Annu. Rev. Public Health. 1999. 20:109–24

ADVANCES IN CLINICAL TRIALS IN THE TWENTIETH CENTURY

Lloyd D. Fisher
Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232; e-mail: lfisher@biostat.washington.edu

KEY WORDS: randomization, sequential analysis, ethics of human experimentation, intention-to-treat, Food and Drug Administration

ABSTRACT

This article considers the rise of the randomized clinical trial during the twentieth century. Before such development could begin, probability and statistics needed to merge. Sir RA Fisher introduced randomization in the 1920s and, beginning in the 1930s and 1940s, randomized clinical trials in humans were being performed by using the statistical-hypothesis-testing paradigm. Randomization gave unbiased comparisons and a way to perform hypothesis testing without model assumptions. To preserve the benefits of randomization, a type of analysis called intent-to-treat analysis is appropriate. Needed development has occurred and is occurring in refining ethical standards, monitoring trials of serious irreversible endpoints while preserving type-I error, and instituting independent data- and safety-monitoring boards. Recent methodology has also been concerned with the appropriateness of using surrogate endpoints. A current area of debate is the appropriateness of using Bayesian statistical methods in this context.

INTRODUCTION

The end of a century and millennium might not be the best time to review and reflect on a specific field of human endeavor, but the symbolism provides additional incentive for us to try. If Pope was correct that "The proper study of mankind is man," then the fields of public health in general and biostatistics in particular are proper endeavors. In this article one of the most outstanding and "proper" contributions of biostatistics to the public health is discussed. That contribution, the randomized clinical trial (RCT), is now a deeply embedded and accepted technique in the evaluation of new therapies, community

0163-7525/99/0510-0109$08.00

interventions, diagnostic techniques, and other areas. This article discusses the RCT in medicine. Biostatistics is of value because of its contribution to other fields; biostatistics is a collaborative, symbiotic field. For that reason any discussion of advances in biostatistics usually and appropriately brings in subject matter concerns, which will also be done in this article. Because clinical trials involve experimentation on humans, the ethical concerns are of primary importance and are addressed below. Of course the emphasis is on the statistical and biostatistical methodology associated with the development of modern RCTs. In addition, areas of current development and debate are presented.

Medical Progress

The consistent progress in medicine over the last 50–55 years gives a misleading impression of the overall history of medicine. The earliest recorded medicine was associated with magic and religion (2, 32). Although there was considerable progress before the twentieth century, the history of medicine also included long periods of adherence to authority, little or no progress, and most importantly, very harmful treatments. Ackerknecht (1) reviews the history of therapeutics, noting that it has been called a "history of errors." Further, "...the history of therapeutics is embarrassing on account of the extraordinary lack of logic, rationality, and openness to experience that is manifest in its history." He attributes the "many reports of success contained in the history of therapeutics where quite obviously the therapeutics could not have produced this success" to four main reasons: (*a*) wrong diagnosis, (*b*) spontaneous recovery, (*c*) the curative effect of suggestion, and (*d*) the forgetting or reinterpretation of failures. Another major reason must be the variability in outcome in most medical situations. Inappropriate behavior is hardest to extinguish when it receives periodic random reinforcement. Such is clearly the case in many medical settings.

STATISTICS AND BIOSTATISTICS

Statistical theory developed from separate paths in probability theory and statistical theory. Only later was it understood that probability theory was the appropriate mathematical foundation of statistical theory.

Probability theory was initially developed to understand gambling. The famous correspondence between Fermat and Pascal in the 1650s (47) is often considered the beginning of probability theory. Cardano (1501–1576) (36) also was instrumental in the early development of probability theory. The history of probability theory is given elsewhere (see 12, 25, 37, 47).

The early field of statistics dealt with methods of estimation (46). The history of statistics (as distinguished from probability theory) up to the twentieth century is given by Stigler (46), with an emphasis on application to the social sciences.

Bayes' Theorem and Inverse Probability

All of those teaching elementary statistics know that, after instruction in statistical inference, students commonly want to interpret the findings as giving probabilities of events. For example, 95% confidence intervals are commonly misinterpreted as indicating a 95% probability that the true value lies within the interval. This problem of inverse probabilities has a long, distinguished history. The first to treat the problem (although not with the generality associated with his name) was the Reverend Thomas Bayes, with the publication of a posthumous memoir in 1764 (5). Pierre Simon Laplace (1749–1827) subsequently published Bayes' Theorem for the special case of prior events with equal probabilities. Bayes' Theorem starts with prior probabilities of events, obtains new data or information related to the events, and then shows how to compute the new posterior probabilities of the events. The longstanding debates about Bayes' Theorem revolve around the necessity to have a prior distribution, that is, prior probabilities for each event. These prior probabilities have been argued to be subjective or personal probabilities (which interact with external probabilities through Bayes' Theorem to give subjective or personal posterior probabilities). The subjectivity of the prior probabilities has been the crux of the debate about the use of Bayes' Theorem in science. It has been persuasively argued that anyone who would bet in a coherent fashion (if forced to bet) would in fact follow Bayes' Theorem with some prior probabilities, regardless of whether this was a conscious act (44).

Another approach to producing posterior probabilities was given by Sir Ronald Alymer Fisher with his introduction of fiducial probability (20, 21). Fiducial probability was generalized mathematically by Fraser (23). This probability avoided the arbitrariness of the selection of a prior probability at the expense of requiring a mathematical formulation that could reasonably be done in a number of ways for the same data, thus leading to different answers. For this reason structural probability has not generally been used and is not even taught in most statistics programs (except sometimes as of historical interest).

Frequentist Probabilities

The debate about inverse probability brings to the fore the philosophic debate about the meaning of probability. The most prevalent view about probabilities is that in many settings probabilities are an external property of the world. For example, quantum mechanics naturally leads to probability models for the decay of atoms and numerous other physical events. This is an example of the frequency interpretation of probability. A probability is defined as the proportion of the time that an event would occur if exactly the same situation were performed (without any interference between different trials of the situation) approaching an infinite number of times. Thus the probability of a cure for a disease would be the proportion of times a cure was obtained among an

infinite number of such patients (i.e. patients with the same disease and the same risk factors as the patient under consideration). Clearly with gambling as the motivation for probability this definition made sense; conceptually one could independently shuffle and deal cards approaching an infinite number of times. For situations that occur only once (e.g. a sporting event or the performance of the stock market over the next six months), the interpretation can lead to conceptual difficulties.

Hypothesis Testing

During the 1920s–1930s, Neyman & Pearson (34) developed an approach to statistical inference that does not require the production of a posterior distribution from a prior distribution. Rather they examined the operating characteristics of procedures when one is forced to make decisions. Some philosophers of science note that truth is never definitively established. A current theory that adequately explains experimental data may subsequently be rejected if the theory does not explain the data from a new data set. Neyman & Pearson developed the testing of "null" hypotheses. This resulted in the now familiar paradigm of null hypothesis, alternative hypothesis, rejection regions, and p-values. The emphasis is on the operating characteristics of tests in which the frequentist probabilities are an inherent property of the external world. The method does not produce posterior probabilities after data are obtained; the p-value instead is the probability, before the experiment is performed and assuming the null hypothesis is true, that a result will be the same as or more extreme against the null hypothesis than that given by the actual data. A key concept is the type-I error—the probability under the null hypothesis that the null hypothesis will be rejected. This probability is called the size of the experiment or the statistical significance level of the experiment. The alternative probability, the type-II error, of rejecting the alternative when it holds, leads to the statistical power of an experiment or observation. Power is the probability of rejecting the null hypothesis when the alternative is true. The p-value is used as a rough measure of the strength of evidence for rejecting the null hypothesis. This is only one possible such measure, and disagreement exists about the appropriate way to summarize the evidence (43). Hypothesis testing is far and away the most widely accepted paradigm for statistical inference within the scientific literature.

RANDOMIZATION

One of the great intellectual advances of the twentieth century was brought about by the English statistician and geneticist Sir Ronald Alymer Fisher (8). He was involved in the analysis of agricultural data sets. The analysis was problematic

because of difficulties inherent in the science of agriculture. For example, the obvious way to compare the yield of different strains of a plant is to plant these strains in adjacent plots and to measure the yield. The problem is that gradients in water drainage, soil, sunlight, wind, and the like can cause major differences in yield. For example, this author observed separate juniper shrubs that were planted along an 80-foot length to establish a hedge. These plants, at least to the uninitiated, appear to have grown under similar conditions, but they now vary in height (in a systematic tall-to-short manner) by a factor of over two. RA Fisher was faced with similar difficulties: Differences in yield could be caused by the strain of plant or might plausibly be caused by environmental differences. His solution seems absurd at first sight. The plot of ground for a comparative experiment was divided into sections. Then the assignment of strain to plot was done at random! That is, the probability of every possible arrangement was the same. This technique deliberately introduced noise into the experiment. However, on average over all the randomized experiments that could have occurred, each strain had an equal chance to get a good or bad plot assignment. Thus the data that actually were observed could be compared with all the possible arrangements to see whether the magnitude of observed differences could have occurred by chance. Adding the "noise" of randomization to the experiment allowed a fair comparison to be made between the strains.

In comparative medical trials of two therapies (including the possibility of a placebo arm in many trials), randomization assigns the two therapies to experimental groups by the flip of coin (as it were). This will be discussed below.

Benefits

There are a number of benefits of the randomized study, and the primary benefits are enumerated below.

BIAS Observational comparison of possibilities is fraught with potential for arriving at a wrong conclusion, not because of a lack of statistical evidence but because one might be comparing relative "apples and oranges." If a comparison has an expected estimated value that is not the value desired, the difference is called bias. Consider a comparison that involves a drug to reduce mortality. Suppose that those who take the drug $\geq$80% of the time have approximately half the mortality of those who do not comply with this medication frequency. Further investigation of $\sim$40 commonly used prognostic variables reveals no difference between the two groups. Finally the observed mortality difference is statistically significant. Most would consider the case proved; the therapy prolongs life. Such a situation actually occurred in the Coronary Drug Project (10), but the drug shown to be efficacious was the placebo! It turned out that the active arm also had the same relationship to drug compliance and

mortality. The biases that can result from observational data analyses can be quite large.

There is ample evidence that humans cannot behave in a fair (that is, statistically unbiased) manner in most situations (28). Simply asking physicians to divide patients into two fair or equal groups for a clinical trial would not usually give an unbiased, or fair, comparison in general. And even if the division were fair, many would not believe in the results.

The process of randomization not only involves a fair process and assures randomly assigned groups are equal (on the average or based on the rules of probability), but there is probabilistic balance even on unknown or unrecorded characteristics of experimental assignments. No unconscious human bias can enter into treatment assignment if the assigned treatment is the result of a randomized assignment.

MODEL ASSUMPTIONS AND THE RANDOMIZATION DISTRIBUTION One of the important methods used in statistics is to model data based on some fixed model and then, typically, an explicit or implicit error term. Important examples of such models are multivariate linear regression models, logistic regression models for binary outcome data, and Cox proportional hazard models for censored time-to-event data. For the analyses to be valid, the data need to conform (at least approximately) to the given model. Although the models can be validated against the data to some extent, there is always limited power for verifying the model assumptions. Consider the source of the variability. This is inherent in the model. For example, in logistic regression the outcome is binary (one of two outcomes), depending only on the variables in the model and in the chosen form. The outcome takes each value with a fixed probability. The variability resides in the external world and needs to follow the assumed form.

For a randomized trial there are two places that random variability can enter into the outcome. As just discussed there is the usual variability associated with a model for the outcome. However, with a randomized experiment there is another source of random variability, for example the randomization process that assigns plant strains to plots or the randomized assignment of a drug or placebo to a patient.

Now consider a different point of view: Suppose that we have a randomized assignment and that the null hypothesis holds. The null hypothesis is that the randomized assignment has no effect on the outcomes of interest. If this is true we may consider that the outcomes observed would have occurred under any treatment assignment. Therefore think of the outcomes of the individual experimental units as fixed. Under the null hypothesis, no matter what the assignment of the randomized process, we would have seen the same results. If there is a statistic to test the null hypothesis that reasonably measures a treatment

effect, it may be used without assumption as follows. Thinking of the patient outcomes as fixed values, the distribution of the test statistic comes from the randomized assignments. For each different possible randomized assignment, the value of the test statistic that would have resulted may be computed; from this probability of each randomized assignment, the distribution of the test statistic may be computed. This distribution is called the randomization distribution. Its validity does not depend on the appropriateness of a statistical model; the test statistic may involve a model with covariate values for adjustment or with complexity in any other manner. In any event the distribution of the test statistic depends only on the randomization process for its validity.

Having noted this valuable property of randomized experiments, one would expect randomization distributions to be the standard method of analysis for randomized studies. Such is not the case. There appear to be two reasons why the usual models are used: (*a*) The computation of the randomization distribution is prohibitive in many situations. If 100 subjects were allocated at random into two groups with 50 in each group, there are over 10^{29} possibilities. This is too many to enumerate the possibilities and compute the randomization distribution. Only with the recent advances in computing power is the approximate answer available. The randomization distribution may be estimated by simulating multiple random samples from the randomization process. Such simulation, or Monte Carlo simulation, may be used to give a p-value without assumptions (7, 14). (*b*) The historical use of the usual tests (which usually are valid) has not led to a perceived need for the randomization distribution.

RANDOMIZED CLINICAL TRIALS

With this background we now turn to the primary subject of this paper, development of the RCT in the twentieth century. This development has primarily taken place since World War II. The ethical issues inherent in human experimentation were at the forefront after the Nazi "medical" war crimes (31). Ethical issues are discussed below; there has been continual development in this area.

Initial Trials

Among the earliest randomized trials was one by Amberson (3) in 1931 (cited in 33; see 33 for a short history of early clinical trials and references). An influential figure in the development of early clinical trials was Austin Bradford Hill, who published a text on clinical trials and entered into the ethical justification for clinical trials (26, 27). The appropriate benefits of randomization (lack of bias or creation of comparable groups on the average and ability to compute p-values without model assumptions) are immediately applicable to human experimentation in medicine. The ethical concerns and the need for physicians

to both discuss their ignorance and allow another to decide on the (random) treatment must have made the early implementation of RCTs a delicate political undertaking. The rapid acceptance of RCTs and the continual growth in their use attest to the scientific cogency and value of the resulting knowledge.

Ethical Issues

Most individuals feel some distaste and concern when they first hear about the concept of human experimentation. The more historical system of the physician bravely (few discussions cite the bravery of the patient) and boldly trying a new therapy seems more appealing and dramatic. Yet without controlled experimentation, much less knowledge is obtained, and ethical concerns may not be appropriately addressed.

BACKGROUND Some argue that an appropriate ethical stance can be derived from appealing to a priori, or at least fixed, principles (6). The Nuremberg war crime trials involved consideration of physicians involved in unethical medical experimentation. This resulted in the Nuremberg Code (42). Subsequently the international physician community addressed medical experimentation in the Declaration of Helsinki and its periodic revisions (49). Among the generally agreed upon principles are the right of patients to informed consent [with some possible exceptions in trials in which informed consent is impossible to obtain (e.g. resuscitation for cardiac arrest)] and the necessity of review by an independent body without the potential for profiting from the particular RCT. In most countries these issues not only are required by law but often have very detailed federal regulations about their implementation (15).

THERAPEUTIC IMPERATIVE Among the many ethical issues of RCTs, the most widely debated is the conflict between (*a*) the implicit contract between the physician and patient that the physician will deliver the best care available in her/his opinion and (*b*) the need for a random assignment to provide scientific knowledge (17). This implicit contract has been called the therapeutic imperative. Some feel that the inherent conflict between the physician as a physician and the physician as a scientist has so much contradiction that randomized trials are unethical de facto. A majority of the medical community feel that such trials are ethical if run according to certain principles that assure appropriate consideration and protection of patient rights.

EQUIPOISE Because of the physician's implicit contractual obligation to deliver to the patient the best possible medical care (at least for serious endpoints), the concept of equipoise has been developed. Equipoise means that there are equal chances of any of the treatment arms of a trial being the most efficacious. Some risk may be allowed for minimal-risk protocols, but for serious

irreversible endpoints equipoise should be in place to make a trial ethical. The patient risk compared to study benefit is important in considering whether a trial is ethical.

SOCIETAL TRADEOFFS Some would argue that, in a world with limited resources, the benefit of a therapy must also be related to the resources consumed [e.g. is it good medicine to spend $500,000 per patient for an average gain in life of 3 weeks? Or should society let such issues become prominent (including which trials should be conducted and whether cost/resource use should also be a part of such trials)]. This author would argue that many more trials would actually both (*a*) help limit the costs of medical care and (*b*) give more reliable information for the public debate and implementation of the health care system. This would lead into a debate far beyond the scope of this paper.

Minimal Level of Proof

Because of the uncertainty associated with small numbers of observations, some level of proof is needed before accepting data as evidence of some fact. Traditionally for scientific publication, a p-value of ≤ 0.05 is satisfactory. Such results are called statistically significant (and may or may not be clinically significant if true). In any setting the level of proof needed is arbitrary, but clearly the concept of a minimal level of proof is desirable.

REGULATORY CONCERNS AND THE TYPE-I ERROR RATE The control of the type-I error is taken particularly seriously in a regulatory setting. A commercial concern with tens or even hundreds of millions of dollars invested has great incentive to get the new drug, biologic, or device approved. (A biologic is a drug made of a compound that naturally occurs in the human body.) It is to be expected that a commercial sponsor will put the best possible light on their data and apply the most favorable possible method of data analysis. For this reason society has decided (through their elected representatives) that the level of proof must meet a very high standard. We discuss this approval process as regulated in the United States by the Food and Drug Administration (FDA). There are a number of reasons for tight control and emphasis on the type-I error rate. First, once regulatory approval is granted it is difficult to withdraw approval unless a sponsor agrees. If a sponsor disagrees, then the issue goes to court, with the US Food and Drug Administration needing to show a lack of efficacy or safety—the shoe is on the other foot. In the United States, with the possible exception of trials with serious irreversible endpoints as the outcome measure, two statistically significant trials have been required for approval. For one large trial this would correspond to a maximum p-value of 0.00125. Currently there is discussion about whether such a level should be required for serious irreversible endpoints such as death.

Because human life and well being are at stake, it is often considered unethical to replicate a placebo-controlled RCT that established the efficacy of a treatment. One basic principle of science is the replicability of results. In medicine this principle does not hold for ethical reasons. Therefore "mistakes" can have very serious and far-reaching consequences. This argues for a strong level of proof before the medical community accepts a therapy as proven.

CONSISTENCY AND MEASURES OF THE LEVEL OF PROOF The level of proof as measured by the p-value has weaknesses. Recently, Bayesian approaches to determining the appropriate level of proof have been advocated (45). If true personal prior beliefs are advocated, then one has the large and controversial task of deciding whose prior belief to use. Use of multiple fixed prior distributions, including various pessimistic prior distributions, has been advocated (45). If these prior distributions do not reflect beliefs of at least a community, then the benefit of Bayesian statistics as bringing past knowledge (belief?) to bear is lost, and the endeavor is essentially frequentist in nature. Reasons for the inappropriateness of true Bayesian statistical methods in the RCT setting have been advanced (16).

One concern that receives relatively little consideration is the implicit conflict between studies being run while in equipoise and the accumulation of proof to some minimal or acceptable level. As soon as information begins to accumulate, equipoise is lost, however slightly. Further, if trials need to accumulate a minimal level of proof, there is assumed to be little or no proof even when a small increment of data will push the level of proof over the line of acceptability. Implicitly it is considered appropriate to place participants at some possible risk before the acceptable level of proof is reached. By requiring proof in federal law and regulation, the United States has decided that the risks associated with drug, biologic, and device approval without adequate levels of proof justify some conflict with the therapeutic imperative, to establish new therapies and their usage before approving the therapies for general use.

Blinding or Masking

Another aspect of RCTs that attempts to minimize bias in therapeutic comparisons is the blinding or masking of therapies. In a single-blind or single-masked trial, the subject or patient does not know what treatment she or he is receiving. In a double-blind or double-masked trial, neither the person delivering nor the one receiving the therapy knows which treatment has been assigned. Placebo treatments are often developed to aid in blinding. The placebo has the broadest indication in medicine; it is effective to a greater or lesser extent in almost all medical settings. If those evaluating the success of treatment arms are blinded

to treatment, then their conscious or unconscious biases cannot enter into the evaluations (when the blinding is perfect).

BIOSTATISTICAL RCT METHODOLOGY

Control of Type-I Error and the Possibility of Early Stopping: Sequential Monitoring of RCTs

When a trial with a serious irreversible endpoint is conducted in humans, it is necessary to monitor the data as it accumulates. If a given arm of a trial can be shown to be superior, then such a trial will be stopped. However if multiple examinations of accumulating trial data are made, then one cannot stop a trial whenever the current value of a fixed sample size p-value reaches the required level of statistical significance for the whole trial. Clearly, with multiple looks, stopping a trial when a required p-value is reached would elevate the probability of accepting a chance finding when in fact the null hypothesis was true (4). For this reason biostatisticians have developed methods of allowing examination of the accumulating data of a clinical trial with the potential of early stopping, while also maintaining the overall type-I error rate for the trial (18, 30, 35, 40, 48). Although formal Bayesian methods have rarely been used in this context, they hold the potential for such use if judged appropriate (45). The use of sequential strategies or monitoring boundaries has been developed largely for specific use in RCTs.

DATA AND SAFETY MONITORING BOARDS The formal, mechanistic monitoring of RCTs for safety and efficacy with serious irreversible endpoints could algorithmically rely upon a sequential monitoring plan. However, all commentators have agreed that the formal stopping rules are only guidelines, which must be used in the total context of the trial. For example, there might be chance baseline imbalances between treatment groups that lower the evidentiary value of a difference that otherwise would terminate a trial. The imbalance may suggest the need for further data collection. Other trials with drugs that have a similar mechanism of action might strengthen or weaken the findings. Because sponsors have a vested interest that might unconsciously influence their decisions, data and safety monitoring boards have been established (13). Such boards are composed of individuals who will not potentially gain or lose from the results of the trial. They thus can make a decision without the potential commercial conflict of sponsor employees. Such boards typically have members who are physicians with appropriate specialties, biostatisticians, and possibly ethicists or lay members. In considering termination of a trial for undue risk or for efficacy, the risk/benefit ratio is appropriately taken into account. Another

function such boards often serve is to monitor for futility, that is, situations in which the trial has little chance of a positive outcome, and stop such trials early to conserve resources.

Intention-to-Treat Analysis

The benefits of randomization were described above. The most important benefit is to construct fair or unbiased treatment groups for comparison. The groups are balanced in a statistical sense, even against unknown or unrecorded covariates. To preserve the benefits of the randomization process, individuals are to be maintained and analyzed in their assigned groups (19, 39). In the (never observed) perfect medical experiment, everyone agrees that the intent-to-treat (ITT) analysis is appropriate. However, if a number of subjects never receive a treatment, does it make sense to include their results in the treatment group? This tension between the biostatistical/scientific need for an unbiased comparison (the ITT analysis) and the biologic/scientific analysis that considers only treated or compliant patients can usually, but by no means always, be avoided by an appropriate experimental design. In most regulatory settings, the ITT analysis is the first analysis expected. Rarely does the ITT analysis miss proving the efficacy of a new treatment and the RCT result in regulatory approval from some other analysis.

Surrogate Endpoints

It would seem obvious that the purpose of giving a therapy is to benefit the patient. To develop beneficial therapies an understanding of the biology, even to the molecular level, is often used. If the understanding were sufficient, including knowledge of potential adverse effects of therapy, it would be enough to show that the drug affected some intermediate associated factor. For example, early trials with placebo controls showed that antihypertensive therapy prevented cardiovascular endpoints. It is currently accepted that lowering blood pressure benefits patients. Current approval of antihypertensive drugs depends on placebo-controlled trials with a duration of 12–16 weeks in moderately hypertensive patients. This is supplemented by longer-term uncontrolled exposure of a year or more. The benefit is inferred from the surrogate endpoint of lowered blood pressure (rightly or wrongly). On the other hand many cancer therapies that shrink tumors (that is, that have been shown to be biologically active) apparently do not benefit patients.

The most famous recent experience with surrogate endpoints was the Cardiac Arrhythmia Suppression Trial (9). Irregular heartbeats (cardiac arrhythmia) have been studied with ambulatory electrocardiograms that monitor cardiac rhythm over $\geq$24 hours. Unfortunately, some individuals being monitored have died by a sudden cardiac death. There were typical findings on the ambulatory

electrocardiogram. Runs of rapid, ventricular, premature beats (ventricular tachycardia) lead to a fluttering irregular motion of the heart (ventricular fibrillation), which leads to sudden unconsciousness and rapid death. Graboys et al (24) showed that individuals with high-risk arrhythmia had a much better survival if their arrhythmia could successfully be treated with an antiarrythmic drug, compared with patients in whom no such drug could be found. On this basis, suppression of arrhythmia on ambulatory electrocardiographic monitoring was considered an adequate surrogate for drug approval as an antiarrhythmic drug. The Cardiac Arrhythmia Suppression Trial studied, in a placebo-controlled fashion, individuals with arrhythmia after a heart attack (myocardial infarction). The investigators were so sure that, if not beneficial, the therapy at least was not harmful, that the study was designed with a one-sided hypothesis test at the 0.025 significance level. The trial was stopped for excess mortality. Three antiarrhythmic drugs were studied; all three proved to be harmful! Thus surrogate endpoints must be carefully chosen.

Prentice gives conditions that would allow appropriate use of a surrogate endpoint (41). Others have studied the use, within a trial, of changes in a potential surrogate in conjunction with the primary endpoint to increase the statistical power (38). With pressure to shorten the period for drug approval, the use of surrogate endpoints is very tempting. History has shown that such reliance must be used judiciously in limited contexts (22).

Time-to-Event Analysis

One of the substantial advances in biostatistics has been the ability to handle data for time to events with different lengths of observations for different subjects. In particular, many subjects have not experienced the endpoint(s) when the period of observation ends. Such censored data were first studied by actuaries constructing life tables-of-survival data. Efficient use of such information was then studied by Kaplan & Meier (29). Later the effect of covariates was introduced into parametric models and also into the Cox proportional hazards regression model (11). These time-to-event models are used in many RCTs, as well as for observational data analysis of biomedical data. Advances in this methodology continue with research to allow multiple events, multiple types of endpoints, etc.

COMMENTARY

The randomized clinical trial is a development of the twentieth century. It builds on a number of historical developments. Among these are the development and then merging of probability theory and statistics. The appreciation of the scientific method became apparent in the “hard” sciences initially but moved over to

the social and biologic sciences. Sir RA Fisher's introduction of the randomized experiment in agriculture and biology was then adapted to medical experimentation in humans shortly after the Second World War. Full application of the method needed development in a number of areas. (*a*) The ethics of human experimentation needed development (especially in light of the Nazi experiments in World War II). Among the accepted principles were the right to informed consent and independent review of experimental protocols for human experimentation. (*b*) For serious irreversible endpoints, there is an ethical mandate for monitoring the results during the trial. This must be done in a manner that preserves the type-I error of the hypothesis-testing paradigm. Multiple such methods have been developed. (*c*) To perform such monitoring, independent data- and safety-monitoring boards are often used. The boards use statistics as guidelines for early stopping but must also consider other relevant factors.

The use of RCTs has led to theoretical consideration and development in multiple areas, including the use of surrogate endpoints, time-to-endpoint analysis, combinations of multiple endpoint measures, and other areas. Further, the considerable governmental and commercial activity in the area of clinical trials has helped to integrate biostatistical methods into the medical research community. As an offshoot of this activity, one would conjecture that observational analyses have been advanced in methodology as well as made more sensitive to the limitations of observational data analysis.

Biostatistics as a field has developed only in the twentieth century. One of the great successes of biostatistics, as a collaborative part of the medical research community, has been the continuing development and implementation of the RCT.

Literature Cited

1. Ackerknecht EH. 1973. *Therapeutics: from the Primitives to the 20th Century*, pp. 1–3. New York: Hafner
2. Ackerknecht EH. 1968. *A Short History of Medicine*, pp. 10–18. New York: Ronald
3. Amberson JB Jr, McMahon BT, Pinner M. 1931. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am. Rev. Tuberc.* 24:401–35
4. Armitage P, McPherson CK, Rowe BC. 1969. Repeated significance tests on accumulating data. *J. R. Stat. Soc. Ser. A* 132:235–44
5. Bayes T. 1763. An essay toward solving a problem in the doctrine of chances, with Richard Price's foreword and discussion. *Philos. Trans. R. Soc. London* 370–418
6. Beauchamp TL, Childress JF. 1989. *Principles of Biomedical Ethics*. New York: Oxford Univ. Press. 3rd ed., pp. 49–55
7. Birnbaum ZW. 1974. Computers and unconventional test-statistics. In *Reliability and Biometry: Statistical Analysis of Lifelength*, ed. F Proschan, RJ Serfling, pp. 441–58. Philadelphia: SIAM
8. Box JF. 1978. *R. A. Fisher: The Life of a Scientist*, p. 146. New York: Wiley
9. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. 1989. Preliminary

report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N. Engl. J. Med.* 321:406–12
10. Coronary Drug Project Research Group. 1980. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N. Engl. J. Med.* 303:1038–41
11. Cox DR. 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc. Ser. B* 34:187–220
12. David FN. 1962. *Games, Gods and Gambling: The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*. London: C. Griffin
13. DeMets DL, Ellenberg S, Fleming TR, Childress JF, Mayer KH, et al. 1995. The data and safety monitoring board and acquired immune deficiency syndrome (AIDS) clinical trials. *Contr. Clin. Trials* 16:408–21
14. Edgington ES. 1995. *Randomization Tests*. New York: Dekker. 3rd ed.
15. Federal Register. 1988. Institutional Review Boards. *Fed. Regist.* 21(I-56)
16. Fisher LD. 1996. Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Contr. Clin. Trials* 17:423–34
17. Fisher LD. 1998. Ethics of randomized clinical trials. In *Encyclopedia of Biostatistics*, ed. P Armitage, T Colton, 2:1394–98. New York: Wiley
18. Fisher LD. 1998. Self-designing clinical trials. *Stat. Med.* 17:1551–62
19. Fisher LD, Dixon DO, Herson J, Frankowski RF, Hearron MS, Peace KE. 1990. Intention to treat in clinical trials. In *Statistical Issues in Drug Research and Development*, ed. KE Peace, pp. 331–50. New York: Dekker
20. Fisher RA. 1930. Inverse probability. *Proc. Cambridge Philos. Soc.* 26:528–35
21. Fisher RA. 1935. The fiducial argument in statistical inference. *Ann. Eugenics* 6:391–98
22. Fleming TR, DeMets DL. 1996. Surrogate end points in clinical trials: Are we being misled? *Ann. Intern. Med.* 125:605–13
23. Fraser DA. 1968. *The Structure of Inference*. New York: Wiley
24. Graboys TB, Lown B, Podrid PJ, DeSilva R. 1982. Long-term survival of patients with malignant ventricular arrhythmia treated with antiarrhythmic drugs. *Am. J. Cardiol.* 50:437–43
25. Hald A. 1990. *A History of Probability and Statistics and Their Applications Before 1750*. New York: Wiley
26. Hill AB. 1962. *Statistical Methods in Clinical and Preventive Medicine*. New York: Oxford Univ. Press
27. Hill AB. 1963. Medical ethics and controlled trials. *Br. Med. J.* 1:1043–49
28. Kahneman D, Slovic P, Tversky A, eds. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge Univ. Press
29. Kaplan EL, Meier P. 1958. Nonparametric estimation for incomplete observations. *J. Am. Stat. Assoc.* 53:457–81
30. Lan KKG, DeMets DL. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–63
31. Lifton RJ. 1986. *The Nazi Doctors: Medical Killing and the Psychology of Genocide*, pp. 18, 48–79. New York: Basic Books
32. Magner LN. 1992. *A History of Medicine*, p. 9. New York: Dekker
33. Meinert CL. 1986. *Clinical Trials: Design, Conduct, and Analysis*, pp. 3–8. New York: Oxford Univ. Press
34. Neyman J, Pearson ES. 1966. *Joint Statistical Papers of J. Neyman & E. S. Pearson*. Berkeley, CA: Univ. Calif. Press
35. O'Brien PC, Fleming TR. 1979. A multiple testing procedure for clinical trials. *Biometrics* 35:549–56
36. Ore O. 1965 (1953). *Cardano: The Gambling Scholar*. New York: Dover
37. Pearson ES, Kendall MG. 1970. *Studies in the History of Statistics and Probability: A Series of Papers Selected and Edited by E. S. Pearson and M. G. Kendall*. Darien, CT: Hafner
38. Pepe MS, Reilly M, Fleming TR. 1992. Auxiliary outcome data and the mean score method. *J. Stat. Plan. Inf.* 42:137–60
39. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, et al. 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34:585–612
40. Pocock SJ. 1982. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 38:153–62
41. Prentice RL. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* 8:431–40
42. Reiser SJ, Dyck AJ, Curran WJ, eds. 1977. The Nuremberg code. In *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*, pp. 272–74. Cambridge, MA: MIT Press
43. Royall R. 1997. *Statistical Evidence: A Likelihood Paradigm*, pp. 35–81. New York: Chapman & Hall

44. Savage LJ. 1972 (1954). *The Foundations of Statistics*. New York: Dover. 2nd ed.
45. Spieglehalter DJ, Freedman LS, Parmar MKB. 1994. Bayesian approaches to randomized trials (with discussion). *J. R. Stat. Soc. Ser. A* 157:357–416
46. Stigler SM. 1986. *The History of Statistics*. Cambridge, MA: Harvard Univ. Press
47. Todhunter I. 1949. *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace*. New York: Chelsea
48. Whitehead J. 1983. *The Design and Analysis of Sequential Clinical Trials*. Chichester, England: Horwood
49. World Medical Association. 1977 (1964). Declaration of Helsinki. In *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*, ed. SJ Reiser, AJ Dyck, WJ Curran, pp. 328–30. Cambridge, MA: MIT Press