Rules of scientific inquiry and clinical trials:

Improvement is needed

Arminee Kazanjian, Dr. Soc.
David C. Hadorn, M.D., M.A.
Carolyn J. Green, B.H.Sc.(P.T.), M.Sc.

British Columbia Office of Health Technology Assessment
The University of British Columbia

Sources of Support: British Columbia Ministry of Health grant through the Office of the Coordinator of Health Sciences, The University of British Columbia

Author Responsible For Correspondence:
Dr. Arminee Kazanjian, Associate Director
Centre for Health Services and Policy Research
The University of British Columbia
429-2194 Health Sciences Mall
Vancouver, British Columbia V6T 1Z3 Canada
Tel: (604) 822-4618 Fax: (604) 822-5690
Abstract

Valid scientific research on the effectiveness of interventions provides the foundation for improving the quality of health care. The social nature of this scientific activity requires a commitment on the part of scientists to follow established rules of inquiry in order to foster objectivity in observation and conclusion. The existence and general acceptance of these rules is in large part responsible for the privileged position awarded to science by contemporary Western societies.

When scientists do not follow prescribed rules, the validity of their observations, studies, and conclusions is compromised. Such rule infractions are not always acknowledged by the relevant scientific community, in which case society at large remains unaware of the problem.

We illustrate this problem by reference to the use (or non-use) of established rules designed to minimize observer bias in clinical trials and ensure unbiased reporting—areas which require subjective judgment and which are therefore modifiable. To the extent that investigators have strayed from these rules, the affected trials should be considered less than fully scientific and society should hesitate to accept their results. The identification of breaches of scientific rules of inquiry through rigorous critical appraisal protects the public from recommendations emanating from flawed science.

Keywords

Research Standards; Experimental Design; Observer Bias
Social context of scientific knowledge

It is well-accepted by philosophers and historians of science that science is a “social enterprise.” (1) Although many concepts are intended by this characterization, perhaps most fundamental is the tenet that observations, studies, and experiments gain scientific status by virtue of following *rules of inquiry* subscribed to by the relevant *community* of scientists.

As noted by Helen Longino in her book *Science as Social Knowledge*:

One does not simply declare oneself a biologist but learns the traditions, questions, mathematical and observational techniques, “the sense of what to do next,” from someone who has herself or himself been through a comparable initiation and then practiced. One ‘enters a world’ and learns how to live there. (2 p.67)

Such initiations also occur among physicians, as noted by Robert Merton:

The profession of medicine... has its own normative subculture, a body of shared and transmitted ideas, values and standards toward which members of the profession are expected to orient their behaviour. (3 p.71)

Indeed, following accepted rules and techniques is one of the hallmarks of science.

To a substantial extent the rules and standards of scientific practice act to ensure objectivity in results emanating from the various forms of scientific inquiry. (4) Unlike the products of artistic endeavors (e.g., paintings, books, poetry), the value and validity of scientific activities is not (or ought not be) “in the eye of the beholder.” Only to the extent that the tools and
procedures of inquiry can, in principle, be wielded by any suitably skilled and equipped individual, and only to the extent those tools and procedures produce the same (or similar) results when wielded in the same (or similar) settings, can the products of inquiry be considered objective—and thus scientific.

Similarly, criticisms of seemingly scientific activities must be based on accepted standards of inquiry:

In order for criticism to be relevant to a position it must appeal to something accepted by those who hold the position criticized. . . . This cannot occur at the whim of individuals but must be a function of public standards or criteria to which members of the scientific community are or feel themselves bound. These standards can include both substantive principles and epistemic, as well as social, values. . . . [I]t is the existence of standards that makes the individual members of a scientific community responsible to something besides themselves. (2 p.77)

Observation as scientific activity

The most fundamental of all scientific processes is observation. As with all scientific activities, to count as scientific, observations must be made in accordance with accepted rules and standards. As noted by Peter Kosso in his book, Reading the Book of Nature:

Science cannot accept just any old report of the senses. As a rule, observations must be carefully done and repeatable. They must be carried out in a controlled way and under the
proper conditions... Responsible observing is observing we are able to justify. This requires an understanding of the relevant conditions under which an observation is done...

... The scientific community must understand what it takes to do the observation properly if the observation is to be accepted as scientific evidence. Science, after all, is a public enterprise, and the matters of justification are matters of community. (5 p.112)

As always, the rules of observation strive to ensure the objectiveness of observations (technically, of observation reports), in the sense described above. The objectiveness of observation reports consists of two dimensions: (1) protection against bias (such as through ensuring reproducibility of observations by multiple independent observers) and (2) independence of the theoretical basis of observation from the theory tested by observation.

Regarding protection against bias, scientists, like everyone, tend to observe what they hope or expect to observe. So much is simple human nature, but the complications wrought by such biased observation (and observation reports) can be harmful. Bias need not be conscious (although this occurs often enough); indeed, the more pervasive problem is due to unconscious bias.

The existence of unconscious bias in scientific observation was perhaps first documented rigorously by Robert Rosenthal and co-workers in the late 1950’s and early 1960’s. (6 7 8) These investigators used a series of paradigms in which experimenter-subjects were led to believe, prior to testing the performance of experimental subjects on various cognitive tasks, that their subjects would perform either better or worse than average. To a substantial and disturbing effect, these expectations were found to be fulfilled, despite the absence of any actual differences in subjects’ abilities.
Subsequent to these pioneering experiments, an entire literature has accumulated documenting the biasing effect of self-fulfilling prophecy, \(^{(9 \ 10 \ 11)}\) including further research into the reasons students perform to teachers' expectations, \(^{(12 \ 13)}\) evaluation of workplace productivity, \(^{(14)}\) and the effects of clinicians' initial expectancies on diagnosis and treatment. \(^{(15)}\)

In the more exact domains of science, such as chemistry, protection against bias is achieved by means of measurement instruments (e.g., thermometers, spectroscopes) whose readings are readily reproducible and accessible to anyone. In less exact sciences, such as clinical research, protection against bias is achieved by the use of additional measurement techniques such as blinding individuals charged with making critical observations (e.g., ensuring observers of patient outcome in clinical trials are unaware of which treatment was applied).

With respect to the second tenet of objectiveness in scientific observation, it is essential that the theory upon which the credibility of a set of observations is based is independent of the theory whose truth (or empirical adequacy) is being tested by those observations.

For example, expansion of a mercury column inside a thermometer provides independent evidence in support of the theory that mixing an acid and a base produces heat because our theory linking heat to mercury expansion is distinct from our theory about acids and bases. By contrast, following Kosso, evidence obtained from a "caloric flow meter" (a machine designed to detect the flow of "caloric fluid"—formerly hypothesized to explain the phenomenon of heat) would fail the test of independence if used to prove the existence of caloric fluid:

\[T\]he problem is not that the observation is very indirect. Lots of good observational evidence for lots of good theories is very indirect. Nor is the problem that the observation
is theory-laden. All observation is theory-laden. It's not even that it is laden with a weak, unproven theory. . . . The problem is that the observation is laden with caloric theory. As evidence for caloric theory, then, it fails on evaluation of independence. It is not objective evidence....Independent evidence is objective evidence, and the requirement of independence is a key ingredient of the scientific process that prevents problematic circularity in the justification of theories. (5 p.172-173; 158)

To summarize, scientific disciplines develop relevant rules of observation according to the specific goals of those disciplines and specific threats to objectiveness. In experimental inquiry, researchers seek to enhance validity by minimizing bias and striving to separate the theoretical foundation of the observation from the theory under study.

**Bias in clinical trials**

In the realm of clinical research, the problem of observer bias is most severe when, as is often the case, observer-investigators have clear prior opinions about the effects of treatments under study. If observers already suspect that Treatment A is better than Treatment B, they are likely to observe more improvement in patients receiving Treatment A than in those receiving Treatment B. Prior beliefs about the relative effectiveness of treatments under study in a clinical trial are particularly problematic when, as is often the case, trials are used to “confirm” the benefits of treatments already in widespread use. Even for new interventions, clinical studies are usually undertaken to establish efficacy, not inefficacy. Rules of inquiry protect a trial from the introduction of bias.
Bias may be variously defined as a systematically wrong estimate of a parameter of interest, the amount by which an estimate differs from the true value, or any process tending to lead to results differing systematically from the truth. (16 p. 176)

In order to provide assurance to those outside the clinical research community - practitioners, payers policy-makers - that rules of scientific inquiry have been adhered to, a number of features of clinical trials are routinely assessed through systematic review and critical appraisal. Some of these rules are often enforced by research funding agencies and publications. For example, random allocation, a technique which ensures that in the long run treatment and control subjects are comparable in relevant ways, is considered so important for controlling selection bias that non-randomized trial designs are considered fatally flawed by many observers.

Although important, we do not address the issue of allocation bias in this paper. Rather, we focus on the problem of observer bias with respect to the observations (and observation reports) upon which inferences are drawn concerning treatment effectiveness (or lack thereof). As it turns out, rules designed to ensure unbiased complete reporting of outcomes are not as stringently enforced as those pertaining to randomized allocation. The neglect of these features is unfortunate, because both sets of rules are equally likely to render trial results invalid. Furthermore, adherence to rules of observation does not represent an undue burden on trialists, especially compared to the expense and logistical difficulties involved in conducting a randomized trial. Insisting that randomized clinical trials adhere to these rules is both appropriate and necessary if the results of the trial are to count as scientific.
Blinding of outcome assessment

To guard against observation bias, rules of inquiry for clinical research have been developed and promulgated in textbooks of epidemiology and clinical research. Specifically, it is widely held by scientists in these disciplines that individuals making observations of relevant treatment outcomes should be blinded or masked with respect to the treatment received. (17 18 19 20 21) That is, they must not know if the patient has received, for example, Treatment A or Treatment B. Schulz et al(22) found that unblinded studies reported significantly higher estimates of treatment effectiveness than blinded trials on the same topic.

Blinding helps to achieve objectivity both by reducing bias and by ensuring independence of the theories underlying observations from the theory being tested. With respect to the latter aim, as noted by Kosso, “an understanding of the causal interaction between the object and the viewer” (5 p.164) is needed to assess independence of evidence. In the case of unblinded observers, this causal interaction includes elements of the theory under test. The theory is “in” the observer, so to speak, and thus in the observation. Such “contaminated” observations do not, therefore, constitute independent evidence of the theory under test. As such, they do not meet the test of objectivity required of scientific evidence.

While this conclusion may seem overly harsh to many investigators, who will protest that many kinds of clinical trials simply “cannot” be performed blind, it is always possible to blind the assessors of key clinical outcomes. This point was made by Thomas Chalmers:

Many people do not realize that in situations where the physician caring for the patient cannot be blinded, it is still possible to blind some observers who are gathering endpoint
data. In other words, the physicians who are making the critical decisions on whether one therapy is better or worse than another can be blinded, even though those taking care of the patient may not be. (23 p.139)

Thus, it is insufficient to claim simply that blinded assessment is “too hard” or “not possible” in some clinical trials. In this same paper Chalmers noted that only 64% of a series of 300 randomized trials reviewed by him and co-workers had been conducted using properly blinded assessment of outcomes.

Blinding of data analysis and reporting

More rigorous scientific standards may also be applied to data analysis and reporting, both processes with a subjective component and therefore prone to investigator bias. Although rare, Gotzsche found grounds for blind analysis and reporting while conducting a meta-analysis of nonsteroidal, antiinflammatory drugs (NSAIDS). Recalculation of p-values in instances where this was possible revealed 12 double-blind trials with miscalculations—all errors favored the new drug over the old. Gotzsche subsequently concluded:

Blinding during data analysis and writing of manuscripts may be important and should therefore be assured by, for example, letting an independent office or agency perform the randomization and hold the randomization codes....[S]ham codes would allow the statistical analysis to be made and two manuscripts to be written, approved by all coauthors and filed with the office, before the code is broken. (25 p.289)
Blinding this final step, in the clinical trial process, fulfills conditions of objectivity at the most stringent level and ensures independence of the evidence.

**Inter-observer reliability**

Another major component of the requisite objectivity of scientific observation is the reliability of observation—that is, the extent to which different observers arrive at the same (or similar) judgments concerning the observation of interest. Reliability of measurement is a prerequisite for validity. Indeed, inter-observer reliability is a prerequisite for measurement. Only when independent observers are able to arrive at the same (or similar) determinations of the quantity involved (e.g., temperature, briskness of deep tendon reflexes) can an attribute be considered measurable. For example, the measurement of temperature in a chemistry experiment is generally considered highly reliable (and therefore valid) because multiple observers are able to take independent readings from and of several standard thermometers which are in close agreement with each other.

The observations used in clinical trials are seldom amenable to such high degrees of inter-observer agreement, of course. With the notable exception of death, in which case inter-observer variation is not much of an issue (unless the cause of death must be specified), almost all clinical outcomes are susceptible to inter-observer variation. Even “precise” tests such as X-rays and blood pressure readings are subject to substantial inter-observer variation. A large literature has accumulated on this topic. (26 27 28 29)

The lack of precision in measurement implied by large inter-observer variation has the effect of increasing random (and standard) error and producing less precise effect estimates. In
other circumstances, large inter-observer variability may compound bias introduced by unblinded observers. In trials dependent on “soft” endpoints known to have substantial inter-observer variability, such as “quality of life” or the “presence or absence of asymmetries in muscle strength or in deep tendon reflexes”, inter-observer reliability testing is especially critical. In these circumstances, when protection against bias, as afforded by blinding, is not in place; the low inter-observer variability compounds the biases inherent in pre-conceived notions of effectiveness by reducing the validity of the measurements in terms of which outcomes are assessed and reported. Agreement among independent observers ensures that the measures are replicable and enhances the validity of the clinical observation.

Because of this generic problem, clinical investigators should provide evidence concerning the inter-observer reproducibility of the outcome assessments used in RCTs, particularly in cases of unblinded trials. This is not to say that every patient must be assessed by two or more independent observers, of course, but rather that inter-observer reliability should be tested in an appropriate subset of study patients.

Reporting of results

The final rule or methodological standard to be considered here concerns the reporting of observations. Specifically, observations should be reported in a manner that provides the most accurate indication of the significance of the results. In the hard sciences this tenet is rarely problematic, but in the softer sciences, including medicine, there is often great discretion available in how observations are reported.

With respect to reporting the results of clinical trials, for example, the most commonly reported parameter by far is relative risk reduction, in which a treatment might be said, for
example, to reduce mortality by 50 percent when applied to some patient population. Proponents of relative risk note that it is useful in applying the results of a trial to populations with differing baseline risk.

As is well known, however, relative risk reduction can be grossly misleading, as this statistic does not distinguish between, say, a reduction in mortality rates from 100 percent to 50 percent (a substantial effect by anyone's lights) versus a reduction from 4 percent to 2 percent—an effect of dubious significance. For this reason, it is widely accepted that effects observed in clinical trials should be portrayed primarily in terms of absolute effect size especially when relative effect size is reported in isolation from information about absolute effect size.

A third measure of effect size has been identified, known as the Number Needed to Treat, or NNT, (= 1 / absolute effect size). (29) (For example, a treatment reducing mortality by 2 percent would need to be administered to 50 patients in order to extend one life.) Intuitively, it seems that this statistic provides an important, even critical insight into the practical significance of the observed effect size. For this reason, Laupacis et al. recommended that "the relative risk reduction should not be cited without simultaneously indicating the absolute risk reduction or the NNT." (36 p.A14)

Several studies have demonstrated that the degree of enthusiasm evinced by clinicians for the results of studies depends substantially upon the mode in which those results are presented, with NNT figures resulting in the least enthusiasm. (28 29 30 37) As noted by the investigators in one such study:

...we believe the most plausible explanation for the present results is that NNT's and similar measures dampen enthusiasm for drug therapy by offering a more clinically
meaningful view of treatment effects. . . . We predict that summary measures other than percentage reductions in relative or absolute risk will emerge as a more relevant and ethically grounded way to summarize and compare treatment effects for physicians, patients, and policy makers alike. (28 pp.919-920)

A good example of how NNT can provide a less optimistic (or less dramatic) portrayal of research results can be found in a recent meta-analysis of six RCTs on the effect of warfarin in preventing stroke in patients with chronic non-valvular atrial fibrillation. (38) Collectively, the observations from these trials implied a 3.1 percent absolute reduction in risk of stroke in treated patients (from 4.5 percent untreated to 1.4 percent treated).

The investigators concluded that their analysis “has confirmed that warfarin dramatically decreases the risk of stroke by a relative risk reduction of 68 percent.” This depiction of effect size seems quite inflated when considered from the standpoint of NNT:

. . . the 3.1 percent absolute risk reduction in all strokes claimed in the pooled analysis corresponds to an NNT of 33 (1/3.1). That is, 33 patients would need to be treated with warfarin in order to prevent one stroke. Of the remaining 32 patients, 31 would not have a stroke even if left untreated (and thus do not need warfarin) and one patient would have a stroke even if treated with warfarin (and thus would not benefit from warfarin). This perspective on the overall clinical yield of treatment is far more realistic and meaningful than the relative risk reduction figure 68 percent implies. (39 p.27)
**Review of recent RCTs**

Although rules of scientific inquiry apply to many different experimental and observational research designs, the randomized controlled trial is regarded as the gold standard and therefore is required to meet more stringent criteria for rigour. In order to determine prevailing practices with respect to the standards of observer blinding, inter-observer reliability testing, and appropriate reporting of results, we reviewed all RCTs reported in the the *Journal of the American Medical Association*, the *Lancet*, and the *British Medical Journal* in 1996. Only studies reporting randomized allocation of treatment were included. We noted whether or not patient assessment was conducted by observers blinded to treatment assignment and whether any attempt was made to assess inter-observer reliability of the assessments. We also determined whether effect sizes were conveyed predominantly in terms of NNT, absolute effect sizes or relative effect sizes.

Altogether, 118 RCTs were published by the three journals during 1996 (*JAMA* = 30; *BMJ* = 30; *Lancet* = 58). Of these, 61 (51.7 percent) were conducted using blinded patient assessment. After eliminating the 11 studies that assessed outcomes only in terms of patient-completed questionnaires, blood chemistries, or 'hard' endpoints (e.g., conception and live birth), 61 out of 107 RCTs (57.0 percent) were conducted using blinded assessment. (Note that separate testing of questionnaire reliability and chemistry calibration was generally reported.)

Eleven studies (9.3 percent) observed no effect with treatment. Of the remaining 107 studies, 83 (78 percent) reported results in terms of absolute effect size. No study reported effect sizes in terms of NNT as a primary effect measure.

Only one study reported inter-observer reliability of patient assessment. An additional three studies conducted test-retest reliability of the patient questionnaires used to assess outcome.
Discussion

The proportion of RCTs using blinded assessment in our sample (57 percent) is less than the proportion reported by Chalmers et al. in 1983 (64 percent). Although we are unable to say whether this represents a true decline over the intervening 13 years, certainly the situation does not appear to have improved since then.

As discussed above, unblinded assessments will frequently be seriously biased—especially, as is often the case, when clinicians have an opinion about which treatment is more effective. Under such circumstances, observations cannot be considered objective and the study cannot, therefore, be considered grounded in science.

We were disappointed to find that only one trial conducted (or at any rate reported) the results of inter-observer reliability testing with respect to outcome assessment. The combination of unblinded observation and absence of reliability testing is worrisome, especially in the context of ‘soft’ endpoints, such as neurological or behavioral deficits.

Regarding the method of reporting study findings, we were again disappointed to find that none of the RCTs reported effect sizes in terms of Number Needed to Treat. Nine years after the advantages of this statistic were first described, (29) clinical investigators have still not received the message. On the other hand, about 80 percent of the studies described effects sizes primarily in terms of absolute risk reduction, thus meeting the prevailing standard in this respect. Some of the remaining studies used statistical methods which virtually require reporting in terms of relative risk reduction, including Cox proportional hazards and Kaplan-Meier survival curves. In view of the discussion above, we must question the appropriateness of these methods in the context of small effect sizes where the lack of clinical significance is obscured.
If scientific inquiry is a social enterprise, we can only conclude that too many clinical scientists are behaving in ways that do not benefit society. In particular, the apparent disregard for the importance of blinding, inter-observer reliability testing, and appropriate reporting standards manifested by a substantial proportion of contemporary investigators is incommensurate with good scientific practice and contrary to the public interest.

We call on clinical investigators, funding agencies, peer-reviewed journals, and the academic community at large to ensure adherence to appropriate rules of inquiry and reporting. A greater respect for the reasons behind these rules will help to ensure the objectiveness and (therefore) scientific character of contemporary clinical research.
References


36 Laupacis A, Naylor CD, Sackett DL. How should the results of clinical trials be presented to clinicians? ACP Journal Club May/June 1992.


