Rational Therapy in the Neurosciences: The Role of the Randomized Trial

DAVID L. SACKETT, M.D., FRCP(C)

SUMMARY How should clinicians select the specific drug, operation, splint, exercise, or conversations that will best achieve the therapeutic goal for a given patient? This essay will examine three strategies for doing so: induction from our own, individual prior experiences; abdication to the authority of our teachers and those who write review articles; and deduction from the reports of randomized clinical trials. By means of examples from the recent past, the fallibility of the first two approaches will be illustrated. Then, strategies for critically appraising the published reports of randomized trials will be described. Finally, the reasons why the results of even the proper trials may not be accepted by frontline clinicians will be introduced.

HAVING DECIDED THAT A PATIENT'S sickness warrants treatment, clinicians must select the specific drug, operation, splint, exercise, or conversations that will best achieve the goals of therapy. How should this selection be made? As previously described1 there are three ways to do it: induction, in which our own, uncontrolled clinical experience, that of others, or the extension of current concepts of the mechanisms of disease are used to identify a treatment that seems to work or ought to work; deduction, in which a review of human experiments (such as randomized trials) is used to identify a treatment that has undergone a formal determination of its clinical value; and abdication (or even seduction), in which the treatment recommendation of a teacher, consultant, colleague, advertisement or pharmaceutical representative is simply accepted on faith.

Although the method of abdication may suffice for the rarely encountered problem, its use for common presentations can only be justified for doctors working with severely restricted resources and it will not be discussed further here. The inability of the method of induction to protect us from erroneous conclusions about efficacy is less obvious, however, and deserves discussion.

Clinicians frequently judge the efficacy of a modern treatment by comparing their current clinical experience (outcomes of patients receiving the new treatment) with their former clinical experience (outcomes of patient treated in other ways, before the new treatment was available). If the outcomes in today's patients are better than those of yesterday's, the new therapy is judged both efficacious and superior to the old. The use of “historical comparisons” in this fashion is risky, as shown in the rise and fall of the “gastric freeze.” Recognizing that gastric cooling to 5 to 10 degrees celsius sharply reduced acid secretion, a former president of the American College of Surgeons lowered the temperature of the coolant entering gastric balloons placed in peptic ulcer patients to −10 degrees celsius. The first few patients so treated experienced sharp reductions in their ulcer symptoms and exhibited x-ray healing of their ulcers. These results, when contrasted with prior experience, so impressed the innovators that they paid the ultimate tribute to the use of historic controls; in their landmark report of their initial success, they stated: "Since April 1961, no patients with duodenal ulcer referred for elective operation have been operated upon on the senior author's surgical service. This circumstance in itself bespeaks the confidence in the method by patients as well as surgeons."

The next few years saw the purchase of 2500 gastric freezing machines and a growing controversy over the efficacy of the procedure. Finally, a proper randomized trial was carried out,4 which documented that subsequent surgery for ulcer disease, gastrointestinal hemorrhage, or hospitalization for intractable pain occurred in 30 (44%) of 68 patients randomized to the sham treatment and in 35 (51%) of the 69 patients randomized to the gastric freeze. These results, published several years after the initial, promising report, sounded the death knell for the gastric freeze and it rapidly was abandoned. In the meanwhile, tens of thousands of ulcer patients had their stomachs frozen on the basis of those historical controls, paying the price for our reliance on induction as a method for selecting the best treatment.

A second inductive approach avoids this problem of “historical comparisons” by contrasting the outcomes
of contemporaneous patients who did and did not receive a particular treatment. This approach can lead us astray as well. Consider an attempt to determine the efficacy of a treatment by comparing patients who took all their medicine with those who took little or none of it. At first glance, this approach appears quite reasonable, for the two groups of patients must have the same disease and appear to be reasonably homogeneous, otherwise they would not have been prescribed the same medication.

Such an approach is exemplified in a study in which men who had survived a myocardial infarction and had been started on medication were carefully followed up for recurrent infarcts or death, and also for how well they were taking their medicine. The results were dramatic; 26% of the 882 men who took less than 80% of their medicine died, compared with only 16% of the 1,813 men who took at least 80% of their medicine. Thus, those who were highly exposed to this medication did very much better than those who were not. Even after statistical adjustment for 40 risk factors and other baseline characteristics, those who took their medicine died, compared with only 16% of the 1,813 men who took at least 80% of their medicine. This determination requires an understanding of certain elements of the scientific method, these will be explained along the way.

Until we learn that this medication was a placebo! These results come from the placebo group of a trial of various lipid-lowering drugs, and this striking relation between mortality and compliance was found both in patients on active therapy and in those on placebo. Patients who faithfully follow instructions appear destined for rosier outcomes, regardless of whether their treatment works. As a result, comparisons between compliant and noncompliant patients constitute another fallacious approach to determining whether a treatment does more good than harm.

The defect in all these inductive approaches is that they fail to sufficiently limit error. They seek convincing evidence that the treatment may be good, but fail to carry out the crucial experiments that will show not just whether the treatment appears to be good, but whether it is truly good, truly bad, or simply useless.

The crucial advantage of the deductive approach is that it provides, in the randomized control trial, the scientifically rigorous opportunity to demonstrate which therapeutic claims are justified and which are plain wrong. Such trials have proven the life-saving merits of coronary artery bypass grafting in selected patients, the ability of aspirin to reduce the risk of subsequent stroke and death in patients with threatened stroke, and the safety of immediate surgery in acute cholecystitis.

Of at least as great importance to clinicians and patients, such randomized trials have also convincingly demonstrated that other treatments are not efficacious. Relevant examples here include the failure of sulfinpyrazone to improve the clinical course of either threatened stroke or unstable angina pectoris, the uselessness of internal mammary ligation in patients with angina pectoris, and the failure of the extra-cranial-intracranial arterial bypass to reduce the risk of ischemic stroke.

This essay will describe strategies for determining the validity (closeness to the truth) and applicability (usefulness in one's own clinical practice) of the results and conclusions of randomized clinical trials. Because this determination requires an understanding of certain elements of the scientific method, these will be explained along the way.

These strategies are six in number, and will be discussed in turn:

1. Was the assignment of patients to treatments really randomized?

Every patient who entered the study should have had the same, known probability (typically 50%) of receiving one or the other of the treatments being compared; thus, assignment to one treatment or another should have been carried out by a system analogous to flipping a coin.

Can we ever be confident that a treatment is efficacious in the absence of a randomized trial? Only when traditional therapy is invariably followed by death. For example, prior to 1946 the outcome of tuberculous meningitis was invariably fatal. Then, when small amounts of streptomycin became available for use in human, a few U.S. victims treated with this new drug survived: this remarkable survival following streptomycin was documented shortly thereafter in the United Kingdom. However, such instances are few, and for most treatments for most diseases, the best means for determining the efficacy of treatment is the randomized trial.

2. Were all clinically relevant outcomes reported?

In an important randomized trial of clofibrate among men with elevated serum cholesterol, some of the outcomes of therapy appeared highly favorable. For example, serum cholesterol — a key coronary risk factor — fell by almost 10%, providing some biologic evidence for benefit. However, any claim of therapeutic benefit based on this cholesterol change is an example of the "substitution game" in which a risk factor is substituted for its associated clinical outcome, and when the investigators looked further, they discovered that total mortality actually rose by more than 20% on clofibrate therapy, a result that has profoundly affected both the use and availability of this drug. Thus, because one's judgment about the usefulness of therapy can depend, in a crucial way, on the clinical outcomes chosen for comparison, readers must be sure that all clinically relevant outcomes are reported.

Furthermore, because clinical disagreement is ubiquitous in medicine, readers should also recognize the necessity for explicit, objective criteria for the clinical outcomes of interest and their application by observers who are "blind" to whether the patient under scrutiny was in the active treatment or control group.

3. Were the study patients recognizably similar to your own?

This guide has two elements. First, the study patients must be recognizable; that is, their clinical and sociodemographic status must be described in suffi-
cient detail for readers to be able to recognize the similarity between them and their own patients. Secondly, the study patients must be similar to patients in the reader's practice. To put it another way, readers should ask themselves, "Are the patients in this study so different from my patients that I could not apply the study results in my practice?" When both recognizability and similarity are satisfied, clinical readers will be able to predict with confidence the clinical outcomes to be expected from the application of the specific therapy on specific patients in their practices.

By applying this guide in the manner described, readers also will avoid being misled by two spurious criticisms often levelled at randomized trials that have yielded unpopular negative results. The first criticism suggests that the trial results are untrue if only a small proportion of eligible patients were actually enrolled in the trial. This criticism is unfounded; as long as all randomized patients have been accounted for properly at the conclusion of the trial (see the sixth guide below), the internal validity of the trial is quite unaffected by the proportion of eligible patients who were enrolled. At issue is the generalizability (sometimes called the "external validity") of the trial result, not its internal validity. And even here, the issue is not the proportion of eligible patients that were entered, but whether the enrolled patients included sufficient numbers of patients throughout the relevant clinical spectrum of the disorder under study to permit clinically useful conclusions to be drawn all along this spectrum. This leads to the second spurious criticism.

One reason that otherwise eligible patients may not be entered into a trial is the insistence, by themselves or their attending physicians, that they must (or must not) receive the treatment under study. Moreover, this insistence may be clustered among the most severely affected patients, reducing their representation in the trial. When this fact is reported, critics of the trial result reject it because "many of the sickest patients never got into the trial." Once again, this criticism misses the point; at issue is how many of the sickest patients were in the trial, not how many were out. If the trial included enough patients with the clinical features and clinical severity at issue to generate either statistically significantly positive subgroup results, on the one hand, or statistically powerful negative subgroup results, on the other, the criticism is defeated.

4. Were both clinical and statistical significance considered?

Clinical significance here refers to the importance of a difference in clinical outcomes between treated and control patients, and is usually described in terms of the magnitude of a result. Suppose that the active treatment group in a randomized trial of aspirin in threatened stroke had 50% fewer strokes and deaths than the corresponding control group. This 50% reduction becomes clinically significant when its publication leads to changes in clinical behaviour*; thus, it is confirmed as being clinically significant when clinicians prescribe aspirin to patients with threatened stroke. By contrast, statistical significance tells us whether the conclusions the authors have drawn are likely to be true (regardless of whether or not they are clinically important).

Suppose an article reports a randomized double-blind trial comparing a new drug (drug A) with an identical appearing placebo (drug B) for the control of an important neurologic disorder. Based on their results (and excluding for the moment the possibility that drug A makes patients worse), the authors of the article will have drawn one of two conclusions: either drug A is better than drug B, or drug A is no better than drug B. They will report this conclusion in their article and will try to convince readers that it is correct.

There is, of course, a true answer to this question, and the clinical trial in this article is an attempt to get at this truth. When we compare the author's conclusions with the true state of affairs, as in table 1, some very confusing statistical concepts become clear.

If the team conducting the trial concludes that drug A is better than drug B, and in truth it really is better, as in cell w of this table, then they have drawn a correct true-positive conclusion and all is well. Similarly, they can draw a correct true-negative conclusion that drug A is no better than drug B when in truth it is not, as in cell z. Trouble arises when they draw the erroneous conclusions of cells x and y. In cell x they have drawn an erroneous false-positive conclusion that drug A is better than drug B when, in truth, it is no better. Conversely, in cell y they have drawn an erroneous false-negative conclusion that drug A is no better than drug B when, in truth, it is better.

These relationships between the conclusions drawn from a trial on the one hand and the true state of affairs on the other, can usefully be thought of as an attempt to "diagnose" the truth. The clinical trial is the diagnostic test, and we can interpret it as positive (drug A is better than drug B) or negative (drug A is no better than drug B). We hope that our conclusions are true-positives and true-negatives (cells w and z, respectively) and hope to avoid the erroneous conclusions that lead to false-positive (cell x) and false-negative (cell y) interpretations.

As you can see from table 1, two sorts of erroneous conclusions can be made. These errors have names, and at least one of them is familiar, as shown in table 2. The false-positive error of concluding that drug A is better than drug B when, in truth, it is not, is called the "type I error," and in reporting the results of a trial the authors almost always tell us the size of the risk that they have made in this false-positive error, for this is what the p value means. The p value (or alpha as it is called before the study begins) is the probability of a false-positive conclusion that drug A is better than drug B when, in truth, it isn't. Thus, the smaller the p value, the more likely we will agree with the conclusion that the experimental therapy really is better than the control or placebo therapy.

Also shown in table 2, the erroneous false-negative
TABLE 1  Comparing the Conclusions Drawn from a Clinical Trial with the True State of Affairs

<table>
<thead>
<tr>
<th>The true state of affairs</th>
<th>Drug A is better than drug B</th>
<th>Drug A is no better than drug B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusion drawn from a clinical trial</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Drug A is better than drug B</td>
<td>Correct</td>
<td>Error</td>
</tr>
<tr>
<td>Drug A is no better than drug B</td>
<td>y</td>
<td>z</td>
</tr>
</tbody>
</table>

TP = true positive; FP = false-positive; FN = false negative; TN = true negative.

From the foregoing, readers of clinical articles can derive two very useful yardsticks for interpreting the report of a clinical trial. First, if the difference is statistically significant, is it clinically significant as well? Suppose that the authors of a report have concluded that drug A is better for patients than drug B, and have calculated the p value at 0.02. This means that their risk of having erroneously concluded that drug A is better, when in truth it is not, is only 2 in 100. The question then becomes, “Is it clinically important as well?” Is the difference big enough for readers to want to put their patients through the rigors of treatment with drug A? If so, they will proceed. On the other hand, the authors may simply have put so many patients in their trial that even a clinically trivial difference would be statistically significant. In this case, readers would not use drug A. This yardstick, then, can help readers decide whether to apply the results of a “positive” trial.

The second yardstick applies to the results of a “negative” trial: if the difference is not statistically significant, was the trial big enough to show a clinically important difference if it had occurred? This goes back to table 2 where I pointed out that it is possible for authors to determine ahead of time how big their study should be.

Unfortunately, the authors of most trials that reach “negative” conclusions either could not or would not put enough patients in their trials to detect clinically significant differences. That is, the beta errors of such trials are very large and their power (or sensitivity) is...
very low. Indeed, when Freiman and her colleagues reviewed a long list of trials that had reached “negative” conclusion they found that most of them had too few patients to show risk reductions of 25% or even 50%.  

This second yardstick can be either easy or difficult to apply. It is easy to apply if the authors have reported on the power of their negative trial. If this power exceeds 80%, it is usually appropriate to accept their negative conclusion as representing a “true negative;” if the power is less than 80%, the trial may, indeed, have been too small and the conclusion may be a “false negative” one.

When the power of a negative study is not reported, the application of this yardstick becomes more difficult. Methods for doing so have been published elsewhere. 1,19

5. Is the therapeutic maneuver feasible in your practice?

The requirements here are four. First, the therapeutic maneuver has to be described in sufficient detail for readers to replicate it with precision. Who did what to whom, with what formulation and dose, administered under what circumstances, with what dose adjustments and titrations, with which searches for and responses to side effects and toxicity, for how long and with what clinical criteria for deciding that therapy should be increased, tapered, or terminated?

Second, the therapeutic maneuver must be clinically and biologically sensible. For example, the dose, route, and duration of drug therapy should be consistent with existing knowledge about pharmacokinetics and pharmacodynamics. Similarly, combinations of different treatment modalities should make clinical sense. Surgical maneuvers should also make sense in terms of current concepts of structure and function.

Third, the therapeutic maneuver has to be available. Readers must be capable of administering the trial medication or executing the trial operation properly and their patients must find it accessible, acceptable, and affordabe.

Fourth, when reading the description of the maneuver in the published report, readers should note whether the authors avoided two specific biases in its application. The first is contamination (in which control patients accidentally receive the experimental treatment, such as when control patients in a surgical trial “cross-over” and undergo the operation); the result is a spurious reduction of any real differences in clinical outcomes between the experimental and control groups if the treatment is truly efficacious. The second is cointervention (the performance of additional diagnostic or therapeutic acts on experimental but not control patients). The result of efficacious cointervention is a spurious increase in the difference in clinical outcomes observed between experimental and control groups. Once again, it should be recognized that cointervention is prevented by “blinding” both study patients and their clinicians as to who is receiving which treatment. 20

6. Were all patients who entered the study accounted for at its conclusion?

The canny reader will note how many patients entered the study (usually the numbers of experimental and control patients will be almost identical) and will tally them again at its conclusion to make certain that they correspond. For example, panel A of table 3 describes clinical outcomes in 151 patients in a randomized trial of surgical versus medical therapy for bilateral carotid stenosis. 21

Among 79 surgical and 72 medical patients “available for follow-up,” a risk reduction for continued transient ischemic attack, stroke, or death of 27% ($p = 0.02$) was reported, a difference that is both clinically and statistically significant. However, closer reading of the report reveals that 167, not 151, patients entered this study and that 16 of them were “not available for follow-up” because they had suffered a stroke or died during their initial hospitalization, and thus were excluded from the foregoing analysis. Furthermore, 15 of these 16 patients originally had been allocated to surgery; 5 of them had died and 10 had suffered strokes during or shortly after surgery. Only 1 patient randomized to medical therapy suffered such a fate. The introduction of these 16 patients into the analysis results in panel B of table 3; the risk reduction from surgery is now only 16% and it is no longer statistically significant ($p = 0.09$).

The authors of the foregoing report were careful to include outcome information on all patients who entered their trial, making the construction and interpretation of both panels of table 3 possible. However, what can the reader do when outcomes are not reported for missing subjects? One approach (admittedly conservative and therefore liable to lead to the type II error) is to arbitrarily assign a bad outcome to all missing members of the group which fared better, and good outcomes to all missing members of the group that fared worse. If this maneuver fails to cancel the statistical and clinical significance of the results, the reader can accept the study’s conclusions.

The application of these six readers’ guides will not
only help the busy clinician focus down on the key scientific requirements for determining the efficacy of medical or surgical therapy; they also can be used in tracking down the most rigorous studies on a topic. For example, the National Library of Medicine now uses random allocation as an indexing term which can be added to a computer search of the recent clinical literature, thereby restricting the output to those studies that are most likely to be valid.

What proportion of papers will satisfy the requirements both for scientific proof and for clinical applicability described in this essay? Not very many, although there is evidence that matters are improving. (Although cohort studies appear to be losing out to less powerful cross-sectional studies in general medicine journals, randomized trials of therapy are on the rise). There are only a handful of ways to do a study properly but a thousand ways to do it wrong.

Moreover, even if a study does satisfy all of these requirements, it will not settle a clinical question for all time. At best, it will contribute a small, sometimes only temporary increment to our ability to relieve suffering and promote health. Moreover, the results and conclusions of even the soundest studies may provoke sharp and continuing controversy.

The reasons for this slow progress and these disputes are several. First, there is the possibility that, despite impeccable design and analysis, the study results are flat wrong. This, of course, is the inevitable (although rare) consequence of testing for statistical significance.

Second, the contemporaneous understanding of human structure, function, and mechanisms of disease that led us to group certain sorts of patients together may be shown subsequently to have been seriously deficient, negating the results or interpretations of the original study, especially when it was judged to be negative.

Third, the study may be misunderstood or misinterpreted by those who read about it. This is especially true when an explanatory trial designed to answer the question "Can treatment X work under optimal circumstances (for example, compliant patients, elaborate dose-setting schemes, and a restricted set of clinical outcomes)?" is criticized for its inability to answer a management question, "Does treatment X do more good than harm under usual clinical circumstances (for example, all comers, usual dose-setting procedures, and the gamut of morbidity and mortality outcomes)"?

Fourth, controversy can arise over the interpretation of even a proper study when a trade-off must be made between the different sorts of results it produces. For example, studies of alternative approaches to managing patients with symptoms of appendicitis have shown that one could minimize deaths from this cause with a liberal policy of operating on all such patients, even those with mild symptoms. On the other hand, if one wanted to minimize unnecessary surgery, hospital costs, or convalescence, one would adopt a more conservative policy and reserve surgery for patients with severe symptoms. Thus, there may be not one, but two, sharply contrasting "best answers" to a clinical question and in such circumstances controversy becomes inevitable.

Fifth, study results and interpretations, even those that satisfy all the scientific requirements set down in this essay, may meet considerable resistance when they discredit the only treatment currently available for a condition. Clinicians faced with such patients still may prefer to do something, even if it is of no demonstrable benefit, rather than do nothing.

Finally, negative study results will be rejected, regardless of their scientific merit, if they threaten either the prestige of those who previously advocated the discredited treatment or the livelihood of those who profited from its execution.

References

IMPS (Intact Months of Patient Survival): An Analysis of the Results of Carotid Endarterectomy

SARAN JONAS, M.D.

SUMMARY The literature on carotid surgery for lesions appropriate to prior episodes of ischemia has been reviewed. Only one randomized study and six non-controlled reports give useful data (this despite more than thirty years of surgical activity in this field). When analyzed by the IMPS (intact months of patient survival) criterion, the randomized study failed to show benefit from surgery. This failure can be attributed to a high (35%) operative stroke and death rate. That sufficiently low operative stroke and death rates are readily achievable is not clear, however, only two of six relevant non-controlled series reported in the literature had operative stroke and death rates below the 10.4% level calculated as necessary for a "break-even" situation. Three of the six non-controlled series contain sufficient follow-up data to permit IMPS comparison against the "standard" of the control group of the randomized study. Against this "standard" only one of the three non-controlled studies would have "shown benefit" from surgery. Barnett, Plum, and Walton have called for audits of endarterectomy results in institutions in which such surgery is "standard" only one of the three non-controlled studies would have "shown benefit" from surgery. Barnett, Plum, and Walton have called for audits of endarterectomy results in institutions in which such surgery is "standard".

Reports on Carotid Endarterectomy

Warlow recently published a detailed review on endarterectomy. It was his aim to focus on the results of surgery on carotid arteries in whose territory TIA (transient ischemic attack), RIND (reversible ischemic neurologic deficit), or minor stroke had occurred. As he indicated, however, only one randomized study devoted to surgery for TIA had been reported at the time he undertook his analysis. In the report of that study, the outcome of the patients who had had carotid TIA cannot be separated from the outcome of the approximately 46% who had had verte-brobasilar TIA only.

Since Warlow submitted his paper, Shaw et al have published a controlled study (to which Warlow alluded in a footnote) of carotid surgery for TIA and minor stroke. Although this study is small (total of 41 patients), it has several virtues: all patients were randomized for surgically treatable stenoses in carotid arteries relevant to their symptoms; much detail was provided; and mean follow-up was long (greater than six years). I herein review the results of Shaw et al and of six noncontrolled studies. Relevant data from these seven reports are summarized in table 1. These seven reports are the only ones in the literature which meet the following rigorous criterion: they specify that they present results of surgery on carotid arteries in whose territory TIA, RIND, or minor stroke had occurred. (There are, of course, many other reports on surgery, but in those other reports data of the sort described above are either inextricably mixed with, or not specified as isolated from, results of surgery on patients with no symptoms, with symptoms in one vascular territory but surgery in another territory, with surgery on vertebral or subclavian arteries, with surgery after major stroke, etc.)

IMPS Analysis

Background
Warlow stated that the proper criterion for evaluating endarterectomy results is the subsequent duration...
Rational therapy in the neurosciences: the role of the randomized trial.
D L Sackett

Stroke. 1986;17:1323-1329
doi: 10.1161/01.STR.17.6.1323

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/17/6/1323

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/