetric. As above, continue to assume a two-armed trial with the goal being to effectively treat as many patients in horizon *N* as possible. Consider the particular forms of prior information about the unknown rates of success that are shown in Figure 33-12. Suppose the success rate for one of the arms, say arm 1, has distribution A and that for the other arm, arm 2, has either distribution A, B or C.

Consider *N* = 100 or greater, as indicated in Table 33-4. The table shows the optimal sample sizes for each of the arms. As advertised above, these increase with *N* in proportion to the square root of *N* (approximately). Consider case *N* = 1000 and distribution A for both success rates. The table indicates that 21 patients should be assigned to one of the arms and 20 to the other. In view of symmetry, either arm 1 or arm 2 could get the extra patient. (The reason these two numbers are not equal is instructive. Adding an extra patient to arm 2 would never change the optimal arm for the patients outside the clinical trial. The only consequence of this patient is to introduce the possibility of ties in the observed success rates, in which case both arms would be optimal outside the trial.) Using this assignment, the resulting success proportion among the 1000 patients in the horizon is 65%. Increasing either or both sample sizes decreases this success proportion. And decreasing either or both sample sizes decreases this success proportion. To consider the extremes, both not running a clinical trial at all and running a trial entering all 1000 patients have expected success proportions of 50%.

For the case distribution A versus B in Table 33-4, arm 2 is more promising than arm 1 and so it is assigned to more patients in the trial—about $\sqrt{3} - 1 = 73\%$ more for large *N*. For distribution A versus C, the success rate for arm 2 is known to be 0.5. The success rate for arm 1 could be greater than 0.5 or less than 0.5. The trial's purpose —in addition to treating patients effectively—is to identify whether the arm 1 success rate is greater than or less than 0.5. To achieve this purpose it would be a waste to allocate patients to arm 2 because its success rate is known. Arm 2 is held in reserve and will be used after the trial should it turn out that arm 1's observed success rate is less than 0.5. (Only rarely is it reasonable to assume that a treatment's success rate is known. As indicated in Chapter 32, patient populations vary over time and so a treatment's effect may similarly vary.)

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

ADAPTIVE DESIGNS OF CLINICAL TRIALS

Chapter 32 addresses the traditional approach to designing clinical trials, particularly as regards sample size. The first step in planning a trial from a Bayesian perspective is to assess the available evidence regarding the hypotheses and parameters of interest. The designer addresses the possibility of using this information in a prior distribution or incorporating it in a hierarchical model along with the results of the trial being planned.

At the planning stage it is important to consider the possible state of affairs when the trial is over. One consideration is the set of implications and consequences of each possible result. Another is the predictive probability of each possible result. The previous section presents an approach in which utilities are assessed for the former and weighed with respect to the latter. The present section deals with designs that are more flexible than the ones in the previous section. Although the

Table 33-3 Sample Sizes of a Clinical Trial Assuming that the Optimal Sample Size has been Calculated when N is One Million and Turns out to be 1000 (for a single trial)

Patient horizon, N	1,000,000	100,000	10,000	1000	100
Single trial sample size	1000	320	100	32	10
Sample size for first of two trials	170	78	36	17	8

Sample sizes provided with two-digit accuracy. The relationship between sample sizes within each row is general, but the relationship across rows (1000 versus 170, for example) applies only for a particular prior distribution of the unknown parameters.

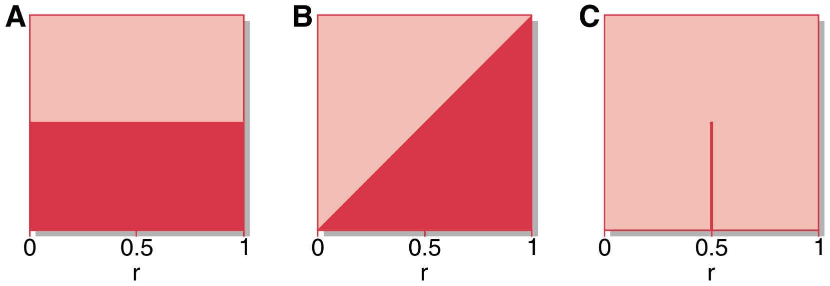

Figure 33-12 Three different prior distributions for r, rate of success. Under distribution A, r is equally likely to be any value between 0 and 1. The density in B is proportional to r, which means, for example, that r greater than 0.5 is three times as probable as r less than 0.5. (These two distributions are the same as the two in Figure 33-1, "a.") Under distribution C, all the probability is concentrated on $r = 0.5$ and so in this case the arm's effectiveness is assumed to be known.

designs in this section are not based on an explicit consideration of utilities, the goals are efficient learning and effective treatment of patients. For explicit decision-analytic generalization of some parts of this section, see *Bandit Problems: Sequential Allocation of Experiments*.[16]

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

Designs of clinical trials are usually static in the sense that the sample size and any prescription for assigning treatment, including for randomization protocols, are fixed in advance. Results observed during the trial are not used to guide its course. There are exceptions. Some Phase II cancer trials have two stages, with stopping after the first stage possible if the results are not sufficiently promising. And most Phase III protocols specify interim analyses that determine whether the trial should be stopped early for sufficiently strong evidence of a difference between competing treatment arms. However, traditional early stopping criteria are very conservative and so few trials stop early.

The simplicity of trials with static designs makes them solid inferential tools. Their sample sizes tend to be large, at least in comparison with alternatives to be discussed in this section. And they usually consider two therapeutic strategies, or arms, thus enabling straightforward treatment comparisons. I do not mean that static trials always give clear answers as to whether one arm is better than the other, but only that they usually allow for an unambiguous quantification of the uncertainty regarding whether one arm is better.

Despite their virtues, static trials result in slow and unnecessarily costly drug development. Hundreds of millions of dollars and many years can be expended in developing a single cancer drug, one that may not make it to market. For a company developing a moderate number of drugs (say 20 or more), this circumstance is tolerated because costs are balanced by profits from other drugs. Smaller companies are at the mercy of the prevailing attitudes toward drug development and risk going belly up.

The tradition of drug development is one at a time. The number of cancer drugs available for development is increasing exponentially. It is inefficient to focus on a single drug while a gazillion others are sitting on the sidelines waiting to be evaluated. The standard types of errors in drug development are false positives and false negatives. These errors apply to drugs actually being tested. Another kind of error applies to drugs not under investigation: Every such drug is a false neutral. A drug not being developed has no chance of helping anyone. Finite resources limit the ability of the medical establishment to develop therapies. But when resources are limited we should approach their allocation in a more rational way. And what makes sense today may well be different from the ways of the past.

Pharmaceutical companies and medical researchers generally must be able to consider hundreds or thousands of drugs for development at the same time. Static trials inhibit the simultaneous processing of many drugs. And they cannot efficiently address dose-response questions when many drugs are under consideration. Dynamic designs that are integrated with the drug development process are necessary for reasonable progress in medical research.

The focus of this section is a family of designs that are dynamic in the sense that observations made during the trial can affect the subsequent course of the trial. The general class of designs is *adaptive* or *sequential*. The focus is clinical trials, but the ideas apply at least as forcefully in the preclinical setting. A main bottleneck of the drug development process occurs at the level of the preclinical animal toxicity/carcinogenicity studies. There are many opportunities for using adaptive designs in the preclinical area that will efficiently identify the best drugs to move forward in trials for humans.

Using an adaptive design means examining the accumulating data periodically—or even continually—with the goal of modifying the trial's design. These modifications depend on what the data show about the unknown hypotheses. Among the modifications possible are stopping early, restricting eligibility criteria, expanding accrual to additional sites, extending accrual beyond the trial's original sample size if its conclusion is still not clear, dropping arms or doses and adding arms or doses. All of these possibilities are considered in the light of the accumulating information. Adaptive designs also include unbalanced randomization where the degree of imbalance depends on the accumulating data. For example, arms that give more information about the hypothesis in question or that are performing better than other arms can be weighted more heavily.[16]

Adaptation is not limited to the data accumulating in the trial. Information that is reported from other ongoing trials can also be used. This is easier to effect if one takes a Bayesian approach, possibly using hierarchical modeling as described in the previous section.

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

Table 33-4 Optimal Allocations of Sample Size to Arms 1 and 2 in a Two-armed Clinical Trial plus Success Proportion among the N Patients for that Allocation

Patient	Prior distributions of rates from Figure 33-12; optimal success proportion								
horizon, N	1: A	2: A	Prop	1: A	2: B	Prop	1: A	2: C	Prop
100	6	5	0.63	4	8	0.71	9	0	0.60
1000	21	20	0.65	16	30	0.74	29	0	0.62
10000	70	69	0.66	56	98	0.75	99	0	0.62
Large N	$\sqrt{N}/2$	$\sqrt{N}/2$	2/3	$\sqrt{N}/3$	$\sqrt{N}$	3/4	$\sqrt{N}$	0	5/8

For each value of N there are three pairings of prior distributions of the arm 1 and arm 2 success rates considered. The optimal asymptotic (large N) success proportion is the expected value of the maximum of the two success rates, where the expectation is with respect to the prior distribution.

Prop = success proportion.

Continuous Reassessment Method (CRM) in Phase I As indicated in Chapter 32, the purpose of Phase I cancer trials is to identify the maximum tolerated dose (MTD). The most commonly used Phase I designs are variants of the so-called "3+3" design. Patients are admitted in groups of 3. If none of the 3 experiences toxicity then the dose is increased one level for the next group of 3. If 2 or 3 of the 3 experience toxicity then the next lower dose is the MTD. If 1 of the 3 experiences toxicity then 3 more patients are added at the same dose level. If 2 or more of the 6 patients at that dose experience toxicity then again the next lower dose is the MTD.[17]

This design is adaptive, but its adaptation is very crude. Such a design is likely to assign low doses and to select an MTD that is ineffective. Moreover, such a design ignores important information that is available in the trial. In particular, dose assignments are not based on *sufficient statistics*.[18] An alternative approach uses Bayesian updating: the continual reassessment method (CRM) of O'Quigley and colleagues[19] Updating takes place assuming a particular model of the relationship between dose and toxicity (such as the logistic function). The CRM too is adaptive. Each patient is assigned to the dose having probability of toxicity closest to some predetermined target value. This is the Bayesian posterior probability calculated from the data available up to that point (and so it is based on sufficient statistics).

The CRM more effectively finds the MTD than does the 3+3 design. The CRM is the standard design used in Phase I trials at MDACC. But it is rather crude and we are improving it in a number of ways. One of these ways is based on the fundamental principle that ignoring information is wrong. (A catch, of course, is that taking information into account is work, and it can require modeling.) There is some information that accrues about efficacy in a Phase I trial. This information is limited, especially regarding the dose-efficacy relationship. But at a minimum, in proceeding to Phase II with a particular dose (usually the MTD), one should use the efficacy information from those patients in Phase I who were assigned to that dose. This notion leads to using a Phase I/II design that addresses safety and efficacy simultaneously, or the focus turns to efficacy after an initial focus on toxicity. Such an approach is efficient from the perspective of both time and patient resources.

A way in which both 3+3 and CRM designs are crude is the need to pause accrual while waiting for toxicity information.[20,21] Such pauses are inefficient and they cause logistical problems. Trials should be paused or stopped if there are safety concerns, not because the design cannot get out of its own way. In getting information about toxicity (or efficacy), there is seldom a magical dose that the next patient must get. All doses are potentially informative. Rather than stopping, one should use a design that models dose-response (toxicity and efficacy) and is able to assign a next dose even though patients previously treated are not yet fully evaluable.

Another way in which both 3+3 and CRM designs are crude is the assumption that toxicity is dichotomous. An approach that is better—again because of using all available information—would be to account for severity of toxicity. Again, it would be better to consider both severity of toxicity and efficacy in a Phase I/II design.[22] Assigning utilities to the various possible health states would lead to weighing these two conflicting desiderata in a decision analysis.

Adaptive Dose-finding in Phase II In many diseases, the standard Phase II dose-finding design is to allocate a fixed number of patients to each of a number of doses in a grid. Such questions are generally of less interest in cancer because of the MTD mentality: administer as much drug as the patient can tolerate. But with the increasing interest in biological agents, dose finding for efficacy is becoming important in cancer research.

After seeing the results of a dose-finding trial, the investigators usually wish they had assigned patients in some other fashion. Perhaps the dose-response curve was shifted more to the left or right than anticipated. If so, then assignment of the bulk of patients on one end or the other was wasted. Or perhaps the slope of the dose-response curve is greater than anticipated and the response of patients assigned to the flat regions of the curve would have been more informative if the doses assigned had been in the region where the slope is apparently greatest. Or perhaps results for the early patients made it clear that the dose-response curve was flat over the entire range and therefore the trial could have stopped earlier. Or perhaps the results of the trial show that the standard deviation of the outcome of interest is greater or less than anticipated and so the trial should have been larger or could have been smaller.

The approach of Berry and colleagues [23] is to proceed sequentially, analyzing the data as it accumulates—see also Malakoff.[5] There are two stages of the trial, first dose ranging (Phase II) and then confirmatory (Phase III), if the latter is warranted. The dose-ranging stage continues until a decision is made that the drug is not sufficiently effective to pursue future development or that the optimal dose for the confirmatory Phase III trial is sufficiently well known. (Switches to Phase III can be effected seamlessly and without stopping accrual—see below, and this is so even if the endpoint of interest is delayed, such as time to progression.) The example trial of Berry and colleagues [23] involves a biological neuroprotective agent for stroke. But the same principles of trial design apply in cancer. Each entering patient is assigned the dose (one of 16, including placebo) that maximizes information about the dose-response relationship, given the results observed so far. This dose could be in the region of greatest apparent slope, or it could be placebo or a high dose. But future patients are not assigned to doses in any region where accumulating evidence suggests that the dose-response curve is flat.

In the dose-ranging stage, neither the number of patients assigned to any particular dose nor the total number of patients assigned in this stage are fixed in advance. The dose-ranging sample size will be large when the data suggest that the drug has moderate benefit, when the dose-response curve is gently sloping, or when the standard deviation of the responses is moderately large. It will tend to be small if the drug has substantial benefit, if the drug has no benefit, if the dose-response curve rises over a narrow range of doses, or if the standard deviation of the responses turns out to be small. (In addition, and somewhat non-intuitively, the dose-ranging stage will be small if the standard deviation of responses is very large. The reason is that a sufficiently large standard deviation means that a very large sample size would be required to demonstrate a beneficial drug effect. The required sample size may be so large that it would be impossible to study the drug and so the trial stops in the dose-ranging phase before substantial resources go down the drain.)

In the stroke trial considered by Berry and colleagues,[23] the ultimate endpoint is improvement in stroke scale from baseline to 13 weeks. If the accrual rate is large then the benefit of adaptive assignment can be limited by delays in obtaining endpoint information. To minimize the effects of delayed information, each patient's stroke scale is assessed weekly between baseline and week 13. Within-patient measurements are correlated, with correlations greater if they are closer together in time. We incorporate a longitudinal model into the analysis of the trial and carry out Bayesian predictions of ultimate endpoint based on current patient-specific information, and we update probability distributions of treatment effect accordingly.

Adaptive dosing is more effective than is the standard design at identifying the right dose. And it usually identifies the right dose with a smaller sample size than when using fixed dose assignments. Another advantage is that many more doses can be considered in an adaptive design. (Even though some doses will be little used and some might never be used, these cannot be predicted in advance.) An adaptive design therefore has some ability for distinguishing responses at adjacent doses and for estimating nuances of the dose-response curve.

The circumstances of the stroke trial are similar to those in many other types of trials. Finding the right dose is a ubiquitous problem in pharmaceutical development, and it is seldom done well or efficiently. The adaptive nature of the stroke trial would be less advantageous if we had not exploited early endpoints. Cancer too is characterized by the availability of information about a patient's performance (local control of the disease, biomarkers, etc.) before reaching the primary endpoint. Finally, the possibility of moving seamlessly into Phase III depending on the Phase II results exists for many types of drugs. That issue leads naturally to the subject of the next section.

Seamless Phases II and III The convention of categorizing drug development into phases is unfortunate. We proceed from one phase to the next when we think we know something: the MTD from Phase I or that a drug's impact on a Phase II endpoint will translate into a benefit in Phase III, and at the Phase II dose. In the Bayesian approach one never takes a quantity to be perfectly known. Instead, the Bayesian perspective is to carry along uncertainty with whatever knowledge is available. Phases of drug development are arbitrary labels that describe a process that is—or should be—continuous.

One of the consequences of partitioning drug development into phases is that there are delays between phases. For example, there is a pause between phases II and III to set up one or more pivotal studies. As mentioned above in the context of the stroke trial, its design allows for avoiding such a hiatus. At each time point, say weekly, the algorithm that guides the conduct of the trial carries out a decision analysis and recommends either (1) continue the dose-ranging stage of the trial, (2) stop the trial for lack of efficacy (inadequate slope of the dose-response curve or, more accurately, evidence of a positive dose-response that is insufficient to justify continuing the trial), or (3) shift into a confirmatory trial. This shift can be made seamlessly, with no break in accrual. Indeed, it is even theoretically possible to effect such a shift in a double-blind trial without informing the investigators: they simply continue to randomize doses, but unbeknownst to them, the only two being assigned are the Phase III dose and placebo.

We designed a trial at MDACC[24] that encompasses both phases II and III. If there is a switch to Phase III, this switch is seamless. The anticipated effect of the drug is on local control. We model survival as it depends on local control and as it depends on treatment. (Though the possibility is remote, we allow for the experimental drug to have a beneficial effect on survival that is not mitigated by local control.) So local control is a surrogate endpoint in a way similar to the way early stroke scale assessments are surrogate endpoints in the stroke trial. But the clear focus is on survival as the main endpoint and the utility of the surrogate endpoint must be demonstrated by the results actually observed in the trial. We exploit any relationships that exist, but do not assume such relationships. We analyze the data in the trial frequently and adapt to the accruing evidence.

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

The trial continues until stopping occurs because (1) continuing would be futile, judged by predictive probabilities, (2) the maximum sample size is reached, or (3) the predictive probability of eventually achieving statistical significance becomes sufficiently large. Should the third event occur, accrual ceases and the pharmaceutical company prepares a marketing application.

The sample size of a conventional Phase III trial with the desired operating characteristics is 900. We take this to be the maximum sample size in the seamless design as well. Actual accrual is very likely to be much less than this maximum sample size, and on average it will be about half as large. On the other hand, incorporating the same number of interim analyses in a conventional design using a conventional type of stopping boundary allows for only a slight decrease in average sample size. Under any hypothesis, null or alternative, the Bayesian design occasionally leads to a relatively large trial (close to 900 patients). However, a pleasant aspect of the design is that the sample size is large precisely when a large trial is necessary. Conventional trials may well (and sometimes do!) come to their predetermined end with an ambiguous conclusion. In a Bayesian approach one may choose to continue such a trial to resolve the ambiguity, and this option has substantial utility. (Carrying this argument to the maximum sample size, there may be times for which stopping at 900 is ill advised, but for logistical reasons we specified a maximum size.)

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

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

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

More than a dozen trials at MDACC have been designed and are being conducted using adaptive allocation. Our standard approach is to randomize treatment assignment, but we shift the weights toward better performing arms as the trial proceeds and the results accumulate. Many of these trials have more than two arms. The arms are sometimes distinct therapies, and sometimes they are closely related. An example of the latter is an MDACC trial involving five doses (including 0) of a drug (pentostatin). The goal is to inhibit graft-versus-host-disease (GVHD) in leukemia patients who are receiving bone marrow transplants. The problem is that the drug may inhibit successful engraftment of the transplant, which is necessary for survival. Such inhibition may be related to dose. We use a combination endpoint: survival at 100 days free of GVHD. The conflict between engraftment and freedom from GVHD means that the dose-response curve may not be monotone. In particular, it may increase for small doses and then decrease. Initially we assign doses in a graduated fashion, climbing the dose ladder slowly. But as doses become admissible, we assign patients to those that have been performing well.

Consider a patient who qualifies for the trial. To decide which pentostatin dose to assign we calculate the current (Bayesian) probabilities that each admissible dose is better than placebo. This calculation uses all information from patients treated to date. We allocate doses randomly, with weights proportional to these probabilities. We consider other allocation algorithms, including assigning in proportion to powers of these probabilities. The assignments involve some degree of randomization, but all patients are more likely to receive doses that are performing better. Doses that are doing sufficiently poorly become inadmissible in the sense that their assignment weight becomes 0. When and if we learn that the drug is effective, we stop the trial. When and if we learn that the drug is ineffective, then again we stop the trial. Patients in the trial benefit from data collected *in the trial*. The explicit goal is to treat patients more effectively, but in addition we learn efficiently. We evaluate each design's frequentist operating characteristics using Monte Carlo simulation, possibly modifying the parameters of the assignment algorithm to achieve desired characteristics.

Process or Trial? Evaluating Many Drugs Simultaneously Using Adaptive Allocation The greatest need for innovation and the greatest room for improving drug development is effectively dealing with the enormous numbers of potential drugs that are available for development. The notion of developing drugs one at a time is part of the pharmaceutical culture. It will change. Companies that are able to screen many drugs simultaneously and do so effectively will survive and others will not.

Many different drugs should be evaluated in the same preclinical experiment or collection of experiments. Information should be updated frequently or even continually. The extent to which any particular drug is used and the order of drugs used will depend on the available data. Drugs that are apparently more promising will move faster through the preclinical setting. Drugs that give disappointing data will languish. And the sample sizes of drugs whose promises and toxicities are not clear will tend to be large so as to enable resolving uncertainties.

These ideas and imperatives apply as well to drugs' clinical development. As an example, at MDACC we are building the foundation for a Phase II trial for evaluating drugs that is more a process than a trial. The idea is an extension of the adaptive assignment strategies described in the previous section. We start with a number of treatment arms plus a control—possibly a standard therapy. We randomize to the arms and learn about their relative efficacy as the trial proceeds. Arms that perform better get used more often. An arm that performs sufficiently poorly gets dropped. An arm that does well enough graduates to Phase III, and if it does sufficiently well it might even replace the control. As more arms become available, they are added to the mix.

The result is that better arms move through quickly and poorer arms get dropped. Patients in the trial are provided with better treatment (when the arms are not equally good). Patients outside the trial get access to better drugs more rapidly.

Extraim Analyses

A common circumstance is that a clinical trial ends without a clear conclusion. For example, a statistical significance level of 5% in the primary endpoint may be required for drug registration and the *p*-value turns out to be 6%. The regulatory agency suggests that the trial was "underpowered" and that the company should carry out another trial. It would be much more efficient to simply increase the sample size in the present trial. The problem is that the possibility of such an extension increases the type I error rate. The principle is identical to that for interim analyses.

The solution is to build into the design the possibility of continuing the trial depending on the results, suitably adjusting the significance levels. In contrast to adjustments for interim analysis, the adjustments for "extra-im" (extraim) analyses are reversed, with much of the overall significance level "spent" at the originally planned sample size. For example, taking equal significance levels at each possible termination point is preferable to O'Brien-Fleming stopping boundaries because the latter are too conservative for extraim analyses. Allowing for extending the trial increases the maximal sample size and also the average sample size. But a modest increase in average sample size (such as 20%) comes with a substantial increase in statistical power (such as 80% increasing to 95%). The reason for this beneficial trade-off is that the trial is extended only when such an extension is worthwhile.

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

The increment in sample size depends on the available data at the time the decision is made to continue accrual. It also depends on the number of possible extensions. In trials I have designed, I base each extension on *predictive power*. The usual definition of power assumes a particular value of the parameter of interest, say *r*. Predictive power considers all possible values of *r*. The data available at the time of the extraim analysis plays two roles. First, they count in the final results of the trial. Second, they are used to update the (Bayesian) probability distribution of *r*. Fix the total sample size *n* and calculate the power for detecting each possible value of *r*. Average this power with respect to the probability distribution of *r* to give predictive power for sample size *n*. Extend accrual by the minimum sample size that gives total sample size having pre-specified predictive power. If there is no such value of *n*, then continuing accrual may be unwise.

There is an aspect of the above development that may be unrealistic. Namely, it assumes that endpoints for those patients treated in the trial so far are available at the time of the extraim analysis. Even if the endpoint is tumor response, there is a delay in obtaining this information. There is no need to stop the trial just because some of the endpoint information is unavailable. Rather, these data can be predicted along with that from patients not yet accrued. If there is some early information (biomarkers, performance status, etc.) that is correlated with the endpoint of interest then this can be used to inform the prediction. A special and important case is when the endpoint is time to event. The fact that a patient has not yet reached an event is useful information in predicting the time to that event. But if there is no patient-specific early information, then patients treated but not yet assessed for response are treated in the same way as patients not yet treated. (This set of issues is sufficiently important that they deserve being addressed separately—see the next section)

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

AUXILIARY VARIABLES, BIOMARKERS, AND BIOLOGICAL AGENTS

The adaptive designs considered in the previous section are based on information on the primary endpoint that accrues during the trial. If the primary endpoint is delayed and accrual is sufficiently fast then adaptive methods are of limited value. This section addresses statistical procedures for designs that exploit information on other than the primary endpoint that accrues during the trial.

Using Auxiliary Variables Information that accrues during a trial has a broad interpretation. Suppose that the endpoint is time to progression and a patient has not yet progressed. That is information, and it can be used to update the distributions of whatever parameters are involved.

In addition, information accrues about each patient's circumstances and each patient's condition. Whether the patient's tumor has responded is information, and this is so even if response is not the endpoint of interest. Moreover, time to tumor response can be informative. A patient's performance status can change over time (or not!) and such information is important and the various possibilities can be used prospectively in designing a trial. There are many such variables that might be considered. They are *auxiliary variables* since they may contain information about the primary endpoint even though they are not themselves endpoints.

The critical issue is how to take advantage of the wealth of information that accrues in a trial. The answer is modeling. A model can relate the early information to the primary endpoint.

There are several benefits of modeling. One benefit was considered in the sections on adaptive dose-finding and on seamless phases. Waiting for long-term endpoints may rule out the ability to modify the design of a clinical trial during its course. Using auxiliary variables can make adaptation possible. Another benefit of modeling is that the relationship between the primary endpoint and auxiliary variables may allow for announcing trial results earlier or for getting earlier regulatory approval of an experimental drug. For example, suppose that survival is the primary endpoint and that modeling its relationship with response was considered explicitly in the design of the trial. Accrual to the trial has ended and all patients have been treated. There is insufficient information to conclude drug benefit on the basis of survival alone. But the drug has a positive impact on tumor response. And it turns out that in both drug and control groups there is a clear relationship between response and survival. A model can utilize this information to conclude a survival benefit.

Chapter 32 addresses surrogate endpoints. An auxiliary variable may or may not be a surrogate endpoint. This distinction is critically important. In the above example, tumor response is an auxiliary variable and it is not assumed to be a surrogate for survival. The focus of the definitive analysis is the primary endpoint and not the auxiliary variable. The conclusion of the trial is that the drug improves survival or not.

A model incorporating early information can be arbitrarily complicated and, in particular, it can contain all the variables discussed above. However, one should tiptoe into model development and consider one auxiliary variable at a time. Especially important will be to consider the possibility that any relationship between the auxiliary variable and the primary endpoint depends on treatment. So treatment must be explicitly considered in the equation. Should it happen that there is an interaction between the auxiliary variable and treatment—such as that tumor response is related to survival in the control group but not in the treatment group—then the model automatically discounts the auxiliary variable and relies on survival data alone.

Little is lost by modeling, and much can be gained, as indicated in the seamless Phase II/III trial design presented earlier. Again, the gains and any losses can be assessed by simulation.

A special type of auxiliary variable is a biomarker. Models relating to primary endpoints can be based on longitudinal models that incorporate biomarker information that accrues over time. An example of a longitudinal model is described in Berry and colleagues,[23] which is set in the context of a stroke trial.

Developing Biological Agents Biological agents present special drug development problems. Historically, oncology drug development has dealt primarily with cytotoxic agents. Drug activity was judged by assessing tumor growth. An effective biological agent may well have an impact that slows tumor growth rather than killing the tumor. Or it might even allow tumor growth but halt tumor spread.

A possible strategy is to include stable disease with partial and complete tumor response as the Phase II endpoint. Another is to use time-to-progression as the Phase II endpoint. The latter has the drawback that sample sizes and length to trial may increase, but not as much as when skipping Phase II entirely. Both strategies may succeed. And both have the advantage that they do not involve a paradigm shift for the oncology research community. But there are better options.

There are two types of biological agents, those with and those without measurable targets (such as specific oncoproteins). I distinguish between these as follows: If there is a target and the drug does not affect the target then the drug cannot be effective. Circumstances are usually more complicated and affecting a particular target may not be a drug's only mechanism. In such a case the drug is in essence a generic biological agent that might affect cancer through a variety of pathways.

Another reason one might view a targeted agent to be generic is when assessing expression of the target is subject to error. An example is trastuzumab, which may well benefit tumors that have normal levels of HER-2/neu because laboratory tests are not perfect in assessing expression levels.[25,26]

Whether or not a biological agent has a measurable target, the section "Using Auxiliary Variables" applies. If there is a target, think of it as a biomarker and develop a longitudinal model to relate its level with the primary endpoint, usually time to progression or overall survival. If there is no measurable target then identify auxiliary variables (biological and otherwise) that *may* be correlated with the primary endpoint. Model the possibility of a relationship should one exist. Again, the goal is to learn early and quickly as to whether the drug has a benefit, and by which route that benefit travels.

ACKNOWLEDGEMENTS:

The Cancer and Leukemia Group B gave permission for using data from CALGB 8541. Peter Thall, Jeff Morris and the editors made helpful suggestions.

REFERENCES:

1. Berry DA. Statistics: a Bayesian perspective. Belmont(CA): Duxbury Press; 1996.
2. Berry DA. A case for Bayesianism in clinical trials (with discussion). Stat Med 1993;12:1377–404.
3. Berry DA, Stangl DK. Bayesian biostatistics. New York: Marcel Dekker; 1996.
4. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. Am Sci 1988;76:159–65.
5. Malakoff D. Bayes offers a "new" way to make sense of numbers. Science 1999;286:1460–4.
6. Hawking SW. A brief history of time: from the big bang to black holes. New York: Bantam Books; 1988.
7. DuMouchel W. Bayesian Metaanalysis. In: Berry DA, editor. Statistical methodology in the pharmaceutical sciences. New York: Marcel Dekker; 1989. p. 509–29.
8. Thall PF, Wathen JK, Bekele BN, et al. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. 2003. [In press]
9. Budman DR, Berry DA, Cirrincione CT, et al. Dose and dose intensity as determinants of outcome in the adjuvant treatment of breast cancer. J Natl Cancer Inst 1998;90:1205–11.
10. Thor A, Berry DA, Budman D, et al. *erb*B-2, p53 and efficacy of adjuvant therapy in lymph node-positive breast cancer. J Natl Cancer Inst 1998;90:1346–60.
11. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworth and Heinemann; 1988.
12. Clemen RT. Making hard decisions. Boston: PWS-Kent; 1991.
13. Berry DA. Decision analysis and Bayesian methods in clinical trials. In: Thall PF, editor. Recent advances in clinical trial design and analysis. New York: Kluwer Press; 1995. p. 125–54.
14. Lewis RJ, Berry DA. Decision theory. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. Vol. 2. New York: John Wiley & Sons; 1998. p. 1109–18.
15. Cheng Y, Su F, Berry DA. Choosing sample size for a clinical trial using decision analysis. 2002. [In press].
16. Berry DA, Fristedt B. Bandit problems: sequential allocation of experiments. London: Chapman-Hall; 1985.
17. Dixon WJ. The up-and down method for small samples. J Am Stat Assoc 1965;60:967–78.
18. Berry DA, Lindgren BW. Statistics: theory and methods. 2nd ed. Belmont(CA): Duxbury Press; 1996.
19. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. Biometrics 1990;52:673–84.
20. Thall PF, Lee JJ, Tseng C-H, Estey E. Accrual strategies for phase I trials with delayed patient outcome. Stat Med 1999;18:1155–69.
21. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics 2000;56:1177–82.
22. Thall PF, Russell KT. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. Biometrics 1998;54:251–64.
23. Berry DA, Mueller P, Grieve AP, et al. Adaptive Bayesian designs for dose-ranging drug trials. In: Gatsonis C, Carlin B, Carriquiry A, editors. Case studies in Bayesian statistics V. New York: Springer-Verlag; 2001. p. 99–181.
24. Inoue LYT, Thall P, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. 2002. [In press].
25. Paik S, Bryant J, Tan-Chiu E, et al. Real-world performance of HER2 testing—National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst 2002;94:852–4.
26. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the Breast Intergroup Trial N9831. J Natl Cancer Inst 2002;94:855–7.

NOTES FROM BONNIE

Please have the equations checked carefully.