III-5

Historical Controls, Data Banks, and Randomized Trials in Clinical Research: A Review[1,2]

Thomas R. Fleming[3,4,*]

SUMMARY

A brief review is presented of the strengths and weaknesses of historical controls, data banks, and randomized trials in the evaluation of clinical treatment. Use of prerandomized versus postrandomized informed consent is discussed. Recommendations are made for the development of an appropriate clinical research strategy.

[Cancer Treat Rep 66:1101–1105, 1982]

The use and the sophistication of scientific techniques in clinical experimentation have increased. In the words of Marvin Zelen (1), "The day is rapidly receding when ex-cathedra judgments dominate therapeutics." However, a continual problem faces the clinical investigator who wants to evaluate a new treatment and to compare its efficacy with that of an existing "standard" treatment. How can information enabling this comparison be obtained in a manner that is ethically acceptable, that allows one to make an unbiased or fair evaluation, and that is the most efficient possible? Here, "efficiency" refers to minimizing the number of patients and medical personnel, the amount of money, and the length of time needed for the assessment. Methods for providing the necessary information include the use of a contemporary treatment group compared with historical controls, the use of data banks in which patients with and without the treatment in question are compared, or the performance of randomized trials. In the last-named instance, informed consent either is always obtained before the time of randomization or is only obtained after. In this volume, these four methods are discussed by Gehan (2), Starmer and Lee (3), Simon (4), and Zelen (5), respectively.

These and many other investigators have explored the strengths and weaknesses of each methodologic approach in clinical research. These issues are considered in the second and third sections of this paper. In the fourth section, the general problem of developing an appropriate clinical research strategy is addressed and recommendations are made.

Strengths and Weaknesses of Methods to Select Controls—A Brief Review

In an attempt to make efficient use of limited research resources, certain clinical trials are designed to use historical controls. This design is particularly appropriate when one evaluates experimental treatment regimens that are not only hoped to be but also expected to be clearly superior to standard therapies.

Unfortunately, substantial heterogeneity often exists in the patient population eligible for any given clinical trial, particularly in cancer research. Thus, whenever historical controls are used, great care must be taken to minimize selection bias, that is, to minimize the lack of comparability between the experimental treatment group and the historical control group. To this end, Pocock (6) has suggested important criteria to be implemented in designing a historically controlled study. Gehan and Freireich (7) have emphasized the role of quantitative models that can be used during the comparison of treatment groups to adjust for the intergroup dissimilarities

[1]Received Mar 4, 1982; accepted Mar 4, 1982.

[2]Supported by Public Health Service grant CA-24089 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

[3]Department of Medical Statistics and Epidemiology, Mayo Clinic and Mayo Foundation, Rochester, MN.

[4]I express my appreciation to the many colleagues at Mayo who provided helpful comments on this paper.

*Reprint requests to: Thomas R. Fleming, PhD, Department of Medical Statistics and Epidemiology, Mayo Clinic, Rochester, MN 55905.

in identified patient characteristics influencing prognosis. Unfortunately, as noted by Simon (4), Byar (8), Pocock (9), and Peto (10), much of the heterogeneity in the study population is not accounted for by the prognostic variables that are known and properly recorded. Thus, when using historical controls, one must exercise great caution in drawing conclusions about treatment comparisons. One cannot distinguish differences truly due to treatment from differences due to systematically occurring intergroup dissimilarities arising from important but unrecognized or unrecorded covariates.

Random assignment of patients to treatment groups eliminates the systematic occurrence of lack of comparability among these groups. This property is referred to as "unbiasedness" by Simon (4). Randomly occurring dissimilarities in recorded or unrecorded covariates can still appear among treatment groups in randomized trials, although in very large samples these dissimilarities also disappear. Simon calls this property "consistency." The property of unbiasedness, particularly in the elimination of systematic dissimilarities that occur among treatment groups and that are caused by unknown or unrecorded prognostic factors, is a fundamental component to a fair, reliable, and reproducible evaluation of treatment. Only randomization guarantees unbiasedness.

Certainly, an increased investment of research resources is required to perform a randomized trial. However, this increased investment, together with the unbiasedness and increased reliability of results obtained from such a design, leads to a greater willingness to publish not only the positive results but also the negative results. This lessens the reporting bias that arises in clinical research when, as Zelen (5) states, "Editors and even authors are reluctant to publish negative results."

With the rapid and extensive development of computer hardware and software as well as statistical methodology, the use of data banks has been supported in some treatment evaluations as an alternative to randomized or historically controlled trials. The data bank or data-based approach, when constructed in the careful manner followed by Starmer and Lee (3), is indeed useful. The data base, a shared repository of data elements describing attributes of patients, interventions, and outcomes and collected uniformly over long periods, is often well-suited for identifying prognostically important variables, for studying temporal shifts in patients' disease and in technology, and for investigating the feasibility and desirability of performing confirmatory randomized trials of certain comparisons of treatment. Data bases that are population-based can be used as well in the assessment of incidence and mortality of cancers.

In chronic disease, Starmer and Lee (3) propose that the data-based approach be used for treatment evaluation and individualized selection of treatment because the prolonged course between onset, intervention, and target outcome and the nonstationary disease process and milieu render randomized designs impractical. However, the concerns of Simon (4) and Byar (8) about bias in treatment assignment, missing data, and lack of standardization of treatment and response assessment should be carefully heeded when data-based treatment evaluation is considered. Bias will arise whenever treatment decisions are made not only on the basis of elements in the data bank but also on other important but unrecorded considerations. In addition, how data-based, individualized selection of treatment could result in the selection of newly proposed therapies is uncertain.

Prerandomized or Postrandomized Informed Consent?

Are the treatments potentially medically equivalent? This is a key ethical question that must be addressed when one is considering whether to randomize patients in a cancer clinical trial. If the answer is yes and the decision is made to conduct such a trial, the Code of Federal Regulations (45 CFR 46) requires, in essence, that each patient who could be asked to receive therapy "which departs from . . . established or accepted methods" give informed consent, the purpose of which is to facilitate and guarantee informed decisions from each participating patient. Information to be provided to the patient includes a discussion of the risks and benefits of each of the treatments in the study and of potentially effective alternative therapies.

The process of obtaining informed consent is often lengthy. This, together with Zelen's observation (1) that "physicians may find it difficult to tell patients that they do not know which treatment is best," has resulted in substantially reduced accrual rates in many randomized trials.

To alleviate this problem, Zelen (1) has proposed that patients be randomized before informed consent is obtained. Supposing randomization is to either a "standard" or an "experimental" treatment, he suggests that consent be obtained only from those randomized to the experimental regimen. Patients randomized to the experimental regimen must be analyzed with that group whether or not they consent to that treatment. Even though this dilutes true differences in treatment, Zelen's design could

yield an increase in efficiency over the classic prerandomized informed consent design if it leads to an increased rate of accrual and if the consent rate is fairly high among those randomized to the experimental group.

Zelen's design (1) presents some ethical problems not present in the classic situation wherein randomization occurs only after informed consent has been obtained. The first problem concerns those patients who are randomized to the experimental group. Recall that, ethically, randomization requires that investigators have decided that the treatments being compared are potentially medically equivalent. In turn, this assessment of equivalence should be conveyed to the patients if they are to be fully informed. However, to increase the efficiency of the study, investigators need as many patients as possible who are randomized to the experimental regimen to elect to receive that treatment. Thus, the design provides a subtle encouragement for the investigator to provide a biased presentation of the relative merits of the treatments to these patients.

The second ethical problem concerns those patients who are randomized to the standard regimen. In Zelen's single randomized consent design (*see* fig 2 in Zelen [5]), these patients are not approached for consent to enter the clinical trial. He states, "The physician need only approach the patient to discuss a single therapy. The physician need not leave himself open, in the eyes of the patient, to not knowing what he is doing. . . ." (1). Important ethical questions must be addressed. Do these patients, as well, have a need for informed consent? Recall that the intent of the informed consent process is to provide the patient with a discussion of the risks and benefits of the various therapeutic options in order to facilitate and guarantee informed decisions. Doesn't this need exist, then, whenever there is more than one potentially effective therapy for a given patient, in fact, even for the individual being considered for treatment in a nonresearch setting? Fost (11) has stated, "To initiate any treatment, standard or experimental, without the patient's fully informed consent is to fail to respect him as a person with a right to autonomy." If one accepts this statement, it follows that patients who are randomized to the standard regimen in Zelen's design (1) continue to have the same need and right to be fully informed and to formally consent. If this right is satisfied, the newly proposed design usually will lose much of its appeal. An example of this issue is early breast cancer in which the standard regimen in a given trial is modified radical mastectomy and the experimental regimen is excision plus radiation therapy. Certainly, under any circumstances, a woman being considered for mastectomy has a right to be informed of potentially effective alternative therapies.

In summary, for Zelen's design (1) to be useful, one must identify situations in which patients receiving a certain treatment are ethically required to be fully informed of alternative therapies and situations in which another treatment exists that is potentially medically equivalent to the first but for which it is acceptable not to inform patients of the alternatives. In addition, one must anticipate that a high proportion of patients who are randomized to the treatment requiring consent will elect to receive it.

Appropriate Strategy in Clinical Research

In the words of Simon (4), "Good clinical therapeutic research identifies useful treatments, provides reliable leads, or appropriately categorizes useless or harmful treatments." If one accepts this statement as the general goal in clinical research, one must consider what Simon refers to as the "thermodynamics of clinical trials," that is, the aggregate results from a program of clinical research. Simon (4) and Zelen (5) have provided clear illustrations of the importance of this consideration.

Well recognized in the design of a clinical trial is the need to properly control the size (α) and the power ($1 - \beta$) of the trial. Here α denotes the false-positive probability, that is, the probability of concluding that differences exist when, in fact, there are no true differences, and $1 - \beta$ denotes the probability of detecting clinically meaningful differences that truly exist. However, not as commonly considered when the trial is designed is π, the pretrial estimate of the probability that clinically important true differences exist. Knowledge of α, $1 - \beta$, and π permits the evaluation of such important questions as "What is the probability that a positive result in our trial will be a true positive result?" One can see in table 1 that the answer is

$$\frac{\pi(1-\beta)}{\pi(1-\beta) + (1-\pi)\alpha}.$$

TABLE 1.—Relationship of experiment results and truth

Result of experiment	Truth		
	Positive	Negative	
Positive	$\pi(1-\beta)$	$(1-\pi)\alpha$	$\pi(1-\beta) + (1-\pi)\alpha$
Negative	$\pi\beta$	$(1-\pi)(1-\alpha)$	$\pi\beta + (1-\pi)(1-\alpha)$
	π	$1-\pi$	1

In many clinical situations (*see* Simon [4] and Zelen [5]), a realistic pretrial estimate of the probability that true clinically meaningful differences exist is $\pi \approx 0.20$ (*or 0.10*). Then, even if trials are constructed to have a size of only $\alpha = 0.05$ and a power as high as $1 - \beta = 0.80$, 20% (*or 36%*) of positive results will be false-positive. Unfortunately, the actual situation is often even worse. Studies may be designed to have sample sizes adequate to provide a size of $\alpha = 0.05$ and a power of $1 - \beta = 0.80$ if the patient population is homogeneous. Typically, however, the sample of patients is quite heterogeneous, causing a loss in power. Additional false-positive and false-negative conclusions are introduced when trials are terminated early as a consequence of interim analyses based on test procedures that are appropriate only if applied at the predetermined time of final analysis. The impact of this is that the clinical trial, as actually carried out, may have a size closer to $\alpha = 0.20$ and a power closer to $1 - \beta = 0.60$, which, if $\pi = 0.10$, would result in 75% of positive trials being false-positive.

One other very common source of false-positive conclusions is the consideration of data-suggested hypotheses arising from "data dredging" or "exploratory data analysis." As mentioned earlier, the population of patients receiving treatment is generally quite heterogeneous. Since it is well recognized that the relative benefits of various treatment regimens may vary from one type of patient to another, it is common practice to perform a retrospective search through the data to find subsets of patients showing large differences in treatment. This retrospective search, referred to as exploratory data analysis, can be instructive, yet all too frequently it provides false leads, as illustrated by Starmer and Lee (3).

What, then, does constitute an effective strategy in clinical research? The key, discussed by Gehan (2), Starmer and Lee (3), Simon (4), and Zelen (5), is the recognition of the need for independent confirmatory trials. Simon speaks of treatment evaluation and comparisons occurring in two stages. Stage I comprises small- or moderate-sized screening trials that would encourage innovation. In some situations, these trials should be randomized. Yet it is in the setting of screening trials that historically controlled or data-based trials are useful as well. The increase in efficiency of carefully performed historical and data-based designs could outweigh their lack of "unbiasedness," since a confirmatory trial (or trials) would be performed to verify results.

Unless results in stage I provide extreme evidence in favor of differences in treatment, all positive trials should then be confirmed in stage II in large, randomized clinical trials. These trials generally should have at least 100–200 patients in each treatment group. Performing these large, definitive stage II confirmatory trials only after positive stage I studies ensures that Zelen's conclusion (5) (α) will be met, that is, definitive clinical trials should be initiated only if π, the pretrial estimate of the probability that clinically important gains do exist, is > 0.05.

Since only positive stage I trials will be confirmed in stage II, the power, $1 - \beta$, of the stage I studies should be high enough that frequent missing of clinically important true differences is avoided. Specifically, in the screening trial, $1 - \beta$ preferably should be at least 0.90. As a result, since sample sizes in screening trials are small or moderate, the size of a stage I screening trial often needs to be > 0.05.

Ethically required interim analyses should be performed during these large stage II trials by methods that maintain the desired upper limit on the number of false-positives and that maintain sensitivity to treatment-caused differences of interest (*see* Fleming et al [12]).

It is in the setting of the large stage II trials that sufficient data exist for one to perform useful exploratory data analyses. Since as much as possible needs to be learned from the data, this practice of retrospective exploration of the data should be continued for as long as investigators recognize that the only hypotheses being definitively evaluated by the stage II trial are those pretrial hypotheses that motivated the study. Data-suggested hypotheses formulated during exploratory analyses of data from stage II trials should in turn be confirmed in subsequent trials.

One of the greatest causes of difficulty in carrying out this two-stage clinical research strategy is the tendency to overreact to positive results obtained in the small- or moderate-sized stage I screening trials. In manuscripts, these results often are presented as though they provide a reliable assessment of treatment rather than screening information requiring confirmation. In fact, this overreaction occasionally has led to changes in medical practice so firmly accepted that it has been judged to be unethical to consider randomization against future alternative treatments. It is useful when obtaining these positive screening results to recall the very high chance that a positive result in a trial of this size will actually be a false lead and the subsequent negative impact of accepting an inferior treatment. In the words of Artemus Ward, "It isn't so much the things we don't know that get us in trouble. It's the things we know that aren't so."

REFERENCES

1. ZELEN M. A new design for randomized clinical trials. N Engl J Med 300:1242–1245, 1979.
2. GEHAN EA. Design of controlled clinical trials: use of historical controls. Cancer Treat Rep 66:1089–1093, 1982.
3. STARMER CF, and LEE KL. A data-based approach to assessing clinical interventions in the setting of chronic disease. Cancer Treat Rep 66:1077–1082, 1982.
4. SIMON R. Randomized clinical trials and research strategy. Cancer Treat Rep 66:1083–1087, 1982.
5. ZELEN M. Strategy and alternate randomized designs in cancer clinical trials. Cancer Treat Rep 66:1095–1100, 1982.
6. POCOCK SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis 29:175–188, 1976.
7. GEHAN EA, and FREIREICH EJ. Non-randomized controls in cancer clinical trials. N Engl J Med 290:198–203, 1974.
8. BYAR DP. Why data bases should not replace randomized clinical trials. Biometrics 36:337–342, 1980.
9. POCOCK SJ. Letter: Randomised clinical trials. Br Med J 1:1661, 1977.
10. PETO R. Clinical trial methodology. Biomedicine 28 (Suppl):24–36, 1978.
11. FOST N. Consent as a barrier to research. N Engl J Med 300:1272–1273, 1979.
12. FLEMING TR, GREEN SJ, and HARRINGTON DP. Performing serial testing of treatment effects. *In* Proceedings of the Heidelberg Conference on Early Breast Cancer, Heidelberg, W Germany, Dec 14–17, 1981. In press.