RANDOMIZATION, STRATIFICATION AND OTHER STUFF

by

Seymour Geisser*

University of Minnesota

Technical Report No. 374

May, 1980

* This work was supported in part by NIH Grant 25271.

RANDOMIZATION, STRATIFICATION AND OTHER STUFF*

Seymour Geisser

University of Minnesota

Summary

The principles for comparative experiments, Randomization, Stratification and Replication are critically reviewed in the light of a new view put forth by a prominent group of statisticians. They propose that trials should be large: randomization with regard to therapeutic agents be balanced; prior stratification on prognostic indicators be abandoned.

It is concluded that if this advice is followed, clinicians will obtain interpretable results using fairly simple techniques, that can be convincing to fellow workers. But it is advised in this paper that important deviations are in order when a competent statistician is collaborating on the trial. In particular, a more intensive modeling of the experiment should be attempted; pre-randomization stratification on key prognostic indicators be employed; P-value analyses be abandoned in favor of Bayes-factor analyses; a predictivistic approach be undertaken.

* Invited address to the 16th Hoechst Medical Symposium on Analgesic, Anti-Rheumatic and Anti-Inflammatory Drugs held at Santa Barbara, California, June 2-4, 1980.

RANDOMIZATION, STRATIFICATION AND OTHER STUFF

Seymour Geisser*

University of Minnesota

Introduction

One of the difficulties in making inferences with respect to certain "measurable" clinical responses in highly complex biological organisms such as rats or man is the inherent variation attendent upon the administration of almost any agent including those that have no claim to being anything but totally inactive. This has led to several principles useful in conducting experiments that attempt to sort out a muted signal in the presence of a good deal of, if not overwhelming, noise.

The principles, proposed by Fisher (1926) originally for comparative agricultural experiments, are Randomization, Stratification, and Replication. Stratifying variable units into more or less homogenous blocks or strata obviously is an error reducing device. Randomly allocating the treatments to the units within strata ensures, on the average, a fair comparison of the treatments with the precision of the comparison increasing with the number of replications. This is traditionally the statistical basis for experimental design as enunciated by Fisher-- all else is commentary.

A Newly Received State of the Art or Advice to the Uniniated?

A recent series of papers (Peto plus 9, 1976, 1977) should be required reading for all clinical investigators (and statisticians) even if only to take issue with certain suggestions contained therein. This product of a collaboration of a British-American Statistici Decemviri promotes

* This work was supported in part by NIH Grant 25271.

guidelines for the conduct of clinical trials based on certain statistical considerations. Amidst a plethora of sensible and illuminating advice, three major points emerge that are pertinent to the topic under discussion here: trials should be large; randomization with regard to therapeutic agents should be balanced; prior stratification on prognostic favors should not be used. The order in which these points are made is especially relevant.

They argue that for a given number of patients and stipulated significance level the increase in power of a few large trials over many small ones, not only permits the detection of smaller differences, but also greatly restricts the number of misleading trials. They then conclude that at the least if a single large trial cannot be sustained for whatever reason, then it would be preferable that any series of small independently organized trials to assess similar questions be competently organized into a large collaborative trial.

Another consequence of large trials, they claim, is the negligible increase in expected power if pre-randomization stratification were employed, as contrasted to its use in a small trial. The necessary logistical effort in achieving an appropriate stratification may also deter investigators from active cooperation in a trial. At any rate, patients may later be subdivided retrospectively into several relevant prognostic factors, so that sharp comparisons may be made within strata. Although prior stratification can automatically achieve the balance between the numbers assigned to each treatment, which for a small trial may be critical, a randomized large trial will on the average ensure that a reasonable balance is also attained. The expected increase in efficiency in attempting a pre-randomization stratification with complete balance is slight when compared to the potential risk attendent on losing a large

fraction of available patients because of lack of cooperation in implementing a more complex trial.

The utility of stratification, the assignment of patients into mutually exclusive reasonably homogenous classes defined by prognostic factors which are presumed to influence the response, is the enhancement of the sensitivity of the comparison between treatments. It may also serve to identify differential effects of the therapies with regard to prognostic factors. But they argue that unless both the trial and the number of strata are small there is no discernible advantage to a pre-entry stratification as opposed to a retrospective stratification analysis. One note of caution -- this advice was proclaimed as intended for clinicians without statistical expertise who presumably design and analyze their own trials. It must be viewed as similar to advice a surgeon would give to a statistician who intended to lance his own boil, "If you use such and such a simple technique and take the following simple precautions, the prognosis should be favorable." Of course if the surgeon himself were to operate, he would ordinarily use much more sensitive tools and techniques depending on a variety of circumstances. The thrust of the counsel is clearly to minimize potential damage while still providing relief.

In a similar vein, the advice given by the Statistici Decemviri is -- if you conduct a large trial, and if you use balanced randomization with respect to the treatments, you will have an excellent chance of obtaining interpretable results using fairly simple statistical techniques. From this point of view, it is conservative but surely sound advice. Some statisticians, Brown (1978a, 1978b), Pocock (1979), Simon (1979), have tended to look upon the Peto plus 9 paper as one advising them how to advise clinicians

with whom they were collaborating on a trial and as such have reacted, not without some irritation, by stressing the importance of stratifying on a few potentially important prognostic variables and adaptive balancing schemes.

Large Trials

All other things being equal, everyone would agree that a large trial is not only at least as informative as a small one, assuming both were conducted in the same manner, but more than likely much more so. All things, however, are hardly ever equal -- there are costs in time, money, and effort relative to the information to be ascertained which often must be taken into consideration in determining an appropriate size for an anticipated clinical trial. It is also of no small matter that the Peto plus 9 opus was primarily directed towards trials in Cancer, Heart Disease, Transplant Rejection, Thrombosis and other extremely serious diseases which study time to death or some other untoward event, clearly attributable to the disease.

In less life threatening diseases one must more carefully consider adverse effects of a therapy which may be more threatening than the disease. The agent itself may promote an unwanted and possibly irreversible state in some fraction of the population at risk. Detection of and protection against such an eventuality is of considerable importance. The smaller the proportion so affected the larger a sample size necessary for its detection. Thus in trials of this kind, aside from the usual assessment of the benefit of the therapy as directed towards the disease, there is the potential for the emergence of far more serious consequences, which may virtually go undetected if only a small fraction are at risk, unless a reasonably large sample size is employed.

Randomization

Why randomize? Seemingly cogent arguments are advanced. The estimates of the critical parameters are unbiased and impartial with regard to selection factors. "Valid" estimates of error are obtained with which to gauge the estimation of the critical parameters. The physical act of randomization induces a distribution which can be used to assess the significance of the result. A dividend so large from so small an investment requires an explanation.

The usual schema for the statistical paradigm is a random sample from some specified population at large with some measured attribute often assumed approximately normally distributed. The basis for tests of significance and confidence intervals for the characteristics of the population under scrutiny rests on these assumptions. In this setup the physical act of randomly allocating the treatments to the units of the random sample serves merely to ensure an impartial trial in that a valid comparison can be made of the agents, unencumbered by conscious or unconscious selective bias. Although a haphazard allocation would theoretically serve just as well, it would not engender the degree of acceptability by others nor psychologically protect the experimenter against self-deception. In clinical trials, however, this model is often an impossible scenario. It is generally very difficult if not impossible to draw a random sample from a well defined population that entirely encompasses the target of the therapy. And even if it were, the attribute may not be approximately normally distributed. In the absence of satisfying this model, randomization now plays a further crucial, if not controversial, role. It permits, for moderate to large sample sizes, the use of normal theory in calculating approximate

significance levels irrespective of the original underlying distribution. Further, in small samples exact significance levels may be determined without great difficulty by a permutation computation. This amply demonstrates the potency of the physical act of randomization -- according to its proponents.

Even this powerful tool has limitations -- the statistical inference that results from the randomization test, whether calculated approximately or exactly is restricted to the individuals in the sample under hypothetically identical reruns of the trial. Only random allocation of the agents is permitted to vary while other factors are fixed at the levels assumed when the original trial was run. Such an inference is regarded by some as being of little or no import, Basu (1980). Advocates of randomization must now make an extra-statistical "logical" argument, by analogy for example, rather than a statistical inference to an interesting target population. Certainly, conclusions based on such reasoning do have a validity of their own in most instances. Indeed, there is no reason to believe that aspirin will differentially effect 50-year-old arthritic females in 1980 as opposed to 50-year-old arthritic females in 1985 under similar circumstances. Nevertheless, awareness of the various types of arguments being marshalled in support of a conclusion by those totally committed to this use of randomization is illuminating. Sooner or later even they must switch into a subjective mode.

Stratification

Randomization ensures that two agents that are balanced are equally administered, on the average, to the various categories of patients who enter the study. Clearly if the number of strata are more than modest, the on average guarantee of balance within each stratum may be illusory because of the great variability over sets of such trials. So much so that being near the average in any single trial can be highly unlikely. For a fixed number of patients the probability of achieving balance or near balance in every stratum declines as the number

of strata increases. Mitigating this is the fact that, for a fixed number of strata, the possibility of near balance increases with the size of the trial. But nobody likes to be caught with his balance down. For large studies and few strata imbalance is unlikely. But it will occur with prescribed frequency, depending on the size of the trial and the number of strata, no matter how surprising it is that it occurred to us. And when it does it can inculcate a bias, as well as a loss of power for a particular trial -- though not on the average. Small comfort indeed is an unattained average.

The possibility that an agent will exhibit specificity for particular prognostic factors, is always a possibility and sometimes a welcome one. More generally the differential response of the agents with respect to prognostic factors may vary considerably and could remain undetected or unaccounted for, unless balance is achieved within and even between strata. Such is the conventional wisdom which stresses the importance of prior stratification.

The Statistici Decemviri argue that the larger the number of strata and the closer the balance within and between strata, the larger the number of patients it is necessary to secure. As a consequence many patients may have to be turned away, especially if equal numbers are required for all strata. Quite often gaps may be filled only with great difficulty, or not at all. Indeed, the trial may become so involved and so expensive that it is useless to even attempt. Further, its analysis, which would be more complex because of the stratification, could be misleading unless this design feature is carefully taken into account. Again it is not very likely in a large trial that a particularly bad balance of agents would result from not stratifying prior to randomization for a few critical prognostic factors. Patients are easier to obtain, the trial is simpler and prognostic factors may be adjusted for after the trial. However, the latter proposal -- an analysis

based on retrospective stratification, is not held in as high repute and is thus less convincing than a prospective stratification. The controversy involves the potential incompatibility of the principles of conditionality and repetitive sampling, as well as the ambiguity insinuated into the interpretation of significance levels and power by posterior data ransacking. A completely consistent resolution of these philosophical issues, within the confines of classical sampling theory, may be impossible to achieve. Because a crucial objective of a trial is to obtain a clear cut evaluation which is persuasive not only to the conductors of the trial but to fellow scientific workers and clinicians, prior stratification is certainly the safer course, if a trial is not large.

McHugh (1980) has attempted to sort out estimators and their variances for pre-and post-stratification designs. He regards the virtue of his effort as enabling one to plan an appropriate prior stratification strategy on the basis of the anticipated relative precision of the estimators. An unsettling feature of such analyses is that if the same configuration of patients, treatments, etc. is achieved by a balanced restricted randomization implicit in prior stratification, as obtained merely by an overall balanced randomization then the analyses would differ, for identical data. McHugh also demonstrates that as the sample size increases, the analyses would tend to be the same, lending credence to large post-stratification. It is my impression that at least some of the Statistici Decemviri, though frequentists all, were disturbed or disagreed sufficiently about such possibilities that they neatly finessed the issue by advocating large trials, thus ensuring that discrepancies due to such anomalies would negligibly, or not at all, influence any important conclusions for the anticipated few retrospective strata that would require scrutiny.

Further Comments And Other Stuff

Much has been claimed for the physical act of randomization e.g. Kempthorne (1977) -- it deals with the inability to obtain a random sample from a target population -- it dispenses with restrictive distributional assumptions -- it avoids or negates selection bias -- it provides valid estimates of error -- it enables the generation of informative significance levels. Others allege these benefits are either illusory, Harville (1975) or illogical, Basu (1980) -- the former arguing in the frequency mode and the latter from a Bayesian viewpoint. Harville proposes that considerably more modeling of the setup either from a design standpoint or from a knowledge of covariates is the appropriate avenue. In a frontal assault, Basu argues that randomization glosses over much information that one neglects at one's peril while other Bayesians such as Jeffreys (1939) insist that sample space notions upon which randomization rests are not compelling. Indeed, why reject a hypothesis because it has not predicted observable results that have not occurred? This is better known as the "tail area syndrome." There is also that well known conundrum that the more control (in terms of information brought to bear on the issue) a scientist exerts on his experiment, the less scope for randomization. This results in a proportionate diminution of the number of elements in the sample space which has the direct consequence (in extremis) of rendering a test incapable of rejecting a false hypothesis at an appropriate significance level. Sample space theorists (frequentists) regard this as irremediable, and counsel limits to the control of an experiment, while Basu and other Bayesians regard such a view as unscientific and illogical. Contrarily, frequentists consider the Bayesian view as unrealistic and held by individuals who either do not conduct experiments or conduct them only to convince themselves. The

personalistic outlook is not reasonable, they maintain, for those who conduct "public" experiments; i.e. to instill conviction or confidence amongst peer groups in the conclusions drawn from an experiment.

A more enlightened, if not regal, Bayesian view espoused by Cornfield (1976) also denies that the sample space induced by randomization is cogent for inference and decision, but perceives randomization as one of several possible ad hoc devices to achieve comparability between treatment groups which come under the heading of "rexing" -- rendering prior distributions exchangeable.

What then is the empirical evidence concerning the value of randomization? Mosteller (1977) infers support indirectly for randomized trials from some 53 non-randomized clinical trials, controlled to a greater or lesser extent, of the portacaval shunt operation. He correlates the degree of enthusiasm for the operation after the study with the degree of control exercised in the trial. The data clearly indicate that the less well controlled the study the greater is the enthusiasm vested in the operation by the investigators. Collateral evidence of this kind, though far from compelling, is mildly persuasive of the merits of conducting a "fair" trial.

The Statistici Decemviri gave "robust" advice to clinicians endeavoring to design and analyze their own clinical trial on patients with a life threatening disease. The first point, and all else essentially flowed from it, was to conduct a large trial. But anyone who has the resources to embark on such a grand enterprise can certainly afford the counsel of a competent statistician on the design and analysis aspects. Hence the guidance provided is to a virtually vacuous class of clinicians. Further an informed prior stratification on a reasonable number of

prognostic indicators can be of great value in sharpening a comparison, when appropriately accounted for in the analysis.

Apprehensive that an analysis made by clinicians would not appropriately take into account design features that had been built into a trial the Statistici Decemviri are persuaded to recommend only the simplest type of design. Or it was an issue (analysis dependent on the design) in which total agreement among them was not possible. Thus, they may have preferred to avoid addressing this matter in detail. They only intimate that if stratified entry were disregarded and a retrospective stratification undertaken, calculated P-values would be conservative.

Another fact, not always of subordinate import, is that sometimes interim analyses are of value in concluding early that some agent is particularly effective, ineffective, or even deleterious. Lack of prior stratification and continual sequential balance could impede such analyses.

In the best of all possible scientific or "clinical" enterprises one brings to bear every relevant bit of information available in the construction of a probability model appropriate to the experiment. Output from the trial is transformed into a probability distribution for the response of the patients to the agents. Combining this output with a utility function (the utility of making a particular decision when a specified hypothesis is true) results in a decision as to which agent is indicated for a particular group of patients. This is the subjective Bayesian decision theoretic approach. However, to fully execute it is often such a complex and difficult affair that it boggles the mind. For example, aside from the actual modeling difficulties, whose utility is to be employed? Is it the patient's, the physician's or society's? These are often at odds.

In the absence of perfection, one must inevitably compromise. Hence, the advice of the Statistici Decemviri is sensible as an initial posture but significant deviations are in order, if guided by a competent statistician. The use of prior stratification has already been indicated as one variation.

A more radical departure is to refrain from drawing conclusions only on the basis of P-values (significance levels) which are antiquated assessments for the comparison of agents. P-value analysis should be replaced by the Bayes-factor, Good (1967), which is essentially a ratio of predictive densities, Geisser (1971), also termed relative betting odds by Cornfield (1972). When the Bayes-factor is multiplied by the prior odds, posterior odds are obtained that one agent is more effective than another. An elaborate Bayesian edifice need not be constructed to achieve such a result. Methods are available which minimize the influence of prior subjective information that is difficult to model or is vague. The time has also come to couch inferences in terms of probabilistic predictions; e.g. as a chance that one agent will enhance some measurable response in a patient or group of patients by a given number of units with respect to another agent, Geisser (1971).

To implement such analyses it is necessary to secure the assistance of a Bayesian oriented statistician with a predivistic outlook.

References

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. J. Am. Stat. Assoc. (to be published).

Brown, B.W., Jr. (1978a). Statistical controversies in the design of clinical trial. Stanford University Tech. Report 37.

Brown, B.W., Jr. (1978b). Designing for cancer clinical trials. Selection of prognostic factors. Stanford University Tech. Report 38.

Cornfield, J. (1972). The Coronary Drug Project Research Group. The coronary drug project: Findings leading to further modifications of its protocol with respect to dextrothyroxine. JAMA, 220, pp. 996-1008.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. Am. J. of Epidemiology 104, 408-421.

Fisher, R.A. (1926). The arrangement of field experiments. Journal of Ministry of agriculture, 33, pp. 503-513.

Geisser, S. (1971). The inferential use of predictive distributions. Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, pp. 456-469.

Good, I.J. (1967). A Bayesian significance test for multinominal distributions. J. Roy. Stat. Soc B. 29, pp. 399-431.

Harville, D. A. (1975). Experimental randomization: Who needs it? Am. Statist. 29, pp. 27-31.

Jeffreys, H. (1933). Probability, statistics and the theory of errors, Proceedings of the Royal Society A, 140, pp. 523-535.

Kempthorne, O. (1977). Why randomize? Journal of Statistical Panning and Inference 1, pp. 1-25.

McHugh, R.B. (1980). Post-stratification in the randomized clinical trial. Unpublished manuscript.

Mosteller, F. (1977). Experimentation and Innovations. Bulletin of Int. Stat. Inst. pp. 559-572.

Peto, R., M.C. Pike, P. Armitage, N.E. Breslow, D.R. Cox, S.V. Howard, N. Mantel, K. McPherson, J. Peto and P.G. Smith (1976, 1977). Design and analysis of randomized clinical trials requiring prolonger observation of each patient. I. Introduction and design. Br. J. Cancer 34, pp. 585-612 and II. Analysis and Examples. Br. J. Cancer 35, pp. 1-39.

Pocock, S.J. (1979). Allocation of patients to treatment with clinical trials. Biometrics 35, pp. 183-197.

Simon, R. (1979). Restricted randomization designs in clinical trials. Biometrics 35, pp. 503-512.