Randomized clinical trials in the evaluation of surgical innovation

STEPHEN J. HAINES, M.D.
Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania

Randomized clinical trials are widely accepted as the standard for evaluation of therapeutic innovation in many fields of medicine. The three basic components of such trials (concurrent comparison, random allocation, and objective observation) are designed to control four forms of bias (chronology bias, susceptibility bias, compliance bias, and observation bias) that may interfere with the interpretation of the results of a study. Only 2% of the articles evaluating therapeutic maneuvers published in the Journal of Neurosurgery have attempted to use concurrent controls. Only one of 863 such articles met the criteria for a randomized clinical trial. Reasons for underutilization of such trials in neurosurgery are discussed and suggestions for their wider use are offered.

KEY WORDS research design • statistics • neurosurgery • clinical trial • randomization

The evaluation and eventual rejection of gastric freezing\(^1\) and internal mammary artery ligation\(^2\) as viable therapeutic techniques are clear examples of the hazards of inadequate early evaluation of new surgical procedures. Similar experiences in other areas of medicine have made the double-blind, randomized clinical trial the standard for evaluation of new medical therapies, especially in clinical pharmacology.\(^8,14,20,24,92\) In a provocative paper by Deniston and Rosenthal,\(^15\) non-randomized designs have been compared with randomized designs using identical data, and shown to produce invalid estimates of the outcomes under study. Widespread acceptance of randomized clinical trials has not been apparent in surgical disciplines.\(^9\) It is the purpose of this paper to discuss the reasons for the evolution of this technique of evaluation, the current extent of its use in neurological surgery, and its implications for future progress in neurosurgery. Throughout the paper the focus is on clinical evaluation of new diagnostic and therapeutic procedures and these comments may not necessarily apply to other forms of research.

The Randomized Clinical Trial

Feinstein\(^21\) describes three basic criteria for satisfactory clinical trials: 1) concurrent comparison, 2) random allocation, and 3) objective observation. Concurrent comparison implies that the group receiving the therapy under evaluation (the treatment group) will be compared to a control group selected by the same criteria from the same population, and treated identically save for the therapy being evaluated. Random allocation means that the allocation to treatment or control group is made by a truly random, not haphazard or systematic, process. By objective observation we mean that neither the preconceived notions of the investigator-observer nor the well intentioned helpfulness of the patient-subject is allowed to influence the outcome of the trial. The study design which most closely approximates these ideal criteria is the double-blind randomized trial.

The Development of Clinical Trials

Some of the earliest controlled trials have been documented by Bull.\(^6\) These include Lind's 1747 trial of a number of treatments for scurvy, which demonstrated the effectiveness of oranges and lemons, and an 1801 trial of a device called "Perkins' tractor." This was a metallic rod which was supposed to attract electricity to cure its holder of many diseases. When identically designed wooden rods were substituted, however, the results were equally spectacular.

There has been a gradual recognition that various forms of bias may distort the results of a study in ways that are not initially obvious. Modifying Feinstein's
groupings,\textsuperscript{18,19} we may classify bias as: 1) chronology bias, 2) susceptibility bias, 3) compliance bias, and 4) observation bias. In the following paragraphs definitions and examples drawn from both neurosurgery and other medical disciplines will be given to clarify these concepts.

**Chronology Bias.** There are many subtle forms of chronology bias,\textsuperscript{19} but the most flagrant is the use of historical controls. It is clear that diagnostic and therapeutic ability may change over time. In addition, supportive therapy not directly related to the treatment under study may improve, thus influencing the outcome. It clearly would be unfair to compare operative mortalities of resection of craniopharyngioma before and after the introduction of steroids as adjunctive therapy. Such comparisons are frequently made, however, but in an informal way which makes them difficult to pinpoint as examples. The case for historical controls has been forcefully stated by Gehan and Freireich,\textsuperscript{22} and Pocock.\textsuperscript{58} The example which follows, however, meets their most stringent criteria for historical controls, and demonstrates the danger of the technique.

Byar, \textit{et al.},\textsuperscript{8} cited the Veterans Administration Co-operative Urological Research Group study on prostatic carcinoma.\textsuperscript{47} Here, patients admitted in the last third of the study and treated with estrogen showed no difference in survival between placebo and estrogen treatment. A steady improvement in both groups led to the spuriously significant result. Here historical comparison over a relatively short time period (7 years) in a homogeneous population treated by the same investigators with the same techniques would have been most misleading.

Rees\textsuperscript{97} described the procedure of multiple bilateral percutaneous rhizolysis for the treatment of low-back pain. His clinical trial consisted of 5000 consecutive patients in whom he implied a high rate of success in pain relief. He implied that this success rate was superior to that of more standard procedures. Controlled studies by King,\textsuperscript{26} however, not only demonstrated facet denervation to have no greater effect than placebo, but demolished the anatomic basis for the procedure as performed by Rees. Many patients could have been spared this procedure had a well designed clinical trial been carried out shortly after its introduction.

**Susceptibility Bias.** For various reasons, the treatment group and the control group may differ in the likelihood that they will have the expected outcome even when the effect of the new therapy is eliminated. This might come about through a non-random selection process, such as a control group consisting of patients believed to be too ill to withstand surgery, or through the operation of an unrecognized cofactor (such as smoking) that, if not equally allocated among the study groups, might bias the results for a number of important medical outcomes.

An excellent example from early in this century that demonstrates some of the subtlety with which this form of bias may operate is Berkson's analysis\textsuperscript{4} of Pearl's study,\textsuperscript{84} which seemed to indicate that tuberculosis had a protective effect against cancer. In the latter study, 816 patients with malignant tumors confirmed at autopsy were matched for age, sex, and race with 816 autopsied patients without malignancies. Active tuberculosis was found in 6.6% of those with, and 16.3% of those without malignancy. It was concluded that tuberculosis was antagonistic to cancer. An equally plausible interpretation, however, which Pearl himself pointed out shortly after the publication of his paper,\textsuperscript{38} was that the malignancy caused the death of the patients before they contracted active tuberculosis. Thus, the lack of malignancy made the control group more susceptible to active tuberculosis. Berkson\textsuperscript{4} noted that an even more subtle mechanism of differential susceptibility for selection into a study may produce similar spurious associations, and the reader is referred to his article for details.

A neurosurgical example of susceptibility bias may be found in the article by Ramsey and Brand\textsuperscript{90} on radiation therapy for glioblastoma. They conducted a randomized study of whole-brain versus limited-volume radiotherapy, and compared the results to an uncontrolled series of patients treated in the same way at the discretion of the therapist. In the uncontrolled group there was a clear preference for limited-volume therapy. Those patients had a mean survival of 20.4 months compared to 8.2 months with whole-brain irradiation. In the controlled study, however, the difference was much smaller (11.5 versus 8.5 months). Although significance tests were not reported for these differences, it is clear that uncontrolled factors markedly altered the susceptibility of the groups to recurrence and death. Had the uncontrolled study been the only one reported, the conclusions would have been quite different.

**Compliance Bias.** The treatment and control groups must be treated in such a way that they differ only in the application of the therapy being evaluated, or systematic bias may be introduced that invalidates the entire study. For example, trials of surgical procedures are liable to bias according to the skill of the surgeons involved. In addition, the patients must be equally able to comply with the treatment regimen. A group receiving monthly injections at a clinic may be more likely to fully comply with the therapeutic plan than a group required to take four pills every day without professional supervision.

An early example of compliance bias is well documented by "Student" in his article "The Lanarkshire milk experiment."\textsuperscript{44} A large-scale experiment involving 10,000 children in 67 schools was designed to assess the effect of milk supplements given at school on growth as measured by height and weight.
Allocation to the treatment and control groups was left to the head teacher at each school. The head teachers were also given the prerogative of adjusting the groups if they felt them to be unbalanced with regard to nutritional status. No specified plan for selecting the groups was made, and many different methods of allocation were used. The end result was that the group given milk had a much larger number of poorly nourished children than the control group. This initial bias obscured the subsequent effects on growth and made the results of the study equivocal.

A medical example of compliance bias is cited by Truelove. In a study of anticoagulant therapy in myocardial infarction, the choice of anticoagulant or no anticoagulant therapy was made by the day of the week. Patients admitted on odd days received anticoagulants. Those admitted on even days did not. This scheme was soon evident to referring physicians and families, and pressure to admit to the treatment group apparently resulted in significantly more patients (544) being treated with anticoagulants than were treated by control methods (442). This compliance bias makes the results of the study difficult to interpret.

Observation Bias. Incorrect observations may bias the results of any study. Poorly designed or malfunctioning equipment may introduce systematic error. Subjective impressions of either observer or patient may influence the reporting of results. A more subtle effect may come about if one of the experimental groups is more likely to have certain outcomes detected than the other, as might be the case if the treatment group were examined more frequently than the control group. A historical example of observation bias is Evans and Hoyle’s study of the value of drug therapy for the prophylaxis of angina pectoris. They studied 14 drugs which had been reported to be effective for angina prophylaxis. None had been subjected to a controlled study. Claims for these drugs included “excellent effects ... in all cases,” and 12 of 18 patients “remained free from attacks for an observation period in no case less than three months.” Ten of the 14 drugs were shown to have no greater effect than placebo. The authors emphasized that the “power of suggestion” played a large role in the efficacy of these drugs.

For a neurosurgical example we turn to a study by Sharrard, et al., on the early closure of myelomeningocele. The authors began a randomized but not blinded trial of early versus delayed closure of myelomeningocele. Although they recognized the difficulty of differentiating voluntary from reflex lower extremity motion, they thought that they could adequately distinguish the two. Early in the study they had the impression that they had attained such significant improvement in neurological function in the early closure group that they could not continue the study. They believed that at least one child would eventually walk as a result of early closure. Their follow-up study appeared to confirm these impressions. Other authors could not reproduce these results. Brocklehurst, et al., found no improvement in 25 early closures. Duckworth and Brown found a few patients who showed transient increased lower extremity activity, but none in whom this persisted long enough to be functionally useful. These authors re-emphasize the difficulty in distinguishing between voluntary and reflex activity in the neonate. Finally, Smyth, et al., conducted a randomized trial of early versus delayed closure. This time the pre- and postoperative functional assessments were made by independent observers who were kept ignorant of the timing of closure. No differences in functional improvement were found between the two groups. It seems most reasonable to relate the difference between the two studies to observer bias on the part of the original investigators who carried out functional evaluations with knowledge of the treatment and preconceived notions regarding its value.

Control of Bias

The three basic components of clinical trials are designed to control these forms of bias (Table 1). The concurrent control group reduces chronology bias. Random allocation (sometimes aided by stratification techniques) usually roughly equalizes susceptibility among the experimental groups and helps to eliminate the kind of compliance bias described above. In addition, it has certain advantages in meeting the underlying assumptions of common statistical tests thereby allowing estimation of the error inherent in the data. Objective observation, using the double-blind technique when possible, helps to control bias.

Concurrent controls, random allocation, and objective observation are necessary but not sufficient criteria for a well designed clinical trial. There is no substitute for a carefully researched and thought-out study plan. Indeed, this is more important in the control of compliance bias than is randomization. However, even the most carefully considered and scholarly study may be undermined by controllable bias if these criteria are ignored.
These articles constituted 18.4% of the papers diagnostic or therapeutic techniques are reported here. Only articles that attempted to assess the methodology in the text, or the entire text, was considered. Where necessary, the section on methodology was classified according to their methodology. In most cases the classification could be made from the title and abstract. Where necessary, the section on methodology in the text, or the entire text, was consulted. Only articles that attempted to assess diagnostic or therapeutic techniques are reported here. These articles constituted 18.4% of the papers analyzed. The remainder consists of case reports, literature reviews, animal experiments, physiological and pathological observations, and technical reports.

Papers dealing with the evaluation of diagnostic and therapeutic techniques were put into one of three categories. "Clinical reviews" consist of retrospective summaries of the authors' experiences with a designated disease entity, for example, "Indications for nonoperative treatment of spinal cord compression due to breast cancer" by Cobb, et al. Articles attempting to evaluate a specific technique were called "clinical trials." If any attempt was made to use concurrent controls, the study was classified as a "controlled clinical trial." For example, "Control of shunt infection. Report of 150 consecutive cases" by Venes is an uncontrolled clinical trial while "Radiotherapy and CCNU in the treatment of high-grade supratentorial astrocytomas" by Weir, et al., is a controlled clinical trial.

Of the 863 articles that fell into one of these three categories, 524 (60.7%) were clinical reviews, 321 (37.2%) were uncontrolled clinical trials, and 18 (2.1%) were controlled clinical trials. Of the 18 controlled clinical trials 10 (56%) used randomization procedures, and only one of these (6%) used blinding procedures in addition. Thus, only one article fulfilled the criteria for a satisfactory randomized clinical trial.

Table 2 presents the classifications by 5-year intervals. There has been an increase in the absolute number of controlled clinical trials in the last 5 years, but they continue to represent less than 5% of the articles. It appears that there is a marked underutilization of available sophisticated techniques for controlling bias in clinical neurosurgical studies, and there may be a small but encouraging trend toward increasing use of these techniques in recent years.

### Randomized Clinical Trials and Surgery

We have examined the rational basis for randomized clinical trials and seen examples of the types of error they are designed to eliminate. We have seen that this type of evaluation has become accepted in many branches of medicine as the most definitive method of testing innovations in therapy. We may reasonably ask how surgery in general, and neurosurgery in particular, measure up to this standard.

Randomized clinical trials have been successfully conducted by surgeons many times in the past. Anderson and Baden have reviewed this subject from the point of view of the gastrointestinal surgeon. Chodak and Plaut cited a number of well designed studies of antibiotic prophylaxis for various surgical procedures. The International Multicentre Trial of low-dose heparin for the prevention of postoperative pulmonary embolism is another example. The studies of McKissock, et al., although flawed in some respects, represented an early neurosurgical attempt at conducting randomized trials. On the other hand, Chalmers gives several examples of surgical procedures for which randomized randomized trials could have been performed but never have been. Spodick documents the lack of controlled studies in the surgical literature on coronary artery disease as of 1973. A closer look at the use of randomized clinical trials in neurosurgery is presented below.

The Journal of Neurosurgery was chosen as the single journal most representative of neurosurgical thought in the United States over the last 30 years. The 4685 scientific articles which appeared from 1948 (Volume 1) to 1977 (Volume 47) were analyzed. They were classified according to their methodology. In most cases the classification could be made from the title and abstract. Where necessary, the section on methodology in the text, or the entire text, was consulted. Only articles that attempted to assess diagnostic or therapeutic techniques are reported here. These articles constituted 18.4% of the papers

### Table 2

<table>
<thead>
<tr>
<th>Years of Studies</th>
<th>Clinical Reviews</th>
<th>Uncontrolled Clinical Trials</th>
<th>Controlled Clinical Trials</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1944-1947</td>
<td>17 (47%)</td>
<td>19 (53%)</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>1948-1952</td>
<td>43 (57%)</td>
<td>29 (38%)</td>
<td>3 (4%)</td>
<td>75</td>
</tr>
<tr>
<td>1953-1957</td>
<td>53 (62%)</td>
<td>33 (38%)</td>
<td>0</td>
<td>86</td>
</tr>
<tr>
<td>1958-1962</td>
<td>89 (72%)</td>
<td>34 (28%)</td>
<td>0</td>
<td>123</td>
</tr>
<tr>
<td>1963-1967</td>
<td>98 (67%)</td>
<td>45 (31%)</td>
<td>3 (2%)</td>
<td>146</td>
</tr>
<tr>
<td>1968-1972</td>
<td>120 (66%)</td>
<td>62 (33%)</td>
<td>2 (1%)</td>
<td>190</td>
</tr>
<tr>
<td>1973-1977</td>
<td>98 (47%)</td>
<td>99 (48%)</td>
<td>10 (5%)</td>
<td>207</td>
</tr>
<tr>
<td>total</td>
<td>524 (61%)</td>
<td>321 (37%)</td>
<td>18 (2%)</td>
<td>863</td>
</tr>
</tbody>
</table>

### Obstacles to Acceptance of Randomized Clinical Trials

Controversy over the use of randomized clinical trials in surgery has been evident in the literature. Chalmers, et al., list the objections to randomized clinical trials as: 1) lack of necessity for such control procedures, 2) excessive amounts of work involved, and 3) the unethical nature of such studies. We have addressed the necessity of randomization, concurrent controls, and blinding procedures above. The argument is essentially that bias in many forms may affect a study in unpredictable ways and that these procedures are the best methods currently known to control this bias.

The amount and difficulty of the work involved in conducting a randomized clinical trial deserves some comment. To fail to use the best available techniques merely because they are difficult and time-consuming runs counter to the best traditions of medicine and surgery. The specific problem of blinding procedures in surgery has, at best, partial solutions. Clearly the
surgeon performing a procedure cannot be unaware of the nature of the operation. This problem can be circumvented by having the assessment of outcome performed by independent observers who are kept ignorant of what procedure the patients received. This has been done by Smyth, et al.41

In many cases, the operations being compared will differ only in technical details that the patients will be unable to detect. When they involve different incisions or obvious differences in postoperative care, it will be impossible to keep the patients in the dark. In this case every effort should be made not to give the patient any impression that one procedure is better than the other in order to minimize bias. However, it is often impossible to completely eliminate such bias, and this fact should be taken into account in interpreting the results of such a study. When no standard procedure will be applied to the control group, the theoretical ideal would be to carry out placebo or "sham" surgery. While this has been done,2 and the practice has been supported,8 it is most difficult to justify except in the most unusual circumstances (that is, when it is virtually without risk). One must usually be satisfied with a randomly selected control group not subjected to surgery and followed by independent observers. The possibility that bias may be introduced this way should be recognized and should not impede efforts to control the other forms of bias mentioned above.

Ethical arguments against randomized trials have centered most often on the aspect of random allocation of the therapy in question. The position against randomization has been stated succinctly in articles by Gehan and Freireich,23 Weinstein,44 and Burkhardt and Kienle.7 In addition to pointing out ways in which randomization can be assisted by techniques such as stratification, and certain circumstances in the preliminary evaluation of technique or dosage where randomization may not be necessary, they seem to conclude that in most circumstances the investigator must have some opinion as to whether the innovative therapy is more or less effective than the standard against which it is to be compared. In such a circumstance it is argued that the investigator ethically must give his patients the therapy he "knows" to be more effective and therefore randomization is no longer ethical. The crux of the matter, then, is what constitutes "knowing" that one therapy is superior to another.

In dealing with this problem, the clinical investigator faces a serious dilemma. On the one hand, he is a physician trained to "administer to an individual patient the treatment that gives him the highest probability of a successful outcome."92 Much of his medical training has involved the weighing of admitted incomplete data with his own experience to arrive at a clinical judgment (or educated guess) as to what the best course of therapy is for his patient. This process has become so ingrained that he constantly evaluates reports of new therapies as they appear in the literature, noting their inadequacies, but forced by the pressures of clinical practice to draw tentative conclusions. On the other hand, as a scientist, he realizes that the inadequacies of the usual data available for such preliminary evaluations make any firm conclusion about the efficacy of the innovation impossible. The clinician's responsibility is to make the best possible decision with the available data; the investigator's responsibility is to make the available data the best possible.

These conflicting responsibilities may be resolved if the clinical investigator carefully analyzes the quality of evidence available to him at the time of designing the trial. In the rare circumstance when the natural history of a disease is monotonously poor and the efficacy of the treatment is spectacular, randomized studies are indeed unnecessary. Such circumstances are thought to be rare.9 Otherwise, the evidence may be classified as suggestive, supportive, or confirmative. Suggestive evidence is that which raises the possibility that a new therapy may be effective. It generally appears in the form of case reports or small personal series of patients. It is the basis for formulating therapeutic hypotheses to be subjected to more formal testing. Supportive evidence is that which adds weight to a hypothesis by testing it in a variety of ways. Retrospective studies, large personal series, case matching studies, and other "quasi-experimental" designs all may be used to evaluate the hypothesis, but all are subject to the unpredictable effects of bias as discussed previously. Animal experiments, because of possible species differences, also fall into this category. Because randomized clinical trials are difficult, time-consuming, and expensive, it is appropriate to use these forms of supportive evidence to eliminate hypotheses that have little merit and to confirm those of such overwhelming value that they require no further study. The vast majority of proposals, however, will fall into the middle ground of possible but unproven value. For these hypotheses confirmative evidence must be sought. The distinguishing factor of confirmative evidence is that the study design controls for bias to the greatest extent possible and allows quantification of the uncertainty of the conclusion. For most statistical techniques in use, such quantification measures assume that randomization techniques have been used. The most commonly used statistical techniques make even more rigid assumptions.29

To return to the ethical question, assume that the clinical investigator has sufficient supportive evidence to conclude that the new therapy may be as good or better than the current standard therapy. If he is intellectually honest with himself, he will also realize that he has no precise idea of the confidence he can have in that conclusion, and his knowledge of the major errors of early evaluations of past innovations will make him uneasy about accepting this latest innovation without rigorous evaluation. He is now in the
situation of not merely being ethically permitted to conduct a clinical trial, but also being ethically compelled to do so. In order to make the best decision for his patients, he must have a higher quality of evidence. He has justification, however, for believing that he will not subject any patient in his study to a treatment markedly inferior to the current standard. The surest and most bias-free method of obtaining the required high-quality evidence is the randomized clinical trial.

It is important that the randomized trial be started as soon as sufficient data are available to justify the use of the innovative technique. To wait past this time is to subject patients to further supportive studies without clear-cut evidence supporting the use of the technique. It also risks creating a climate of opinion which supports the innovation in the absence of adequate data and may make it impossible to conduct a properly designed trial.

Throughout this discussion of ethical issues, the role of the patient in the decision-making process has been purposely ignored. Clearly, informed consent for participation in a randomized clinical trial is required. A formal discussion of informed consent is beyond the scope of this paper. A properly designed trial conducted at the appropriate time by an investigator who honestly believes that he cannot make a clear choice between the therapeutic alternatives in the study should not meet with serious objection. A better understanding of the rationale behind such studies on the part of physicians and the public is necessary if objections to randomized clinical trials are to be overcome.

Conclusions

The problems of randomized clinical trials are multiple. They are complicated, expensive, and time-consuming. Where long-term results are important, the evaluation time is lengthy. However, with current analytical techniques, the randomized clinical trial is also the best available technique for bias-controlled evaluation of new therapies, each of which carries some risk and therefore must not be arbitrarily applied to our patients. Until more elegant and sophisticated techniques are available, we are scientifically and ethically compelled to make wider use of the randomized clinical trial in the evaluation of surgical innovation.

Acknowledgments

I wish to thank Drs. Peter Jannetta, Norman Guthkelch, Leland Albright, and Rosa Lynn Pinkus for their cogent criticism of the manuscript.

References

25. Kakkar VV, Corrigan TP, Fossard DP: Prevention of fatal postoperative pulmonary embolism by low doses
Randomized trials in surgical innovation

34. Pearl R: Cancer and tuberculosis. Am J Hyg 9:97-159, 1929
42. Spodick DH: Numerators without denominators. There is no FDA for the surgeon. JAMA 232:35-36, 1975