the equal allocation probabilities used initially in the Harvard study.

Cornell also presented an expression which accelerates the divergence of the allocation probabilities as the results in favor of one of the treatments becomes more pronounced. It is a function of the number of balls of each type in the urn and has the allocation formula of Wei and Durham for simple random selection as a special case. Acceleration would tend to compensate for a large initial value of u in terms of its effect on the expected number of patients allocated to the inferior treatment. It is this acceleration feature which distinguishes Cornell's proposal and makes it a viable alternative to the design presented by Ware and used in the Harvard ECMO trial.

The last suggestion made by Cornell concerns rejection of the null hypothesis of equal success probabilities on the control and the new treatment when more balls for the new treatment are added to the urn. A balls is added to the urn for one type of treatment whenever success is obtained on that treatment, or failure on the other. Cornell proposed that the null hypothesis be rejected only if the number of balls added for the new treatment exceeds that for the control to the extent that the posterior probability of correct selection, denoted by PPCS and conditional on observed frequencies, is high. This along with selection of a large value of u would enable the results of a randomized-play-winner trials to be used for hypothesis testing as well as for treatment selection.

An analysis based upon the PPCS would apply for any adaptive randomization scheme that depends only on the observed results without knowledge of the identity of the two treatments. It is a posterior probability in that it conditions on the observed allocations and frequencies of success on the two treatments. It is not a posterior probability in the sense of Bayesian inference, but is a function of the success probabilities under the control and new treatment.

The formula for PPCS could be used to evaluate the power of an RPW trial after completion of the experiment by substitution of the null and minimum alternative values of success probabilities. It could also be used to calculate an empirical significance level by substitution of the null and maximum indifference values of success probabilities. The approaches to analysis described by Ware would also be appropriate.

Although the accelerated convergence feature of this alternative design is attractive, detailed procedures for specifying u, the acceleration parameter, and the rule for discontinuance of randomization have not been developed. Neither has it been compared with the design presented by Ware. His design for the Harvard ECMO study meets the need for an adaptive design which responds to ethical considerations, yet provides adequate protection against an erroneous conclusion.

In closing I commend Professor Ware for the sensitivity to ethical issues and attention to scientific rigor which he has displayed in his work on the evaluation of the ECMO procedure. His discussion of statistical issues raised by the study will be especially helpful to anyone considering an adaptive design in a similar critical situation in the future.

ADDITIONAL REFERENCES

CORNELL, R. G. (1987). Play-the-winner in pediatrics: The ECMO trial. *ASA Proc. Biopharmaceutical Section* 243–245.

TOOMASIAN, J. M., SNEDECOR, S. M., CORNELL, R. G., CILLEY, R. E. and BARTLETT, R. H. (1988). National experience with extracorporeal membrane oxygenation for newborn respiratory failure: Data from 715 cases. *ASAIO Trans.* **34** 140–147.

Comment: Recent Progress in Clinical Trial Designs That Adapt for Ethical Purposes

Janis Hardwick

1. INTRODUCTION

Controlled medical trials are conducted for a variety of reasons, but in general the desire to validate new treatments that will, overall, decrease the suffering of the afflicted motivates their use. One classical and accepted approach to controlled research is the randomized clinical trial (RCT) in which patients are allocated randomly to competing therapies in such a way that an approximately equal number of patients is assigned to each regimen. But, as the current controversy illustrates, disparity often exists between the environment assumed necessary for a formal scientific inquiry and that of many real-life research situations.

The conflict between the need to conduct research and the desire to attend to the needs of individual

Janis Hardwick is Assistant Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1027.

patients has motivated statisticians to develop adaptive experimental techniques that address ethical dilemmas inherent in RCT's. However, despite a large base of statistical literature concerning designs featuring ethical allocation, there are only a few well-documented examples in which adaptive sampling was adopted due to ethical considerations. Among these are the Michigan and Harvard studies of ECMO. I believe that the issues raised by these studies would be less volatile and complex if the properties of adaptive designs were better understood.

Current concepts of what constitutes high quality medical research are presently in a state of flux. In 1977, Simon wrote that "Any design that does not provide for a convincing test of the null hypothesis has little chance of being adopted for general use." In the current article, Ware notes that "Many investigators believe that trials with unusual designs will not be persuasive." Challenging these perspectives, Berry (1989) argues that "The fact that classical inference is impossible to do in a legitimate scientific enterprise means to me that we should abandon classical inference rather than abandoning the enterprise."

During the last decade, the objective of much statistical research has been to provide the medical community with options that reduce the divergences of these views. I'd like to take this opportunity to point to some recent progress, and I hope, at the same time, to clarify some of the misconceptions that impede acceptance of this remarkably flexible class of designs. To support the latter goal, I've made ample use of direct quotations, the purpose being to exhibit the nature of the objections to adaptive techniques with acceptable precision. Because of space restrictions, only recent work is emphasized, although certain fundamental articles are also noted. There have been, of course, numerous contributions in this area, and among these are a number of well written books and survey articles which provide a wealth of information for the interested reader. See Hoel, Sobel and Weiss (1975), Weinstein (1974), Pocock (1983), Shapiro and Louis (1983), Siegmund (1985), Wetherill and Glazebrook (1986), and Berry and Fristedt (1986).

2. QUANTIFYING ETHICS: THE HARDEST PART

Some researchers insist that ethical considerations fall outside the purview of statistical analysis. To me this is arguable. It seems natural that statisticians should take part in the process of *quantifying* philosophical factors. Tymchuck (1981, 1982) offers insight into such structuring of decision processes. He bases models for ethical decision making on a variety of guidelines for social scientists, classifying the issues into seven *process factors* and four *decision criteria.* The process factors include 1) balancing the rights of individuals against the public interest, 2) avoidance of illegal or unjustified acts while also avoiding "bad laws," 3) using humanitarian and scientific knowledge in novel cases, 4) justice and equality, 5) multilateral decision processes, 6) use of guardians and advocates, and 7) levels of division and supervision. The decision criteria include A) cost, B) time and effort, C) benefits and risks, and D) "other aspects." While Tymchuck's work provides no rigorous techniques for quantifying these factors and criteria, it does provide useful guidelines with which to regulate the formation of decision theoretic models.

A number of the above items (2, 5, 6, 7) are likely to be dealt with by investigators working in areas of protocol development quite removed from that of the statistician (although, in the article under discussion, it appears that use of the randomized consent design brought process factors 5 and 6 into play as well.) Others, like factors 1, 3, 4 and any of the decision criteria could certainly arise as issues to be addressed through statistical modeling. Among all of the items listed, however, process factor 1 seems to represent the greatest challenge to statisticians trying to model ethical criteria. The crucial ingredient to quantifying this balancing factor is the patient horizon. If one could successfully evaluate the patient horizon and the impact that a trial will have on it, then the problem of measuring ethical costs would reduce to defining a mechanism for weighting the relative worths of current and future patients. But, assessment of the size and role of the patient horizon has been a topic of debate for decades, and no resolution seems close at hand. All clinical trials directly affect at least some segment of the patient horizon. In addition, since the actual effect of some trials often turns out to be surprisingly large or small, the role of the patient horizon needs to be given explicit attention. Issues that are relevant to the estimation process include the prospective rate of adoption of new therapies, the prevalence of the affliction being researched, and the projected audience of physicians. Discussions may be found in Armitage (1960), Anscombe (1963), Colton (1963), Upton and Lee (1976), Hoel, Sobel and Weiss (1975), Simon (1977), Petkau (1978), Lai (1984), Armitage (1985), Chernoff and Petkau (1985), Bather (1985), Hardwick (1986a, b), Woodroofe and Hardwick (1988) and Hardwick (1989a).

3. ADAPTIVE DESIGNS: WHAT ARE THE ISSUES?

Adaptive techniques are not new to clinical trials; however, the term covers such varied practices that it rarely suffices as a useful descriptor. Van Ryzin (1986)

provides an informal definition when he refers to adaptive procedures as having "the common feature that they adapt to approximate some ideal or optimal statistical procedure as observations accumulate in number and/or time." (See also Kulkarni, 1988.) The information observed for each patient in a trial consists of a vector of random variables with three components: the measured prognostic factors, an indicator of treatment, and the measured responses to treatment. An adaptive technique, then, is one that uses some subset of all information known at any given time to determine how to allocate the next patient or whether to stop the trial. Of particular interest here are designs that adapt for the specific purpose of minimizing an ethical cost.

The adaptive nature of the design must be designated in a formally structured set of protocols, otherwise, possible ad hoc decisions might invalidate the trial. For example, trials in which patient responses are merely monitored for extreme behavior are not adaptive in the current sense because resulting capricious termination would destroy the intended statistical environment. (See, for example, Pocock, 1982, Goldman, 1987 and Berry, 1988).

Although there are often compelling reasons to prefer adaptive experiments to those with predetermined allocation schemes, as a rule, the design and analysis of such experiments is more complicated, largely because sampling distributions are influenced by sequential design or optional stopping. Further complications arise when the physical limitations of the data collection process must be incorporated into an already intricate design. In order for sequential procedures to be considered as acceptable foundations for clinical trial protocols, they must be shown to be general enough to encompass the complex phenomena that arise in medical research.

The skepticism evoked by adaptive techniques is both practical and philosophical in nature; the former relating to what we are able to do, and the latter to what we should do. Criticisms of adaptive designs address their ability to handle randomization, balance, group sampling, delayed responses, and covariates, while at the same time allowing desired tests of hypotheses and inference procedures. Philosophical concerns are the usual sort involving: type of analysis (frequentist, Bayesian, other); the use of prior information (historical controls, informative priors, initial parameter estimates); ethical criteria (loss functions, horizon size, myopic allocation); and, more generally, the difficulties faced in any attempt to get a new type of design accepted in practice (what constitutes convincing evidence, possible exposure to liability, and so on). With the above factors in mind, I'll focus on several approaches, each with a different mechanism for combining statistical, practical and ethical concerns.

3.1. General Approaches

A useful method of formalizing the design goals of a trial is to use decision theory. With this approach, one can attempt to construct a loss function that represents costs associated with treating patients suboptimally during the trial, invasiveness and side effects of each treatment, reaching incorrect conclusions at the end of a study, and, more generally, making "poor" inferences about parameters.

In allocation problems, where the "sampling" costs are a function of the therapy prescribed, good solutions are apt to depend on two factors: knowing when to stop and achieving a fixed balance in treatment assignments. Usually, the "optimal" balance depends on unknown parameters and, thus, the ability to update parameter estimates as the trial proceeds is likely to be rewarded with a reduction in overall risk. In realistically formulated allocation problems, identifying a class of optimal procedures is extremely difficult. This is not a critical problem, however, since optimal rules are frequently too intricate to be practicable.

A first step, then, may be to seek a class of rules that is asymptotically optimal (in an appropriate sense). Suppose, for example, that θ is the parameter of interest and that, at time k, $\hat{\theta}_{m_k,n_k}$ is the estimate based on m_k and n_k observations from treatments I and II, respectively, with $m_k + n_k = k$. Then if $p(\theta)$ is the (unknown) proportion that we wish to allocate to treatment I, an allocation rule suggested by asymptotic analyses might proceed as follows:

Rule 1: Sample from T_I at stage $k + 1$ if and only if $m_k/k \leq p(\hat{\theta}_{m_k,n_k})$.

Rule 1 has the "practical" disadvantages of requiring that new computations be made after every observation and that each patient's response must be available before the next patient may be assigned. The next move, therefore, may be to determine conditions under which the class of asymptotically optimal rules will include designs that are more realistic and easy to use. Ideally, such conditions, while constraining the theoretical solutions, will not seriously hinder the functionality of the problem formulation. In the present context, it is primarily designs such as these, which encompass utilitarian rules, yet retain some degree of approximate optimality, that merit further study.

In many cases, locating the minimum achievable risk, if not the rules that achieve it, can be a relatively easy task. If so, when evaluating approximate rules, it is useful to determine both the extent of their deficiencies relative to optimal designs and the magnitude of their improvement over designs in current use.

3.2. Group Allocation

A simplifying constraint that is desirable (particularly to investigators involved in multi-center trials) is the allocation of subjects as part of a group rather than as individuals. Data from previous patients is not always available when a new patient needs to be assigned, and grouping the patients allows more time for data to arrive. Moreover, procurement and analysis of data from a group sequential trial is often more straightforward than analysis of a fully sequential trial. During each stage of a multi-stage design, patients are ordinarily allocated to the treatments in fixed proportions, although the allocation process may be either deterministic or random. After each stage, adjustments to the sampling scheme may be made or the trial may be stopped.

Multi-stage designs comprise a huge family. The simplest of these, at least intuitively, are two-stage designs which allocate to both treatments during the first phase, but to a single treatment thereafter. The design used at Harvard falls into this class, and, despite the apparent lack of use of such designs, they have been proposed numerous times as a simple variation of the RCT (Colton, 1963; Begg and Mehta, 1979; Witmer, 1986; Berry and Pearson, 1985).

Designs referred to as "group-sequential" form another set in this class. In group-sequential designs, patients are usually sampled according to the same fixed scheme during every stage. The ultimate goal of such designs is usually a test of hypothesis; so, after each group is sampled, the data are checked for "significant" results. A problem of theoretical interest raised by group sequential designs is how the repeated examinations of the data affect significance levels. DeMets and Lan (1984) review several approaches to the concept of "spending" the type I error level, and more recent attention has focused on obtaining confidence intervals following group sequential tests. (See DeMets and Ware, 1980; Jennison and Turnbull, 1983, 1984; DeMets and Kim, 1987a, b; Geller and Pocock, 1988; Rosner and Tsiatis, 1988).

A further class of multi-stage designs can be derived from the type of strategies exemplified by Rule 1. Here, however, the updated parameter estimates are computed only after groups of patients have been observed. For simplicity, the number of groups is usually small and, as before, the periodic examinations serve the two-fold purpose of allowing for adjustment of the allocation scheme or for determining whether to stop the trial. The decision theoretic formulation allows either testing or estimation, of course, and, recently, there has been some work on models in which ethical costs are explicitly expressed in the loss function (Woodroofe and Hardwick, 1988; Hardwick, 1989a). There are a variety of potential applications for adaptive multi-stage rules, and they appear to offer promising compromises between fully sequential and fixed proportion allocation rules. For further discussion see Hall (1981), Siegmund (1985), Clayton and Witmer (1988), Woodroofe (1988), Lorden (1988), and the references for Section 5.1.

3.3. Randomization

The quotations below typify often-occurring excessive expectations for RCT's.

"The data of Bartlett et al. do not, therefore, perform the function of a randomized trial, which is to convince practitioners that extracorporeal membrane oxygenation is definitely superior to conventional treatment ... more importantly, the inference from this statistical calculation is only valid if the two treatment groups were identical in all respects except treatment" (Panath and Wallenstein, 1985).

"Their experience shows, however, the need for caution in replacing conventional randomization with adaptive schemes.... The imbalance in treatment assignments illustrates a danger with adaptive designs, namely, that the design may have good performance on the average but fail to provide a good comparison between the regimens in a particular instance" (Ware and Epstein, 1985).

Randomization never guarantees that the "treatment groups will be identical in all respects except treatment," and it frequently fails "to provide a good comparison ... in a particular instance." While it does provide an appropriate environment for tests of statistical significance, randomization should not be viewed as a panacea for the disbalancing effects of covariates, known or unknown. Regardless of whether a trial has been randomized, good statistical practice always includes analysis of treatment groups for biases in pertinent prognostic factors. Ware specifically mentions post hoc analyses for biases done on the phase variable (random versus nonrandom), and it is likely that similar exploratory analyses were carried out on other potentially confounding variables. Of course, such analysis does not necessarily protect the data from biases caused by lurking variables, but then, one is likewise never assured that a standard randomization scheme will do so.

Although the option of randomizing is not always available, it is preferable whenever possible to utilize designs with randomization sufficient to protect as well as can be against selection bias, time trends, and any of a wide variety of unknown factors. The significant lag time that exists between the time new ideas are published and when they are assimilated into general practice is demonstrated by Ware's comment that "Sequential methods are not especially effective in this setting ... They do not use adaptive randomization."

There is no dearth of techniques available for incorporating randomization into deterministic adaptive schemes. One fairly straightforward method, involves superimposing the allocation strategy with a "biased coin" scheme (Efron, 1971; Wei 1977, 1978). As an illustration of this procedure, consider again Rule 1, and note that the following rule has similar properties and also is randomized:

Rule 2: Sample from treatment 1 at stage $k + 1$ if and only if $U \leq p(\hat{\theta}_{m_k, n_k})$, where U is a uniform random variable on (0, 1).

The main problem accompanying this technique is the sparseness of published theoretical examinations of the impact brought about by adding randomization. Research in this area is being conducted, however. For example, there are several cases for which it has been established that the deterministic version of a rule such as Rule 1 is asymptotically optimal, and it is believed that the randomization will affect the operating characteristics of the original rule only marginally. This view has been confirmed for the sequential version of the Behrens-Fisher problem in the sense that first order asymptotic efficiency is retained when an adaptive biased coin scheme is overlaid on the adaptive allocation scheme suggested by Robbins, Simons and Starr (1967) (see Eisele, 1989). Other cases under study include estimation problems with ethical cost (Hardwick, 1989b) and modified repeated significance tests (Siegmund, 1983, 1985; Hardwick, 1989a).

Also relevant to this topic are articles presenting comparisons of nonrandomized adaptive, randomized adaptive and/or randomized nonadaptive rules. (See, for example, Berry and Eick, 1989; Bather, 1981; and Glazebrook, 1980.)

4. TESTING AND INFERENCE

4.1 The Null Hypothesis

Because of its inclination toward the use of classical 5% tests for rejecting a true null hypothesis, the medical community may be placing excessive confidence in this one interpretation of what the "correct" statistical analysis should be. The traditional emphasis usually placed on whether and how to test hypotheses often causes other retrievable information in the data to be overlooked. Clinical trials can be initiated with a variety of different objectives, and some of these goals may depend on aspects of the therapies that are not considered in the simple success-fail loss function. Such neglected factors should be an argument in favor of the utilization of less standard and/or more exploratory techniques.

One problem that plagued the original ECMO trial seems to have resurfaced to haunt the present one. What I find perplexing is that, in each study, the design parameters were specified in a way that seems to contradict the assumption of equipoise described by Ware. In the Michigan study, the sample size calculation was based on a discrimination parameter (call this Δ_M) defined to be 0.4, and on the initial estimate of the survival rate for ECMO, $P_2 = 0.8$. In the Harvard study, the research question may be inferred from the statement of the hypotheses: $H_0: P_1 = P_2 = 0.2$ and $H_1: P_1 = 0.2$, $P_2 = 0.8$. In this design, the value $P_2 - P_1$ assumed under the alternative hypothesis (call this Δ_H) serves essentially the same purpose as does Δ_M in the Michigan design.

Specification of a discrimination parameter such as Δ_H or Δ_M should be based, not on a prior estimate of $P_2 - P_1$, but rather on the smallest difference in survival rates that the investigators either would have found meaningful or had the resources to detect, i.e., the *clinical significance*. If $\Delta_H = 0.6$ and $\Delta_M = 0.4$ were estimates of $P_2 - P_1$ *before* the trial began, then the assumption of equipoise was clearly not met. On the other hand, it is difficult to imagine that 0.4 or 0.6 could be the smallest clinically significant difference for a therapy where failure means death. Suppose the true difference in efficacy rates at Harvard had been 0.3. Surely this is a difference worth notice. But, had the Harvard null hypothesis been rejected, the study conclusion would have been that there was insufficient information to conclude ... *what*? To conclude that there was a greater than 60% difference between therapies or to conclude that there was a *meaningful* difference between therapies? I suspect that the answer is neither. The alternative hypothesis was probably specified as it was to reduce the sample size needed to find a statistically significant difference, and there are clearly reasons why a small sample size would have been a desirable goal. These comments are not intended to raise doubts about the validity of the assumption of equipoise made in the Harvard study, but to investigate whether the research question, currently stated in terms of a hypothesis test, would not have been better answered through the use of some other analytic technique.

Adaptive trials with ethical constraints have not been used much, leading to a natural lack of familiarity with related techniques of design and analysis. Following the Michigan study, the investigators and discussants apparently felt obligated to try to analyze their results as if the data had come from a standard RCT. Panath and Wallenstein (1985) state: "Had the results of Bartlett et al.'s study emerged from a randomized clinical trial, a conventional analysis would be to use the Fisher-Irwin test. This yields a one-tailed probability of 0.083, which is not statistically significant." Ware and Epstein (1985) add: "In particular, the type 1, or false positive, error rate for this design is 0.5!"

Even Bartlett and his colleagues contribute: "... the lower 99% one-sided confidence interval on the probability of survival with ECMO is 78.5.... This conclusion is based on all the data on ECMO, both before and after the discontinuance of randomization, and on standard confidence interval calculations."

As a graduate student, I took great interest in the published discussions of the ECMO study at Michigan. This is largely because the design used, based on Durham and Wei's randomized play-the-winner rule, has certain properties in common with the modified bandit, a highly adaptive rule that I was proposing in my dissertation. Out of this interest grew a discussion that was included in the dissertation, and in it, I responded (a bit overzealously, perhaps) to the above interpretations of the study data: "Because the ECMO data were not collected with the intention of performing tests such as those described, the above are rather ill-adapted analyses. If there were test statistics of the sort described above, they would not inherit the distributional laws ascribed to them in these references because sampling distributions are affected by study designs. Unless one sticks entirely to a likelihood interpretation of the data, one must interpret the study results in terms of the design that produced them, i.e., the probability was 0.05 that the inferior treatment would have been designated as the better one when the difference between the success rates was at least 0.4. If a test of the significance of the HEQ (the null hypothesis of equality) is an important objective of the trial, such a test should be incorporated in the design at the outset, not concocted after the results are in. It is worth noting, however, that another option was available, although apparently not considered. The statisticians could have performed randomization tests of the desired hypotheses. Given the small expected sample size, the computational problems would have been quite manageable, and the resultant tests would have provided the requisite p-values and error probabilities.... The allocation rule employed has little bearing on whether the data are eventually analyzed using the classical HEQ. Rather, it is the overall formulation of the design that affects the availability of the desired test. Nevertheless, because the medical community evinces a strong preference for the use of HEQ's in the designs of experiments, innovative designs for comparing standard therapies with new ones are more likely to be adopted if they also include HEQ tests."

Of course, permutation tests for these data were eventually carried out, and the algorithms used were concise and elegant (Wei, 1988). Wei, responding to his own question of whether "... the design can be ignored in the analysis," adds that "... the degree of significance of the treatment effect is exaggerated if the design is ignored in the analysis." (See also Wei and Lachin, 1988.) Finally, it may be worth noting that some selection designs (the class in question during the Michigan discussions) do incorporate tests of the null hypothesis (see Kiefer and Weiss, 1971, 1974; Hoel, Weiss and Simon, 1976; Hardwick, 1986b).

4.2. Repeated Significance Tests

The assumption that adaptive designs preclude making frequentist inferences is not entirely correct. Among adaptive rules, repeated significance tests (RST's) have, perhaps, the greatest potential for use in clinical trials. These tests have received considerable attention, and their properties are quite well understood.

Traditionally, subjects are allocated evenly between the two treatments, either in pairs or in larger groups. After each group, the test statistic of interest is updated and checked against a stopping boundary which is a function of the total number of patients observed to date. If the test statistic exceeds the boundary then the result is significant, but if a truncation point is reached first then the null hypothesis is not rejected. Numerous modifications of RST's have been studied, many of which are discussed in Siegmund, 1985, Chapters 4–6. (Of particular interest is Section 3, Chapter 6, in which the author discusses adaptive allocation.) Siegmund carefully addresses the practical issues related to use of these tests and, by means of Monte Carlo studies, demonstrates approximate gains and losses associated with the various modifications. Among modifications currently under study are RST's that incorporate ethical costs (Hardwick, 1989a). These include both group and fully sequential adaptation, in the sense that adjustments to the allocation strategy take place only after groups of patients have been observed, but stopping criteria are checked after each observation. It is anticipated that the results will be relatively insensitive to the design, and that approximations of the operating characteristics of these procedures will be similar to those derived by Siegmund. In a recent paper, Hu (1988) reviews analytic techniques for working with modified RST's and extends some of the results that had previously been shown only for normal random variables to random variables from exponential families. (See also Takahashi, 1987.)

4.3. Confidence Intervals

"One difficulty with adaptive designs is that methods ... for calculating confidence intervals at the completion of the study, are not presently available" (Ware and Epstein, 1985).

This comment deserves at least a brief response. While moving away from RCT's may imply giving up familiar analytic techniques, it does not necessarily mean that less common procedures do not provide

legitimate alternatives which may be substituted when the former are either not available or not accurate. As mentioned in Section 3.2, substantial research has been carried out on the formation of confidence intervals following group sequential tests. Progress in this area has also been made with regard to other types of sequential designs:

- Many authors have discussed fixed width confidence intervals, in which sampling continues until a predetermined level of precision is estimated to have been obtained (Robbins, Simons and Starr, 1967; Siegmund, 1985; Woodroofe, 1988; Meslim, 1987; Eisele, 1989).
- Using the Michigan data, Wei, Smythe, Lin and Park (1990) compare a number of methods for generating confidence intervals for adaptive designs. In fact, the profile likelihood methods described by Wei and colleagues were used to derive confidence intervals for the Harvard study.
- Using a relatively new approach, Woodroofe exhibits methods for adjusted confidence intervals using estimates whose sampling distributions have been affected by the use of a stopping rule or an adaptive sampling procedure. These bias corrections are based on very weak expansions for the distribution of an appropriately transformed parameter estimate, and they have been examined in several situations where first order asymptotic approximations do not account for biases introduced by an unusual sampling scheme (Woodroofe, 1986, 1989; Hardwick and Woodroofe, 1989).

5. OTHER APPROACHES

5.1. Bandit Problems

"... the dividing line between experiment and routine should really not exist at all. In terms of the accumulation of clinical knowledge, an experiment should not be thought of as a discrete entity but rather as one phase of the ongoing experience with a clinical procedure ... Thus, the experiment never really ends, and concern for the subject as a patient begins the first time a procedure is used" (Weinstein, 1974).

Clearly, in most research settings, the medical community is reluctant to respond to such remarks by changing dramatically their concepts of scientific investigations. However, in settings where ethical considerations are dominant but where it has been decided that some sort of comparative trial must take place, it may be reasonable to consider design formulations based on bandit problems. In bandit problems, optimal policies are those that indicate which treatment should be assigned at each stage in a, possibly infinite, series of trials, with the goal being to maximize the sum of discounted patient responses. The vast assortment of discount sequences available suggests a variety of ways to quantify ethical criteria. The attitude expressed by Weinstein is well represented by geometric discount sequences, $\{1, \beta, \beta^2, \ldots\}$, $0 < \beta < 1$, for these reflect a version of equity with regard to patient treatment: regardless of when the patient is treated, the weight given their response, relative to the sum of the weights of all future patients, is a constant function of β, $((1 - \beta)/\beta)$. Another advantage of the geometric sequences is that the problem of obtaining optimal strategies is far more tractable mathematically (Gittins and Jones, 1974); however, the infinite horizon of the geometric bandit is a drawback for, without a formal end to the data collection phase, it is difficult to perform classical statistical analyses. A couple of options exist. One can either truncate at a predetermined time or simply overlay an ad hoc stopping rule. These options are discussed in Hardwick (1986a, b). The entire area of bandit problems offers approaches to sequential design problems that, as yet, have been impossible to examine in depth. But, given recent advances, numerical solutions of such problems are now feasible in more practical situations (Chernoff and Petkau, 1983; Lai, 1987; Gittins, 1989). Many extensions to conventional bandit problems have been examined, and they include rules that incorporate covariates (Woodroofe, 1979; Clayton, 1988; Sarkar, 1989); delayed responses (Eick, 1988; Flournoy, 1989); and multi-stage sampling (Lai, Levin, Robbins and Siegmund, 1980; Witmer, 1986). Also, in their monograph, Berry and Fristedt (1986) provide an excellent annotated bibliography of contributions to research on bandit problems.

5.2 Desperate Measures

There are unquestionably cases that arise in which controlled studies are judged completely unethical. This doesn't mean that everyone should give up and go home. There are options available. Attempting to discourage randomized trials in such situations, in Hardwick (1986a), I suggested that if, in the design stages of a trial, "it became apparent that certain treatment assignments would present insurmountable ethical problems to participating investigators, they should have decided against the trial. The option of continuing to provide their own patients with ECMO, and, in time, of sharing their carefully documented records with other physicians was always available and could have been substituted for a trial." One way of supporting the validity of information collected in such a fashion is to work with predictive models or Bayesian analyses.

Logistic regression has been used to model survival of ECMO patients (Toomasian, Snedecor, Cornell,

Cilley and Bartlett, 1988). Similar models based on historical data collected by Bartlett prior to his study are being developed for non-ECMO patients (Berry and Hardwick, 1989). Preliminary results indicate that, even while controlling for the effect of technological advances in treatment, these models predict survival extremely well (90%). In addition, it is anticipated that models utilizing more recent data will support the contention that there were respectable alternatives to the randomized study performed at Michigan. See Doll and Peto (1980) and Gehan (1984) for discussions on the use of historical controls.

Tackling the Bayesian viewpoint head-on, some statisticians are trying to formalize methods for eliciting priors from "experts," anticipating that such information can be used to generate "prior" distributions. This approach is quite removed from the ideas discussed here, but it represents a departure from classical methods and often does so with the intent to increasing the well being of the patient horizon. Related references include Kadane (1980), Chaloner (1983) and Freedman and Spiegelhalter (1983).

Obviously the problems mentioned above must be approached with great caution, for there are tangible difficulties that go hand in hand with inferences based on Bayesian analyses and historical controls. In fact, many obstacles must be overcome before inferences based on nonfrequentist arguments will find acceptance. Even so, I hope such work will provide examples of how such inferences can be made when classical randomized methods are not viable.

6. THE ULTIMATE GOAL: ARE ADAPTIVE DESIGNS WORTH IT?

In his discussion, Ware refers to an attitude that can easily catch one off guard: "Some statisticians believe that randomization with constant randomization probabilities should be continued so long as randomization is ethically justified, and that adaptive rules are an insufficient response to evidence that the therapies are not equally effective." My immediate reaction to this statement is to agree, but on closer inspection, I think that such statements can imply contradictory postures depending on how one interprets the word "evidence." As statisticians, we are often faced with the task of delineating the difference between data that "seem to show something" or "look worrisome" and data that are statistically meaningful. If, to an ordinary person, a physician said that current evidence points to a 10% survival rate for CMT and a 90% survival rate for ECMO, then that person would say "I want my baby to have ECMO." A statistician, on the other hand, would begin to ask question upon question, trying valiantly to assess the value of the comparison given by the physician. Prior to the randomized study at Michigan, it was decided that whatever was meant by those words was not evidence of a difference between ECMO and CMT; yet, clearly, whatever evidence the investigators had at that time, while not being statistically significant in an acceptable sense was strong enough to keep the human subjects review board from allowing an RCT to be conducted. Where does that leave us? Certainly, if we believe that the treatments are not equally efficacious, the assumption of equipoise is violated and we should not be conducting a randomized comparative trial. Suppose however, that prior data show a trend in favor of ECMO and that, once the trial is started, the data seem to point strongly in the same direction, but are not extreme enough to invoke the stopping rule? Without adaptation and without nonfrequentist alternatives, what are the options? The trail can be stopped before statistically convincing results are obtained or the trial can be continued with 50%–50% randomization, and, if the data trend is confirmed by eventual significant results, an excessive number of patients will have been exposed to the inferior treatment. As long as we stick to the philosophy that data are informative only when they can be shown to represent statistically significant results, our options are few, and adaptation is one of the only precautionary measures available.

Recall that the main attribute of an adaptive rule is that it allows the trial to continue to its conclusion while limiting exposure of the trial patients, and, in this sense, adaptive rules truly offer a compromise between an RCT (which may harm the trial patients unnecessarily) and no trial at all (which may result in harm to future patients who won't have the benefit of scientific results). In medical trials we deal with the unknown, so everything is a gamble. But if, by using a certain trial design, we expect even an extra 1 or 2% of the patients to live than would otherwise (and if we don't pay for this gain by obvious losses in another area), then I believe that the answer is yes: *yes, it is worth all the controversy and all the hours that we spend studying such possibilities.*

7. CONCLUSIONS

Acknowledging the lack of faith that most investigators place in nonclassical RCT's, Ware defends the Harvard researchers by noting that they recognized the limitations of their study "... relative to the classical RCT, and made special efforts to maintain other strengths of randomized trials, especially standardized accrual, treatment, and data collection methods, throughout the study." It's unfortunate that such a statement was necessary.

A substantial obstacle preventing adoption of new design techniques is the prevalent view that because patients are not randomly assigned to treatment groups, all other aspects of the study must then be

suspect. Correct practices such as controlling for external biases, maintaining strict data collection standards, and generally enforcing consistent behavior throughout an experiment are independent of randomization; they are simply good scientific practice. It is possible, in fact, that the excessive confidence placed in randomization even generates a kind of lazy negligence regarding other aspects of rigorous procedure.

Quite often and unexpectedly, challenges associated with a new clinical trial are found to be unprecedented or unique, demonstrating that no single class of clinical trial designs should have a universal role in the ongoing process of medical research. In striving to retain the high level of rigor that is popularly considered to go hand in hand with the classical RCT, the profession sometimes underrates the applicability of alternative techniques. Rejection of new methodology often results from misinformation or lack of easily available information regarding its applications and properties.

Choosing a trial design that will serve the needs of patients affected by that trial requires a significant degree of open mindedness and courage. Regardless of whether "the right thing was done," Ware and his colleagues are to be commended for their willingness to try something new; there were no loss-free decisions in the project that they undertook. Lebacqz (1983) neatly covered the essentials of the matter when, after discussing a debate not unlike the current one, she remarked that "The preceding study illustrates well two conclusions to be drawn from this discussion of ethical issues in clinical trials; first, that the issues must be decided on a case by case basis, and second, that public discussion and debate is important for clarifying the issues."

ADDITIONAL REFERENCES

ANSCOMBE, F. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58** 365–383.

ARMITAGE, P. (1960). *Sequential Medical Trials.* Thomas, Springfield, Ill.

ARMITAGE, P. (1985). The search for optimality in clinical trials. *Internat. Statist. Rev.* **53** 15–24.

BATHER, J. (1981). Randomized allocation of treatments in sequential experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **43** 265–292.

BATHER, J. (1985). On the allocation of treatments in sequential medical trials. *Internat. Statist. Rev.* **53** 1–13.

BEGG, C. and MEHTA, C. (1979). Sequential analysis of comparative clinical trials. *Biometrika* **66** 97–103.

BERRY, D. (1988). Monitoring safety data in a clinical trial with the possibility of early stopping. Unpublished.

BERRY, D. (1989). Inferential aspects of adaptive allocation rules. Unpublished.

BERRY, D. and EICK, S. (1988). Decision analysis of randomized clinical trials: Comparison with adaptive procedures. Unpublished.

BERRY, D. and FRISTEDT, B. (1986). *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall, New York.

BERRY, D. and HARDWICK, J. (1989). Using historical data in a Bayesian analysis of clinical trials. Unpublished.

BERRY, D. and PEARSON, I. (1985). Optimal designs for clinical trials with dichotomous responses. *Statist. in Medicine* **4** 497–508.

CHALONER, K. (1983). Assessment of a beta prior distribution PM elicitation. *The Statistician* **32** 174–180.

CHERNOFF, H. and PETKAU, A. J. (1983). Numerical methods for Bayes sequential decision problems. Technical Report No. 83-26, Dept. Applied Mathematics and Statistics, Univ. British Columbia.

CHERNOFF, H. and PETKAU, A. J. (1985). Sequential medical trials with ethical cost. *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. LeCam and R. A. Olshen, eds.) **2** 521–537. Wadsworth, Monterey, Calif.

CLAYTON, M. (1988). Bernoulli bandits with covariates. Technical Report, Dept. Statistics, Univ. of Wisconsin-Madison.

CLAYTON, M. and WITMER, J. (1988). Two-stage bandits. *Ann. Statist.* **16** 887–894.

COLTON, T. (1963). A model for selecting one of two medical treatments. *J. Amer. Statist. Assoc.* **58** 388–400.

DEMETS, D. and KIM, K. (1987a). Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43** 857–864.

DEMETS, D. and KIM, K. (1987b). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74** 149–154.

DEMETS, D. and LAN, K. (1984). An overview of sequential methods and their applications in clinical trials. *Comm. Statist. A—Theory Methods* **13** 2315–2338.

DEMETS, D. and WARE, J. (1980). Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika* **67** 651–660.

DOLL, R. and PETO, R. (1980). Randomised controlled trials and retrospective controls. *Brit. Med. J.* **280** 44.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417.

EICK, S. (1988). The two-armed bandit with delayed responses. *Ann. Statist.* **16** 254–264.

EISELE, J. (1989). An adaptive biased coin design for the Behrens-Fisher problem. Technical Report, Dept. Statistics, Univ. Michigan.

FLOURNOY, N. (1989). Adaptive designs in clinical trials. *ASA Proc. Biopharmaceutical Section.* To appear.

FREEDMAN, D. and SPIEGELHALTER, D. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **32** 153–160.

GEHAN, E. (1984). The evaluation of therapies: Historical control studies. *Statist. in Medicine* **3** 315–324.

GELLER, N. and POCOCK, S. (1988). Design and analysis of clinical trials with group sequential stopping rules. In *Biopharmaceutical Statistics for Drug Development* (K.E. Peace, ed.) 489–508. Dekker, New York.

GITTINS, J. C. (1989). *Multi-armed Bandit Allocation Indices.* Wiley, Chichester.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (J. Gani, ed.) 244–266. North-Holland, Amsterdam.

GLAZEBROOK, K. (1980). On randomized allocation indices for the sequential design of experiments. *J. Roy. Statist. Soc. Ser. B* **42** 342–346.

GOLDMAN, A. (1987). Issues in designing sequential stopping rules for monitoring side effects in clinical trials. *Controlled Clin. Trials* **8** 327–337.

HALL, P. (1981). Asymptotic theory of triple sampling for sequential estimation of the mean. *Ann. Statist.* **9** 1229–1238.

HARDWICK, J. (1986a). Adaptive allocation in clinical trials, the

ECMO controversy. Technical Report, Dept. Statistics, Univ. Michigan.

HARDWICK, J. (1986b). The modified bandit: An approach to ethical allocation in clinical trials. Technical Report, Dept. Statistics, Univ. Michigan.

HARDWICK, J. (1989a). Randomized repeated significance tests with unbalanced sampling. *ASA Proc. Biopharmaceutical Section.* To appear.

HARDWICK, J. (1989b). Practical adaptive allocation rules with loss functions that incorporate ethical costs. *Contemp. Math. Proc. of the Conference on Statistical Multiple Integration* (N. Flournoy and R. Tsutakawa, eds). To appear.

HARDWICK, J. and WOODROOFE, M. (1989). Bias corrections for randomly censored survival data. Technical Report, Dept. Statistics, Univ. Michigan.

HOEL, D., SOBEL, M. and WEISS, G. (1975). A survey of adaptive sampling for clinical trials. In *Perspectives in Biometrics* **1** 29–61. Academic, New York.

HOEL, D., WEISS, G. and SIMON, R. (1976). Sequential tests for composite hypotheses with two binomial populations. *J. Roy. Statist. Soc. Ser. B* **38** 302–308.

HU, I. (1988). Repeated significance tests for exponential families. *Ann. Statist.* **16** 1643–1666.

JENNISON, C. and TURNBULL, B. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **25** 49–58.

JENNISON, C. and TURNBULL, B. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clin. Trials* **5** 33–45.

KADANE, J. (1980). Predictive and structural methods for eliciting prior distributions. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Sir Harold Jeffreys* (A. Zellner, ed.). North-Holland, Amsterdam.

KIEFER, J. and WEISS, G. (1971). A truncated test for choosing the better of two binomial populations. *J. Amer. Statist. Assoc.* **66** 867–871.

KIEFER, J. and WEISS, G. (1974). Truncated version of a play-the-winner rule for choosing the better of two binomial populations. *J. Amer. Statist. Assoc.* **69** 807–809.

KULKARNI, R. (1988). Adaptive statistical procedures. *Enc. Statist. Sci.* **8** 399–404.

LAI, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry. A sequential approach. *Comm. Statist. A—Theory Methods* **13** 2315–2338.

LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114.

LAI, T. L., LEVIN, B., ROBBINS, H. and SIEGMUND, D. (1980). Sequential medical trials. *Proc. Nat. Acad. Sci. U.S.A.* **77** 3135–3138.

LEBACQZ, K. (1983). Ethical aspects of clinical trials. In *Clinical Trials: Issues and Approaches* (S. Shapiro and T. Louis, eds.). Dekker, New York.

LORDEN, G. (1988). Testing in stages. Abstract 206–66, *Inst. Math. Statist. Bull.* **17** 234.

MESLIM, A. (1987). Asymptotic expansions for confidence intervals with fixed width. Ph.D. dissertation, Dept. Statistics, Univ. Michigan.

PETKAU, A. J. (1978). Sequential medical trials for comparing an experimental treatment with a standard. *J. Amer. Statist. Assoc.* **73** 328–338.

POCOCK, S. (1982). Interim analysis for randomized clinical trials: The group sequential approach. *Biometrics* **38** 153–162.

POCOCK, S. (1983). *Clinical Trials.* Wiley, Chichester.

ROBBINS, H., SIMONS, G. and STARR, N. (1967). A sequential analogue of the Behrens-Fisher problem. *Ann. Math. Statist.* **38** 1384–1391.

ROSNER, G. and TSIATIS, A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75** 311–318.

SARKAR, J. (1989). A one-armed bandit problem with a concomitant variable. Technical Report, Dept. Statistics, Univ. Michigan.

SHAPIRO, S. and LOUIS, T., eds. (1983). *Clinical Trials: Issues and Approaches.* Dekker, New York.

SIEGMUND, D. (1983). Allocation rules for sequential clinical trials. In *Mathematical Learning Models Theory and Algorithms* (U. Herkenrath, D. Karlin and W. Vogel, eds.) 203–212. Springer, New York.

SIEGMUND, D. (1985). *Sequential Analysis. Tests and Confidence Intervals.* Springer, New York.

TAKAHASHI, H. (1987). Asymptotic expansions in Anscombe's theorem for repeated significance tests and estimation after sequential testing. *Ann. Statist.* **15** 278–295.

TOOMASIAN, J. M., SNEDECOR, S. M., CORNELL, R. G., CILLEY, R. E. and BARTLETT, R. H. (1988). National experience with extracorporeal membrane oxygenation for newborn respiratory failure: Data from 715 cases. *ASAIO Trans.* **34** 140–147.

TYMCHUCK, A. (1981). Ethical decision making and psychological treatment. *J. Psychiatric Treatment Evaluation* **3** 507–513.

TYMCHUCK, A. (1982). Strategies for resolving value dilemmas. *Amer. Behav. Scientist* **26** 159–175.

UPTON, G. and LEE, R. (1976). The importance of the patient horizon in the sequential analysis of binomial clinical trials. *Biometrika* **63** 335–342.

VAN RYZIN, J. (1986). Preface to *Adaptive Statistical Procedures and Related Topics* iii–vi. IMS, Hayward, Calif.

WEI, L. J. (1977). A class of designs for sequential clinical trials. *J. Amer. Statist. Assoc.* **72** 382–386.

WEI, L. J. (1978). The adaptive biased coin design for sequential experiments. *Ann Statist.* **6** 92–100.

WEI, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* **75** 603–606.

WEI, L. J. and LACHIN, J. (1988). Properties of the urn randomization in clinical trials. *Controlled Clin. Trials* **9** 345–364.

WEINSTEIN, M. (1974). Allocation of subjects in medical experiments. *New England J. Med.* **291** 1278–1285.

WETHERILL, B. and GLAZEBROOK, K. (1986). *Sequential Methods in Statistics*, 3rd ed. Chapman and Hall, London.

WITMER, J. (1986). Bayesian multistage decision problems. *Ann. Statist.* **14** 283–297.

WOODROOFE, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74** 799–806.

WOODROOFE, M. (1986). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14** 1049–1067.

WOODROOFE, M. (1988). Fixed proportional accuracy in three stages. In *Statistical Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.). **2** 209–221. Springer, New York.

WOODROOFE, M. (1989). Very weak expansions for sequentially designed experiments: Linear models. *Ann. Statist.* **17** 1087–1102.

WOODROOFE, M. and HARDWICK, J. (1988). Sequential allocation for an estimation problem with ethical costs. *Ann. Statist.* To appear.