ED 327 548 TM 015 655

AUTHOR Berry, Donald A.
TITLE Inferential Aspects of Adaptive Allocation Pules.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE 90
CONTRACT NSF-DMS-8803087
NOTE 16p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Mathematical Models; *Patients; *Research Methodology; Research Projects; Scientific Research; *Statistical Inference; Therapy
IDENTIFIERS *Adaptive Allocation (Testing); *Clinical Trials; Variance (Statistical)

ABSTRACT

In clinical trials, adaptive allocation means that the thera,ies assigned to the next patient or patients depend on the results obtained thus far in the trial. Although many adaptive allocation procedures have been proposed for clinical trials, few have actually used adaptive assignment, largely because classical frequentist measures of inference are difficult or impossible to calculate when the allocation is adaptive. The general problem of making inferences in classical trials, whether randomized, adaptive, or open, is discussed; and Bayesian inference is described as being well-suited to the scientific method. Bayesian analyses of adaptive and other studies are illustrated with examples drawn from the following studies: (1) a study by R. H. Bartlett and others (1985) involving 12 patients; (2) a 39-patient study by J. H. Ware (1989); and (3) a study by D. O. Dixon and others (1989) involving 16 patients. These studies illustrate that Bayesian inference may be possible in clinical trials, but adjusting for variance is essential. Three data tables and two graphs are included. A 27-item list of references is provided. (SLD)

INFERENTIAL ASPECTS OF ADAPTIVE ALLOCATION RULES

Donald A. Berry, School of Statistics
270 Vincent Hall, 206 Church Street SE, Minneapolis, MN 55455

INFERENTIAL ASPECTS OF ADAPTIVE ALLOCATION RULES

by

Donald A. Berry[*], School of Statistics
270 Vincent Hall, 206 Church Street SE, Minneapolis, MN 55455

Key Words and Phrases: Adaptive methods, Statistical inference, Scientific method, Early stopping, Sequential analysis, Bayes's theorem, Likelihood.

ABSTRACT

Many adaptive allocation procedures have been proposed for clinical trials. Few trials have used adaptive assignment. A principal reason is the inability to use classical statistical inferences with adaptive procedures. The general problem of making inferences in clinical trials, whether randomized, adaptive, or open, is discussed. Bayesian inference is described and illustrated in three actual trials, each with a different design.

1. Introduction

The focus of adaptive allocation methodology is on design. A principal reason such methodology is so infrequently used in actual clinical trials is the difficulty in making classical frequentist inferences when using an adaptive design. The focus of this paper is on inference. So I will address designs other than adaptive; these include open studies and other studies that do not have a particular design.

The design of an experiment is the set of actions taken by the investigator during the course of the experiment. The design is adaptive if these actions can depend on results that are observed while the experiment is in progress. It is nonadaptive if they cannot (that is, if actions are constant functions of results). Few clinical trials have adaptive designs.

In a typical randomized clinical trial (RCT), half the patients are randomly assigned to an experimental therapy and the other half serve as controls. The number of patients in the trial is part of the design. P-values are calculated by adding the probabilities of results more extreme than those observed, assuming no treatment difference. This calculation requires that the planned design was actually followed, otherwise what is "extreme" changes and so does the P-value, perhaps in an unknown way. In general, deviations from the design invalidate classical statistical inferences.

Practically every clinical trial deviates from its design in one way or another--the most common deviation is probably a different number of patients from that planned. It seems ludicrous not to be able to draw conclusions from data honestly collected. So in calculating P-values, for example, we pretend that the resulting design was the one planned! I see nothing really wrong with this practice. The problem is that many statisticians fail to see that

[*] Research supported in part by the National Science Foundation under grant DMS-8803087.

essentially every P-value that's ever been calculated is necessarily flawed as a measure of inference. A consequence is that when it comes to other closely related but better understood and more openly discussed practices, they take P-values too seriously.

These latter practices are controversial and include adjusting for multiple comparisons, multiple tests, and interim analyses (O'Brien 1983, Berry 1985, 1987, 1988a, 1988b). The analogs of calculations based on the resulting design are P-values that ignore multiplicities; these are called nominal P-values.

Interim analyses are especially appropriate in any discussion of adaptive methods. Accumulating data are analyzed periodically with the possibility of early stopping. But interim analyses must be planned in advance so "more extreme" results can be specified, and the probability of such data calculated under the various hypotheses. If they are not planned then literally correct P-values and literally correct confidence intervals cannot be calculated, even if early stopping did not occur (Dupont 1983). Nominal P-values can of course be calculated. These serve as perfectly fine descriptive statistics. But, as Brown (1983) and Canner (1983) make clear, nominal P-values are irrelevant as measures of inference. We've already seen that essentially every P-value ever calculated is similarly flawed, though perhaps not as openly or obviously. It is splitting hairs to object in some instances and not in others.

The subject of this session is adaptive allocation. Adaptive allocation means that the therapy assigned to the next patient, or therapies assigned to the next group of patients, depend on results obtained thus far in the trial. Most published adaptive allocation procedures tend to assign therapies that have been performing better (for many examples see the Bibliography of Berry and Fristedt 1985).

Current biostatistical practice dictates that analyses of clinical trial data are tied as closely as possible to the trial's design. As I indicated earlier, classical frequentist measures of inference are difficult or impossible to calculate when a trial's design is adaptive--with the accent on "impossible" when allocation is adaptive. This is one of many reasons adaptive allocation is so infrequently used in actual trials. Some of the other reasons given by Simon (1977), Armitage (1985), and Peto (1985), among others, are quite valid. These latter reasons substantially limit the practical usefulness of adaptive allocation methods. But the fact that classical inference is impossible in a legitimate scientific enterprise means to me that we should abandon classical inference rather than abandoning the enterprise!

I will expand on this statement in the next section, showing that classical inference is counter to the scientific method. In Section 3 I will describe how Bayesian inference applies to adaptive designs. And in Section 4 I will give some examples of Bayesian analysis.

2. The Scientific Method and Adaptation

The process of scientific research is given in the following six steps:

1. Ask a question or pose a problem.
2. Assemble and evaluate the relevant information.
3. Based on current information, design an investigation or an experiment (perhaps the null experiment) to address the question posed in step 1. Consider costs and benefits--including information content--of the available experiments. Recognize that step 6 is coming.
4. Carry out the investigation or experiment.

5. Use the evidence from step 4 to update the previously available information; draw conclusions, if only tentative ones.
6. Repeat steps 3 through 5 as necessary.

Questions addressed in clinical research usually deal with the effectiveness of therapies. The set of available experiments includes clinical trials. Costs are in terms of time and resources, just as in other scientific research. But, and this sets clinical research apart from other scientific research, costs also include ineffective medical therapy--for patients in and out of the trial.

In this scientific process, learning takes place as the experimental results accrue. Suppose an experiment can be decomposed into two separate experiments with no additional costs. After the first of these is carried out the available information is updated (suppose at no or negligible cost). Based on this new information the second half of the original experiment may now be unnecessary, or perhaps a radically different next experiment is appropriate. Continue in this way to partition a contemplated "experiment" into its smallest possible pieces, with information updated continuously. There is a net benefit provided by the possibility of deviating from the original plan. (To see this notice that one option that's always available is to stick with the original plan.) This assumes that updating is costless. In clinical trials this assumption is at best approximately true. But the assumption may be reasonable in those clinical trials where the cost of ineffective treatment far outweighs other costs. It also assumes that there _is_ information that accrues during the trial; in some trials the responses are not observed until the trial is over (though in survival studies at least partial information becomes available at each analysis epoch).

The scientific process described here is the motivation behind the recommendations to use adaptive allocation procedures. In standard approaches to RCTs investigators are supposed to close their eyes to accumulating data, and that seems unscientific. Adaptive procedures seem more scientific. But most adaptive allocation procedures are as arbitrary and as unscientific as RCTs. For example, consider the play-the-winner rule: the same therapy is used after a success and therapy is switched after a failure. The investigator has eyes open, but is made to wear glasses that induce extreme myopia. Throwing out all previous knowledge and remembering only the last thing learned is hardly what I mean by updating.

What kind of adaptive procedures are scientific? In deciding which experiment to carry out the investigator should consider costs and benefits explicitly. For the sake of discussion let's restrict cost considerations to effective therapy. The question is, effective for whom? The answer gives rise to the "patient horizon", N, introduced by Anscombe (1963). The patient horizon is the number of patients (in the trial and not) who are in the population being treated and who will eventually be treated with one of the competing therapies. Anscombe describes the solution by dynamic programming. This is consistent with the scientific method. (He describes it in context of adaptive stopping, but the method applies as well to adaptive allocation.) The current experiment is designed knowing that later experiments are possible (cf. step 6). The value of information to be gained in an experiment--information that will help treat later patients--is weighed against the possibility of ineffective treatment of patients involved in the experiment.

The patient horizon is never perfectly known. And it clearly depends on the safety and effectiveness profiles of the competing therapies, which are also unknown. If one of the therapies turns out to be very effective then, while still unknown, N will be larger than otherwise. This makes allocations to the apparently inferior therapy more worthwhile. This effect is ignored by all the adaptive methods I have seen proposed.

It is not my objective to recommend particular adaptive allocation procedures, nor to ~utline a possible role for such procedures in clinical trials. But I venture the opinion that adaptive allocations will never be widely used in clinical trials, and that this is appropriate. When ethical considerations are not primary and N is large, the RCT is quite a satisfactory design (Berry and Eick 1989), but good RCTs are much smaller than those typically carried out. And in those settings where ethical considerations are of prime importance, which can be accommodated by taking N = 1, well-documented open studies are best, and I believe they will be used increasingly (Berry 1989b). Open studies are scientific, but they are at least as problematical for classical inference as are adaptive studies. Bayesian inference may be possible in open studies, depending on the degree of documentation, particularly as regards reasons for treatment assignment.

I want to make one additional point about design and the scientific method. Partitioning an experiment as described earlier means that it is better to rethink the experimental process as frequently as possible. (I'm assuming that thinking is costless--which it's not--and I'm assuming that the thinker is not constrained by an unscientific process of inference.) In particular, large trials that don't allow adaptation are bad, and small trials are good. Stringing together small trials is flexible. The design of the next trial can be based on the results from previous studies, or the experimental plan can be abandoned. Using small studies is globally adaptive. Small studies are frowned upon by classicists (Peto et al. 1976). Making inferences requires analyzing data from various studies, each with its own peculiar characteristics: metaanalysis. The Bayesian approach is ideally suited to this endeavor (DuMouchel 1989). (However, publication bias and other similar biases can make correct inferences difficult or impossible in any approach: if I hide the smallest numbers in a variable sequence from you, and you think you've got the whole sequence, you're not going to do well in guessing how the sequence was generated! Of course, a Bayesian who understands that there may be publication bias will tend to do better than one who does not.)

3. Flexibility of Bayesian Inference

I indicated in the introduction that the problem of multiplicities makes classical frequentist inference unsuited for adaptive designs; this statement applies for other scientifically valid designs as well. On the other hand, the scientific process outlined in the previous section is ideally suited for Bayesian inference. For example, updating one's state of knowledge is a Bayesian notion. Also, step 3 requires evaluating the information content of possible experiments. Information content usually depends on the results of an experiment. Predictive probability distributions of observable results are anathema to classical inference, but they are easily and naturally formulated using Bayesian methods. I don't want to rule out the possibility that there are other approaches that are consistent with the scientific method described in the previous section, but classical frequentist methods are not.

In the Bayesian approach the design used is irrelevant once the data are at hand (Berger 1985; Berger and Berry 1988a, 1988b; Berry 1987, 1988b). Here I mean "data" in the broadest possible sense; in particular, in an open study the data includes all information about the patients available to the clinician who assigned therapy. (The only problem I see with this is the impossibility of quantifying some types of such information. For example, the clinician might sense characteristics of a patient that are difficult to communicate and use as covariates. A possible solution is to have the clinician who assigns therapy be different from the one who diagnoses.)

Consider the following trial. There are two possible therapies, A and B, and two responses, success (S) and failure (F). Patients in the trial are assumed to be exchangeable insofar as their anticipated response is concerned. The following are the results:

Therapy	A	A	A	B	B	B	B	B	B	A	A	A
Response	S	S	F	F	S	F	F	S	F	S	S	S

In the Bayesian approach the only information needed to analyze these results are the sufficient statistics: 5 of 6 successes on A and 2 of 6 successes on B. (The assumption of exchangeability is critical here.) In particular, the design is irrelevant. Many different designs could have produced these data, here are a few:

(i) An RCT planned for 12 patients assigned randomly in blocks of six, three on each therapy.

(ii) Randomized play-the-winner assignment (see Example 1 below for a description) where sampling stops as soon as the absolute difference in sample success proportions is at least 1/2.

(iii) An open study in which the clinician plans to use 3 A's, 6 B's, 6 A's, 6 B's, etc., until concluding that further use of either therapy would be unethical, or until becoming tired.

(iv) An open study in which the clinician assigns therapy in an arbitrary fashion, with some lance in mind, and the data given are interim results.

The only reservation I have about the design affecting my conclusions is that there might be hidden data that would violate the assumption of exchangeability. For example, in design (iv) I would worry that the clinician might have assigned therapy to patients based on covariates to which I am not privy (not that it's wrong to do this, it's just that I want to know about it); had this happened then I could not draw conclusions from the data unless I were told what the covariates were (and perhaps not even then!). Similarly, in design (iii) I would worry that the clinician had juggled the order of admission to, in effect, assign the sicker patients to one of the therapies.

We do need to know the design to calculate P-values (and confidence intervals). For design (i) I get 1P ≐ 0.12 (exact test). For (ii), the probability that A wins if there is no difference in therapies is 1P = 1/2. P-values cannot be calculated for designs (iii) and (iv).

4. Examples

In this section I will illustrate Bayesian analyses of adaptive and other studies in the context of examples. The examples I give are clinical trials with dichotomous responses; the ideas generalize easily to other types of trials.

Example 1 (Bartlett et al. 1985)

This is one of the few clinical trials in which adaptive allocation has been used. The analyses I present here are far from the final word. More complete analysis is forthcoming in Berry and Hardwick (1989). A randomized play-the-winner scheme was carried out as follows. An A and a B were placed in an urn (figuratively speaking). One was selected randomly and the corresponding therapy administered, A = experimental (ECMO) and B = control (conventional therapy). If the response was survival (S) then the treatment letter was replaced in the urn and another letter of the same type was added--response time was effectively immediate. If the response was death (F) then the treatment letter was replaced and a letter of the other type added. Stopping was to have taken place when ten balls of either type had been added to the urn. The second phase of the study was to be nonrandomized, with all patients assigned to the therapy that performed better in the first place.

The responses reported by Bartlett et al. were as follows:

Therapy	A	B	A	A	A	A	A	A	A	A	A	A
Response	S	F	S	S	S	S	S	S	S	S	S	S

(Note the deviation from the stopping rule.) After the trial, 8 more patients were administered A and all survived, and 2 more were administered B and both died.

Suppose the patient population is homogeneous, so the patients are regarded to be exchangeable. Let p_A and p_B be the probabilities of success on treatments A and B. (In the next example I will describe a model in which these probabilities depend on the patients' prognoses.)

Taking the classical frequentist point of view, Ware and Epstein (1985) observe that the Bartlett et al. trial had a "50% false positive rate", or type I error rate: if the null hypothesis $p_A = p_B$ is true, then the probability of obtaining 10 more A's than B's is 1/2. They say this rate is "unacceptably high". This is an instance of what I mean by taking hypothesis testing too seriously; in particular, it applies no matter how strongly the actual data favors either therapy. Ware and Epstein conclude: "Further randomized clinical trials using concurrent controls...will be difficult but remain necessary." (Hence the study described in Example 2.)

A Bayesian approach requires a prior distribution on (p_A,p_B). For illustrative purposes only, suppose this is uniform. Such an assumption is consistent with assuming the treatments to be exchangeable and independent a priori, with little information available about either. (None of these assumptions is correct--see below.) The posterior density (on the unit square) given data from the trial is then

$$f(p_A,p_B) = 24\, p_A^{11}(1-p_B).$$

Consider the *conditional relative improvement* due to ECMO (compared with conventional therapy): p_A-p_B. Define the (unconditional) *relative improvement* to be the probability that this is greater than c:

$$RI(c) = \int_c^1 \int_0^{x-c} f(x,y)dx$$

$$= \frac{90}{91} - \frac{2}{13}c - c^2 + \frac{2}{13}c^{13} + \frac{1}{91}c^{14}.$$

This is labeled with an asterisk in Figure 1; in particular, the posterior probability that ECMO is better than conventional therapy is $RI(0) = 90/91$.

Consider also the 10 patients reported by Bartlett et al. that were treated after the trial. According to the protocol, all 10 should have been assigned to ECMO. My understanding is that all met the eligibility criteria for the study but the ECMO device was not available for the two who were assigned to conventional therapy. Considering these 10 to be exchangeable with the patients in the trial means that

$$f(p_A,p_B) = 80p_A^{19}(1-p_B)^3.$$

The relative improvement function, RI(c), for this density is labeled with a double asterisk in Figure 1; now $RI(0) = 0.9999$.

Bartlett et al. claim that the patients in the trial would have had at least an 80% death rate on conventional therapy. A Bayesian analysis can incorporate historical controls (Berry and Hardwick 1989)--indeed, the scientific method requires using all available information. But in an ostensibly scientific report, any such statement should be backed up by evidence. In this instance the issue is critical. If p_B is known to be 0.2, say, then

$$RI(c) = \int_{.2+c}^1 f(p_A)dp_A;$$

this is $1-(.2+c)^{12}$ for 11 successes out of 11 patients on A, and $1-(.2+c)^{20}$ for 19 successes out of 19 on A. These are shown in Figure 2, using the same labeling system as in Figure 1. The relative improvement of A is dramatic under this assumption. For example, in the second case, $RI(0.5) > 0.999$, so ECMO is very likely to save an additional 50% of the patients as compared with conventional therapy.

The patients in this trial were not actually exchangeable. (Not incidentally, the patient who received conventional therapy in the randomized phase happened to be the sickest of the 12.) Berry and Hardwick (1989) carry out a Bayesian analysis accounting for the patients' characteristics, as well as incorporating historical controls.

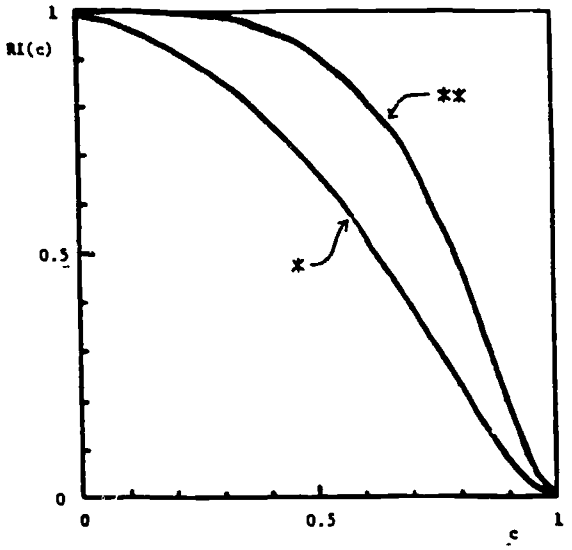

Figure 1. Relative improvement for A over B assuming independent uniform priors for p_A and p_B.

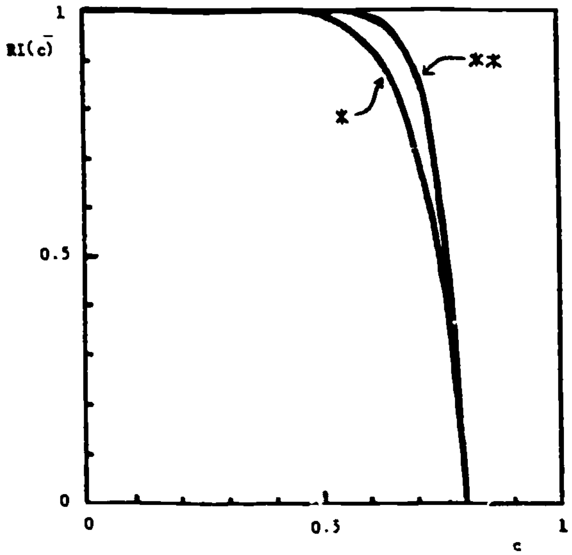

Figure 2. Relative improvement for A over B assuming $p_B = 0.2$ and uniform prior for p_A.

Example 2 (Ware 1989)

Experimental (A) and control (B) were the same as in Example 1. The trial was in two phases. Phase 1 was balanced randomized and would stop when either therapy accumulated 4 deaths. The other therapy would be used exclusively in phase 2, which would end when this other therapy accumulated a total of 4 deaths. (In view of the Bartlett study (Example 1) and other available information on ECMO and conventional therapy, I think this trial--or any trial randomizing to conventional therapy--was unethical; cf. Berry 1989b.)

The results are shown in Table 1. Note that phase 2 stopped with only one ECMO death. See Ware (1989) for the way he gets around this obvious stumbling block for classical inference.

Table 1. Data from Ware (1989)*

	Patient number	Treatment	Initial prognosis**	Response	$P(p_A > p_B \mid \text{data})$
	1	A	0.754	S	0.59
	2	B	0.695	S	0.48
	3	A	0.899	S	0.51
	4	B	0.747	S	0.46
	5	B	0.720	F	0.71
	6	A	0.882	S	0.74
	7	A	0.886	S	0.77
	8	B	0.842	S	0.74
PHASE 1	9	B	0.937	S	0.73
	10	B	0.844	F	0.87
	11	A	0.874	S	0.88
	12	A	0.877	S	0.90
	13	B	0.788	F	0.950
	14	A	0.902	S	0.956
	15	A	0.922	S	0.961
	16	B	0.826	S	0.949
	17	B	0.874	S	0.940
	18	A	0.871	S	0.948
	19	B	0.838	F	0.974
	20	A	0.900	S	0.978
	21	A	0.716	F	0.918
	22	A	0.960	S	0.922
	23	A	0.902	S	0.931
	24	A	0.826	S	0.943
	25	A	0.801	S	0.954
	26	A	0.854	S	0.960
	27	A	0.874	S	0.965
	28	A	0.774	S	0.971
PHASE 2	29	A	0.941	S	0.973
	30	A	0.615	S	0.981
	31	A	0.825	S	0.984
	32	A	0.865	S	0.985
	33	A	0.775	S	0.988
	34	A	0.832	S	0.989
	35	A	0.792	S	0.990
	36	A	0.874	S	0.991
	37	A	0.770	S	0.992
	38	A	0.735	S	0.994
	39	A	0.921	S	0.994

*The order of patient reponses and covariates used to calculate prognoses are not given in Ware (1989); Professor Ware was kind enough to provide these to me.

**Predicted probability of success on treatment A from Toomasian et al. (1988).

Table 1 also shows the individual patients' prognoses. The method of computing these was taken from Toomasian et al. (1988) who report on a national registry of 715 ECMO cases and calculate a logit for survival (= success). Their model is

$$\log\left(\frac{x}{1-x}\right) = 20.054 - .918(\text{birthweight}) - 2.465(\text{pH}) - .386(\text{MAS}) + .597(\text{renal failure}) + .304(\text{female}).$$

The last three variables are indicators; MAS means meconium aspiration syndrome as primary diagnosis; renal failure was defined as creatinine ≥ 1.5. Since I did not have access to the last two variables, I used only the first four terms. (Dropping the last two would have no effect if all the patients were male and none had renal failure--about 10 percent of the 715 cases reported by Toomasian et al. had renal failure.)

There was evidence that ECMO was more effective than conventional therapy--see Example 1 and Ware (1989). But I calculated $RI(0) = P(p_A > p_B|\text{data})$ in Table 1 assuming that A and B were exchangeable initially, and using a technique proposed by Berry (1989a) with $\sigma = 2$. All previously treated patient responses, treatments, and prognoses are included in "data". Such a measure can be calculated at any time during the trial, even if it may result in early stopping, without compromising the eventual conclusions (Berry 1985, 1987).

The probabilities in Table 1 are not P-values. Rather, they have a direct interpretation concerning the two therapies. Namely, $P(p_A > p_B|\text{data})$ is the probability that therapy A is the better treatment to assign to the next patient.

The probabilities in Table 1 assume tha the prior distribution of (p_A,p_B) remains unchanged during the trial. Any evidence that becomes available from outside the trial can be used to update the current distribution of (p_A,p_B).

The ECMO patients in phase 1 had better prognoses (on ECMO therapy) than did their counterparts on control: averages of 0.126 and 0.189, respectively. So the probabilities in the rightmost column of Table 1 are larger than they would be had the covariates been ignored.

In clinical trials in which one therapy is used exclusively for a period of time (phase 2 in the example), one worries that there may be a time trend in the patient population which then is confounded with treatment. (Indeed, this is a standard argument against using adaptive allocation.) The calculations shown in Table 1 adjust for any time trends that are manifest in the covariates used to calculate prognoses. Of course, it does not account for "silent" covariates. (The average prognosis in phase 2 is 0.173, giving an overall average of 0.158 for ECMO patients, so there seems to be at most a slight time trend in the example data.)

Table 2 gives the updated prognosis of patients on A and B using maximum likelihood (see Berry 1989a). The fact that $\hat{p}_A(x) < x$ means that the ECMO patients in the current study had better results than did their counterparts in the national registry. (This difference cannot be the result of dropping "renal failure" and "female" from the logit model since their coefficients are positive.)

Publishing the results of this trial can include an updated prognosis for both therapies, such as given in Table 2. If RI(c) or any other characteristic of the current distribution of (p_A, p_B) is published, the prior distribution of (p_A, p_B) should also be published. It is also incumbent on the authors to indicate the sensitivity of the current distribution to the prior with, perhaps, an indication of what the current distribution would be for different priors.

Table 2. Updated prognosis $\hat{p}(x)$ based on data from Table 1; x = initial prognosis.

x	$\hat{p}_A(x)$	$\hat{p}_B(x)$
0.95	0.991	0.863
0.90	0.980	0.750
0.80	0.956	0.571
0.70	0.928	0.437
0.60	0.892	0.333
0.50	0.846	0.250
0.40	0.785	0.182

Therapies A and B are very different. ECMO is radical, invasive therapy whose use could itself result in death. So it seems reasonable to assume $P(p_A = p_B) = 0$, as I have done in this example. But this assumption seems less appropriate in most settings, in particular, in that of the next example.

Example 3 (Dixon et al. 1989)

This is a balanced randomized trial comparing two treatments for adult acute leukemia: A = amsacrine/cytosine arabinoside and B = mitoxantrone/cytosine arabinoside. The responses are reported in Table 3. Success (S) is complete remission and failure (F) is any response other than S. Initial prognosis in the probability of complete remission based on a logistic model. The calculation of $P(p_A < p_B|\text{data})$ uses Berry (1989a), as in Example 2.

The stopping rule used by Dixon et al. (1989) was based on a Bayesian calculation after pairing A and B patients on the basis of prognosis. Their method has the advantage of being easy to understand: 4 of 8 preferences for A with the other 4 pairs tied. The method of Berry (1989a) does not assume exchangeability of pairs, and it does not require matching patients on prognosis. (Berry (1989a) gives an extension of the method to analysis of survival times with the possibility of censoring.)

Table 3. Data from Dixon et al., (1989)

Patient number	Treatment	Initial prognosis*	Response	$P(p_A>p_B \mid \text{data})$
1	B	0.93	S	0.46
2	A	0.78	S	0.54
3	B	0.59	F	0.77
4	A	0.44	S	0.89
5	B	0.81	S	0.84
6	A	0.68	S	0.88
7	B	0.87	S	0.86
8	A	0.87	S	0.87
9	B	0.49	F	0.933
10	A	0.78	S	0.945
11	B	0.87	F	0.982
12	B	0.74	S	0.971
13	A	0.59	S	0.982
14	A	0.50	S	0.989
15	B	0.40	F	0.993
16	A	0.93	S	0.994

*Predicted probability of success.

5. Conclusion

Scientific research is planning and learning. Learning is adaptive. The scientific method prescribes how learning takes place efficiently. Bayesian inference is consistent with the scientific method. In particular, it is an ideal prescription for learning. Classical frequentist inference is inconsistent with the scientific method.

Adaptive allocation may not have a place in medical research Trials in which there are no ethical concerns are perhaps best carried out with randomized, concurrent controls. But these trials should be small. This allows for global adaptivity, rethinking and modifying strategies between trials, which can save time, resources, and increase the chance of delivering effective medical therapy to more people.

When ethical concerns rule out RCTs, treatment should be assigned in an open fashion, with patients followed to ascertain effect. Correct inferences are difficult in open studies, at least in part because of the possibilities of bias in assigning treatment. Classical frequentist methods are not available; Bayesian inferences may be possible, but adjusting for covariates is essential. These inferences will be better if based on control information from historical data. Appropriately weighing historical data is one of the biggest challenges in the analysis of clinical trials.

References

Anscombe, F.J. (1963). Sequential medical trials. Journal of the American Statistical Association 58, 365-383.

Armitage, P. (1985). The search for optimality in clinical trials. Int. Statist. Rev. 53, 15-24.

Bartlett, R.H., Roloft, D.W., Cornell, R.G., Andrews, A.F., Dillon, P.W., and Zwischenberger, J.B. (1985). Extracorporeal circulation in neonatal respiratory failure: A prospective randomized trial. Pediatrics 76, 479-487.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag.

Berger, J.O., and Berry, D.A. (1988a). The relevance of stopping rules in statistical inference (with discussion). Statistical Decision Theory and Related topics IV 1, 29-72. (Ed. by J.O. Berger and S. Gupta.) New York: Springer-Verlag.

Berger, J.O., and Berry, D.A. (1988b). Statistical analysis and the illusion of objectivity. The American Scientist 76, 159-165.

Berry, D.A. (1985). Interim analysis in clinical trials: Classical vs. Bayesian approaches. Statistics in Medicine 4, 521-526.

Berry, D.A. (1987). Interim analysis in clinical trials: The role of the likelihood principle. American Statistician 41, 117-122.

Berry, D.A. (1988a). Multiple Comparisons, Multiple Tests, and Data Dredging: A Bayesian Perspective (with discussion). In Bayesian Statistics 3, 79-94. Oxford, England: Oxford University Press. (Edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith.)

Berry, D.A. (1988b). Interim analysis in clinical research. Cancer Investigation 5, 469-477.

Berry, D.A. (1989a). Monitoring accumulating data in a clinical trial. Biometrics. (To appear.)

Berry, D.A. (1989b). Ethics and ECMO: Comments on a paper by Ware. Statistical Science. (To appear.)

Berry, D.A., and Eick, S.G. (1989). Decision analysis of randomized clinical trials: Comparison with adaptive procedures. (Submitted for publication.)

Berry, D.A., and Fristedt, B. (1985). Bandit Problems: Sequential Allocation of Experiments. New York: Chapman-Hall.

Berry, D.A., and Hardwick, J. (1989). Using historical data in a Bayesian analysis of clinical trials: Application to ECMO. (In preparation.)

Brown, B.W., Jr. (1983). Comments on the Dupont manuscript. Controlled Clinical Trials 4, 11-12.

Canner, P.L. (1983). Comment on 'Statistical inference from clinical trials: choosing the right P-value'. Controlled Clinical Trials 4, 13-17.

Dixon, D.O., Gehan, E.A., Estey, E.H., and Smith, T.L. (1989). Bayesian stopping in a randomized trial. (Unpublished.)

DuMouchel, W.H. (1989). Bayesian metaanalysis. Statistical Methodology in the Pharmaceutical Sciences, 511-531. (Ed. by D.A. Berry.) New York: Marcel Dekker.

Dupont, W.D. (1983). Sequential stopping rules and sequentially adjusted P-values: does one require the other? Controlled Clinical Trials 4, 3-10.

O'Brien, P.C. (1983). The appropriateness of analysis of variance and multiple-comparison procedures. Biometrics. 39: 787-788.

Peto, R. (1985). Discussion of papers by J.A. Bather and P. Armitage. Int. Statist. Rev. 53, 31-34.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. Br. J. Cancer 34, 585-612.

Simon, R. (1977). Adaptive treatment assignment methods and clinical trials. Biometrics 33, 743-749.

Toomasian, J.M., Snedecor, S.M., Cornell, R.G., Gilley, R.E., and Bartlett, R.H. (1988). National experience with extracorporeal membrane oxygenation for newborn respiratory failure. Trans. Amer. Soc. Artificial Internal Organs 11, 140-147.

Ware, J.H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). Statistical Science. (To appear.)

Ware, J.H., and Epstein, M.F. (1985). Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. Pediatrics 76, 849-851.