THE ROLE OF SUBJECTIVITY IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

by

Donald A. Berry*
University of Minnesota

Technical Report No. 490
May 1987

* Research supported in part by the National Science Foundation under grant DMS 85-05023. Invited paper for Biometrics Section of the 1987 Joint Statistical Meetings, San Francisco, August 17-20.

Abstract

Subjectivity plays a role in the design and analysis of every clinical trial. I discuss its present role (for example, in the context of analyzing at haphazard times as data accumulate), and suggest ways that subjectivity could be incorporated in an explicit way and in a variety of settings. One such setting is a trial sponsored by a pharmaceutical company in which accumulating data can be used to update previously available (subjective) information. Another is in planning and analyzing trials which use adaptive (or sequential) treatment assignment schemes. Another is in planning and analyzing an ethical clinical trial, one in which each patient receives the treatment best suited to him or her.

Introduction

Most biostatisticians involved in clinical trials strive to ensure that a trial's design, conduct, and analysis are objective. This goal is unattainable: objectivity is impossible in science generally. I don't want to get too deeply into philosophical issues so I won't expand on this. Instead, I'll tell you a few of the ways that subjectivity creeps surreptitiously into current clinical trials, and then I'll focus on the advantages of recognizing the presence of subjectivity and show how it can be exploited to best accomplish the purpose of the trial.

The design of a clinical trial should depend on its purpose. Generally speaking, there is only one legitimate reason to conduct a clinical trial: to deliver effective medical treatment. The question is, to whom? There are essentially two answers:

- Patients in the trial, and
- Patients who are treated outside the context of the trial based on information learned during the trial.

Randomized clinical trials (RCTs) are consistent with the second to an extent, but they address the first only in that investigators will stop a trial should it become sufficiently clear that patients in the trial are being treated badly. But the trial continues when there are suggestions in that direction without being "sufficiently clear" that some patients are being treated badly. On the other hand, the trial may end when the evidence does not begin to suggest which treatment is best.

Sample sizes for RCTs are calculated on the basis of power considerations (a process which, incidentally, is far from being objective). RCTs have inflexible designs, most with treatment allocation balanced among the constituent treatments. Neither of these aspects has much to do with delivering good medicine--whether to patients in or out of the trial.

RCTs are saddled with the straightjacket of classical statistics. The classical design of clinical trials is tied to the eventual analysis. This results in a very inflexible view of design. The design must be specified completely in advance, including what analyses will be done and when they will be done; changes are formally impossible and in practice require excuses and approximate analyses, with little ability to assess the accuracy of the approximations.

Inflexibility of the Classical Approach

One of the most objectionable aspects of the classical approach is that it gives essentially the same prescription for every malady. Consider two extremes. The first is coronary bypass surgery compared in a clinical trial with a treatment combination consisting of a regimented diet, exercise and drugs. The second is an experimental drug compared with placebo in the treatment of a rare form of cancer (say there are only 100 new patients in the U.S. per year). In the first setting there are millions of people who could benefit from the results of the trial; while in the second setting, all the patients in the U.S. who contract the disease in the next several years (depending on n) will be in the trial. (The patient is told that the only possibility of receiving the "new treatment" is to participate in the trial.) Assume in both cases that we're interested in the same end point, say five-year survival. Then the same power calculations apply for both. Suppose we find the required sample size to be $n = 400$. In the first setting, 200 patients are randomly assigned to surgery and the other 200 to the diet/exercise program; in the second, 200 are randomly assigned to drug and the remaining 200 to placebo.

There are at least two reasons that these settings should not be dealt with the same. First, the information gained from the trials will have very different impacts. What is learned in the first setting will apply to millions, and so there is much to be gained (in terms of good medical treatment for these millions) from knowing which is the better therapy. But what is learned in the second will apply to very few patients. Actually, it may apply to no patients: the trial will last about four years, and new, obviously superior treatments may well be discovered during that time. In the interim, as many as 200 patients may be treated with an inferior treatment for no worthwhile purpose. Statistical power is obtained at the expense of effective treatment of some of the patients in the trial; in some cases it may be worth the expense, and in others not.

The second reason for dealing with the two settings differently is that the available information is so different. Much is known concerning the effectiveness of coronary bypass surgery and of diet/exercise. In particular, if one is better than the other (in a population of patients who are candidates for coronary bypasses), it is not markedly better. But little is known concerning the effectiveness of a new drug in the treatment of a rare disease. It is conceivably very effective, but it might be distinctly worse than no treatment. So perhaps $n = 1000$ may be required in the first case while $n = 10$ or 20 is sufficient in the rare disease setting; or better, in both cases the data might be examined continuously with no particular trial size in mind! (I'll return to this latter possibility later.)

Small trials have small power: they are able to reject the null hypothesis of no difference with high probability only if one treatment is much better than the other. Classical statisticians complain about trials too small to have a reasonable "chance of detecting differences of therapeutic value" (Mosteller, Gilbert and McPeek, 1983). They contend that many actual trials are too small to be worthwhile. Some even believe that it is unethical to conduct a small trial since some of the patients will be exposed to inferior treatment with little hope of rejecting a false null hypothesis. This is true, but it's not the point. Such a piecemeal approach allows clinicians to digest the information currently available and perhaps decide that further investigation is inappropriate--the experimental treatment may be clearly bad or clearly good. Classical statisticians cannot advocate such an approach because the inverted nature of classical inference makes it woefully inept at combining results from different trials--called meta-analysis by some. Indeed, for a formally correct classical analysis of multiple trials, all the trials have to be designed in advance of the first trial, complete with rules for stopping, initiating other trials, etc. This is obviously impossible, and so, strictly speaking, classical meta-analysis is impossible.

In my opinion, small trials are the rule in medicine today because the associated flexibility is so important to clinicians; they bypass statistics as they know it, with its obscure p-values, in favor of the important inferences: Does this drug work?, Is drug A better than drug B for Mr. Smith?, etc. They make these inferences in an informal, subjective way. (An unfortunate consequence is that they come to view statistics as data analysis and of little help in making medical decisions.)

Classical Interim Analysis

From the point of view of classical statistics, a piecemeal approach is possible in a single trial by splitting the trial into subgroups of patients, provided the options and intentions of the investigator are completely specified in advance. Using a "group sequential" approach, the available data are analyzed at various times during the course of the trial. In Berry (1985, 1987a), I argue against such an approach in very strong terms. The approach harbors a dangerous kind of subjectivity. For example, consider two investigators: A plans interim analyses and B plans none (or says he planned none!). They get identical data. It turns out that B can claim statistical significance, but, because she looked at the data with the possibility of early stopping, A cannot. To avoid A's dilemma, classical statisticians warn investigators not to look at accumulating data because doing so endangers the ability to draw conclusions--sounds crazy!

What about the investigator who does not plan interim analyses but makes them nonetheless, looking at the accumulating data periodically with the possibility of stopping? Strictly speaking, no classical inferences are possible. For example, consider p-values. A p-value is the probability of a result more extreme than that observed. But the investigator has no hope of being able to say accurately what results would have been more extreme at earlier looks and also at subsequent looks that never took place, as well as when they were to take place. In reality, few investigators have even the vaguest notion that merely looking at the data can affect inferences drawn. Most feel that looking is necessary in case the interim results indicate that the trial should be halted or modified, and that not looking is unethical. (Using logic that completely escapes me, some doctors and statisticians argue that investigators should not look at interim data because they will face an ethical dilemma should one of the treatments appear to be better than the other.

If there are possible data for which continuing the trial would be unethical, then not looking seems unethical: Would it not be unethical for a doctor to refuse to read a medical report that might contain vital information for the treatment of a patient?)

The Flexibility of the Bayesian Approach

Clinicians usually have information that can help design a clinical trial. This information can be used in a completely explicit way to design a trial that effectively treats patients in or outside the trial, or both. What is required is that the clinicians quantify their information into probability distributions of the various unknown parameters. (There are numerous published methods which aid in doing this.) These parameters include the degrees of effectiveness of the various drugs and the size of the patient population that will benefit from knowledge gained during the trial. (I want to stress that the clinicians do not have to know or to think they know these parameters, they need only be able to convey the information they have, even if it's limited.)

Assessing probability distributions of the various parameters means that these can be updated at any time using Bayes's theorem. There are two appealing aspects of this approach that are in stark contrast with the classical approach. First, probability statements are direct. For example, while a p-value is the probability of results as or more extreme than those observed assuming the null hypothesis, a Bayesian approach allows one to give the probability that the null hypothesis is true given the data. Secondly, these direct probability statements do not depend on.the trial's design. As such they can be calculated at any time and for any purpose, even to determine the future course of the trial! I turn to an extreme example of this.

Effective Treatment for Patients in the Trial

Consider a trial involving two treatments for a rare disease. All patients who contract this disease in the next several years are to be entered into the study. Say they number n. I shall assume for convenience that n is known, but it need not be. Since all patients with the disease are in the trial, the objective is to treat these n as effectively as possible. Assume that responses are success-failure and that each patient responds before the next patient is to be treated. (Eick (1985) considers the more realistic setting in which a patient's response is a survival time and so at any time only partial information may be available from previously treated patients.)

The treatment assigned to patient i is allowed to depend on the information available at the time of treatment; this information includes the responses of patients 1, 2, ..., i-1. (Obviously, this information is available only if the earlier patients have responded.) A treatment assignment strategy is a sequence of n A's and B's such that the i^{th} symbol (indicating the treatment used for patient i) can depend on the first i-1 treatment assignments and responses. Each treatment assignment strategy has an associated expected number of successes among the n patients. A strategy which maximizes that expected number of successes can be found via backward induction (Berry and Fristedt 1985). Quite generally, an RCT is an unsatisfactory solution of this problem.

Effective Treatment for Patients Outside the Trial

Now let's leave the rare-disease setting and suppose that there are patients with the condition in question outside the trial. Suppose that N is the number of patients with the disease who are in the trial or who will be given the treatment found to be best among those in the trial. This number is obviously unknown, and assessing its distribution is no mean task, but in practice only a rough estimate is required. Suppose n of the N patients are in the trial. After the n patients in the trial respond, the remaining N-n will be assigned the treatment which has the greater probability of success at the end of the trial. Again the problem is a bandit and can be solved using backward induction. With these assumptions, Berry and Eick (1987) find the optimal strategy for various n and N and compare it with other assignment strategies, including RCTs, on the basis of expected number of successes lost as a function of the (unknown) probabilities of success. Summarizing the results, an RCT is quite unsatisfactory when N is small, but if N is at least moderately large, an RCT, while not optimal, is a very reasonable solution.

Ethical Clinical Trials

Is there such a thing? Not in the context of a classical approach to statistics. But they are quite possible in the Bayesian approach. I'll describe one way that an ethical trial can be conducted (see Berry (1987b) for a critical discussion, and technical details). You'll have many reservations; those related to difficulties in making classical inferences such as p-values do not concern me.

To be specific consider a trial involving breast cancer. There are at least five types of therapy available, though they have many variants and certain combinations are possible: mastectomy, lumpectomy, chemotherapy, radiation, and no treatment. A patient is admitted to the trial. The clinician evaluates her condition and considers other relevant information (for example, the patient's age and general state of health). The clinician assesses the patient's prognosis for each available therapy and combination thereof. These assessments are based on all the information available to the clinician: data concerning the various therapies published in the literature and in the clinician's experience, including the current trial!, the patient's condition, etc. These assessments are subjective and are modified continually (or as often as possible) using Bayes's theorem.

The clinician informs the patient of these assessments for each possible combination of therapy. The patient is given complete information: probability of remission (for various lengths of time), probabilities of various side effects and adverse experiences, cosmetic consequences, the patient's responsibilities, inconveniences, costs, etc. In addition, the patient can be told the various components of the clinician's probabilities; for example, the proportions of the clinician's own patients similar to the current patient who are still in remission. The only thing she is not given is the clinician's recommendation of treatment. Armed with all this information the patient chooses her therapy.

The patients are followed as in current trials. Unlike many current trials the patient can change her mind based on the course of.her disease, new data that have come to light, etc., and opt for a modification of her therapy. The data base for the trial is kept current to enable informed later treatment.

The results can be published at any time, even periodically. Classical statistical inferences are impossible. But Bayesian posterior probability distributions can be published (along with the sufficient statistics to separate out the "subjective" part of these distributions); there is never a penalty for interim analysis in the Bayesian approach--the data are taken at face value. In addition, the clinician can calculate and publish such quantities as the posterior probability ("current" is a better modifier because these probabilities are constantly subject to change) that mastectomy + chemotherapy + radiation, for example, is better than no treatment for a typical patient.

Perhaps your biggest objection at this point is that treatment allocation is bound to be unbalanced since similar patients will tend to select the same treatment. So the ability to make certain inferences can be severely limited. There's no arguing with this. It is not possible to have a trial that is both ethical and guaranteed to be balanced. Still, there can be a substantial amount of evidence available even with great imbalance. For example, suppose there are 106 patients in a trial. These patients were regarded as exchangeable before treatment--they all had the same values of any covariates, for example. During the course of the trial the data seem to suggest that treatment A is better than B and so it is chosen by most of the patients: 98 to 8. It turns out that there are 93 successes of the 98 on A and 3 successes of the 8 on B. Then the probability that B is better than A (assuming independent uniform priors on the two probabilities of success) is only 0.00002.

There are complications that will arise in any real trial. I address many of these in Berry (1987b). My point here is that ethical trials are possible by considering subjectivity in a completely explicit way.

A Pharmaceutical Company Decision Problem

Until now, I've been discussing clinical trial design from the point of view of treating patients effectively; in particular, I've ignored monetary considerations. I now turn to a setting in which monetary considerations are primary, namely, planning a drug development program for a pharmaceutical company. I want to make it clear, however, that maximizing profit for pharmaceutical companies is not inconsistent with delivering good medicine.

Only a very short-sighted company would market a drug it knows to be ineffective or unsafe. A company that markets an ineffective drug risks losing marketing and other developmental costs; in this age of litigation, marketing an unsafe drug risks being forced into bankruptcy.

Consider a pharmaceutical company that is developing an experimental drug. It has spent a great deal of money on the drug and has to decide whether to spend still more. If the company continues development and the evidence shows that the drug is safe and effective, then it will eventually try to obtain regulatory approval for marketing the drug. Even if it succeeds in marketing the drug, it may actually lose money, depending on the drug's effectiveness (and side effects). If the company stops development then, of course, it will lose whatever profits were possible. The question is, As a function of current information, should the company continue or stop development?

I'll address this question from three points of view: (1) the status quo, (2) using classical statistics, and (3) using subjectivity in a Bayesian approach. As to (1), such decisions are usually made as follows. A team headed by a company executive (usually an M.D.) uses an approach roughly similar to (3), but in a very informal way, a way filled with perils. They examine the available information, assess the chances of regulatory agency approval, and evaluate the market. The executive makes the final decision with input from the team. There are statisticians on the team but they play a rather minor role in the decision process. The greatest peril in this process is what business analysts call entrapment: the executive has made previous decisions to continue development, so to stop development now is to admit that those previous decisions were wrong. The executive is "trapped" and, according to Staw (1981), is much less likely to change course than would someone who was not previously involved.

I can dispose of (2) even more quickly. Statisticians in the pharmaceutical industry in the U.S. are mainly non-Bayesians. The reason statisticians don't have much input under the status quo is that classical statistics simply cannot address questions such as, How likely is it that the drug is ineffective? A decision maker must be able to answer this and similar questions. Such questions require a subjective interpretation of probability.

I'll devote the rest of my discussion of this problem to a Bayesian approach. Much of what I say is along the lines of the analysis in Berry and Ho (1987). Assume that the experimental drug is being compared with a control in a clinical trial. The trial has a parallel design with about the same number of patients receiving drug as receive control. The trial's costs are assumed to be proportional to the number of patients involved. The setup is similar to the classical problem of interim analysis. The data are to be examined periodically during the course of the trial. (More generally, the "trial" can be viewed as a drug development program comprising various trials, possibly taking place concurrently. In this case the periodic examinations can occur during trials or between trials.)

At these periodic analyses, the question the pharmaceutical company addresses is: Should it continue or stop the trial? If the interim results are very positive then the trial and drug development will continue. But if the interim results are sufficiently negative then the trial and further development will cease.

The company carries on a collateral process of statistical analysis for convincing the regulatory authorities that the drug is safe and effective; at least in the United States, this analysis has to be classical. The maximal trial size is selected to be convincing to the authorities. Premature stopping can occur for negative results but not for positive results. So the appropriate adjustment in one-sided P-values is negligible (Ho, 1986), and is downward in any case.

The company's objective is to maximize profit. Use δ to denote the average advantage of the drug over control. The expected profits from marketing the drug depend on δ. The company assesses this functional dependence, averaging over any unknown parameters. The initial (subjective) distribution of δ is also assessed.

Suppose the total number of analyses is k. (The only reasons for not analyzing after each patient responds are logistical and not statistical.) At the final analysis the decision will be made whether to pursue marketing, and at the k-1 interim analyses the decision will be made whether to continue. Each period includes 2n patients, n on the experimental drug and n on control. A particular model for the responses is assumed, such as normally distributed with known variance.

This is a typical problem in dynamic programming. We calculate for what data the decision to market is optimal at the final analysis. We do this by comparing the profit from marketing averaged with respect to the posterior distribution of δ, with the expected profit from stopping (we can ignore sunken costs). The maximal expected profit for each datum is then the greater of these two quantities. Then we back up to the penultimate analysis time, calculate the predictive distribution of the future data given the present, and evaluate the expect profit from continuing (including -2n in sample costs). If the expected profit from continuing is greater than that of stopping then it is optimal to continue <u>if we ever find ourselves with these data</u>, and we would stop otherwise.

Proceeding backward in this way we pass through each analysis time and end up at the first one. We then know the optimal decision (stop or continue) at each analysis time and for all possible data. In particular, we know whether it is optimal to start the trial.

Conclusion

I have shown several settings in which the flexibility gained by explicitly using subjectivity allows for better fulfilling the aims of clinical trials. But my greater message is to think hard about the purposes of the clinical trial during the design phase. Who is going to be harmed by your design and who is going to be helped--is the sacrifice warranted? The first constructive thing for you to do is to throw out your tables of power vs. sample size!

References

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

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

Berry, D.A. (1987b). Designing ethical clinical trials. Submitted for publication.

Berry, D.A. and Eick, S.G. (1987). Adaptive vs. balanced randomized designs in clinical trials: A decision-theoretic approach. Submitted for publication.

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

Berry, D.A. and Ho, C.-H. (1987). One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. Submitted for publication.

Eick, S.G. (1985). Sequential Experimentation with Delayed Responses. Ph.D. dissertation, University of Minnesota, Minneapolis, Minnesota.

Ho, C.-H. (1986). One-sided Sequential Stopping Boundaries for Clinical Trials: Classical and Bayesian Approaches. Ph.D. dissertation, University of Minnesota, Minneapolis, Minnesota.

Mosteller, F., Gilbert, J.P. and McPeek, B. (1983). Controversies in design and analysis of clinical trials. In Clinical Trials, 13-64, edited by Shapiro, S.H. and Louis, T.A. New York: Marcel Dekker, Inc.

Staw, B.M. (1981). The escalation of commitment to a course of action. Academy of Management Review 6, 577-587.