

Practice Guidelines and Outcomes Research (Part II): Scientific Pitfalls

Jane M. Orient, MD

The purpose of practice guidelines and outcomes research is to optimize the outcome for the collective. As with other designs of "scientific" socialism, serious methodologic flaws are papered over with scientific terminology and quantified opinions.

A societal imperative to follow guidelines should be rejected on moral grounds alone, for it repeals and supersedes the Oath of Hippocrates, which states: "I will prescribe regimen for the good of my patients according to my ability and my judgment and never do harm to anyone."

Voluntary guidelines may sometimes be useful and instructive; however, physicians need to examine them critically. The purpose of this article is to review the basic principles of sound scientific research, which should be applied to the evaluation of outcomes research and practice guidelines.

Requirements for a Properly Designed Clinical Study

To reach valid conclusions, a scientific study requires the following:

1. A precisely defined question,
2. A defined patient population,
3. Standardized data collection,
4. Reproducible measurements,
5. Precisely defined outcomes,
6. Control groups,
7. Unbiased observers,
8. Adequate sample size,
9. Hypotheses defined *in advance* of the analysis,
10. Valid statistical methods.

It should be recognized from the outset that most practice guidelines are not derived from a scientific study but are based solely on the consensus of a select group of "experts." As extensive clinical experience has shown, the experts can be wrong. To name a few examples of protocols that have long been discarded: the Sippy diet for peptic ulcer disease; prolonged bedrest after myocardial infarction, childbirth, or major surgery; and intermittent boluses of high-dose insulin for diabetic ketoacidosis.

Practice guidelines must be validated by actual observations, a step that is usually omitted. Once is not enough; the standard of care should be subject to constant revision as knowledge advances.

Unfortunately, the purpose of many guidelines is actually to thwart scientific advances. At a forum¹ sponsored by various managed-care plans and attended by

congressional representatives, a well-respected academician said: "Research is like nuclear war. As John F. Kennedy said, 'The fruits of victory will be ashes in our mouth.' If we don't stop research, care will become more and more sophisticated."

Technology assessment and outcomes research is one part of his answer. As he pointed out, it takes 10-12 years to bring a drug to market because of FDA regulations. (He didn't mention the cost of more than \$400 million.) At present, "we don't do that with procedures." (He believes American medicine is *underregulated*.)

If the very purpose of outcomes research is anti-scientific, it is not surprising that scientific principles are often ignored in the design of its studies.

What's the Question?

For outcomes research, the question is: "What works...?" Many who advocate such research appear to believe that they are the first people in the history of the world with the insight to come up with this new question. In fact, the nonscientific nature of the inquiry should be immediately apparent from the broadness and vagueness of the question.

A scientific study must begin with a specific question that the study is capable of answering. For example: Does cimetidine speed the healing of duodenal ulcers? Does early ambulation change the mortality of myocardial infarction or the incidence of deep venous thrombosis or pulmonary embolism? Does diabetic ketoacidosis respond to continuous infusion of low-dose insulin, and what effect does this treatment have on the incidence of hypoglycemia?

The "What Works...?" question might usually be translated as: "What decreases expenditures without getting us into deep trouble?" Not coincidentally, it is generally raised in a political context that emphasizes "cost containment."^{2,3}

For example, a prominent academician and medical editor would like to have kept the Swan-Ganz catheter off the market. It was introduced in 1973, "and by 1980 we were spending \$1 billion per year on it," he said, noting that the hospital mortality and one- and five-year survival is about the same whether similar patients have the catheter placed or not.¹

What works...? This question always has an ellipsis. *What works to block new technology?* (Ham-fisted regulation? Fines or imprisonment for violating guidelines?) *What works to limit expenditures for medical treatment?* (Global budgets? Rationing? Destroying research pro-

grams? How about restoring a free-market price mechanism?)

Scientific questions are not of the "fill-in-the-blanks" style. A scientific question about Swan-Ganz catheters might be formulated thus: "Does monitoring the pulmonary wedge pressure result in shorter ICU stays, or lower mortality, or fewer episodes of congestive heart failure or renal shutdown in patients with major trauma, coronary-artery bypass, or some other specified condition?" The blanks must be filled in at the design stage. One cannot determine whether an intervention accomplishes its purpose without knowing what the purpose is.

Who are the Subjects? (Or Human Beings Are Not an Inbred Strain of Mice)

Research on human subjects is always very complex because of the large number of confounding variables. Even diseases that produce similar histologic findings (for example, specific cancers) may behave very differently in different individuals. Thus, it is crucial to have a defined population, with control and experimental groups matched as closely as possible. One must also collect detailed information about all factors that might be important.

The set of persons who had a certain code number on an insurance claim is not likely to be suitable for answering even a well-defined question, much less a "What Works...?" question. And the 34 pieces of information demanded by the Maryland Health Care Access and Cost Commission will not meet the needs of any serious researcher.

How are the Data Recorded? (Or, The Clinical Record Is Not Like a Laboratory Notebook)

As generally kept, medical records are unsuitable for scientific research. They are used in retrospective studies because they contain the only data available. Clinical data are usually collected by a large number of individuals in unstandardized format at variable intervals. The most important data

may be missing in a large proportion of cases. Sometimes, checklists are placed in the chart to help overcome this problem. But I am not aware of any studies that address the question of how much confidence should be invested in a scribble on a checklist. Does the observer carefully weigh each answer or just race through the list to meet a procedural requirement by the end of the shift?

Well-designed clinical research studies invest substantial time and money in developing the observer protocols and training personnel in their use. Each protocol is designed with a particular set of narrowly defined hypotheses in mind. It is logistically impossible to adhere to a research protocol in a busy clinic that sees all comers.

Even if all clinical personnel kept meticulous and legible records, the reliability of their data would still be in question. A stethoscope is not like a Mettler balance.

Even observations that are sanctified as revealed truth by virtue of being typed on an official report and based on a photograph from a high-technology imaging device may in reality be quite subjective.^{4,5} The interobserver variability for physical examination can be enormous. The reported sensitivities of palpation for splenic enlargement range from 28% to 100%, depending partly on the size of the spleen.⁴ In one study, five experts could agree on the presence or absence of a heart murmur in only 53% of the cases.⁶ The reliability of the clinical history is not well studied and open to serious question. And in scientific studies or surveys that use data from questionnaires, each questionnaire must be validated.

Putting the record on a computer does not solve these problems. It simply makes it easier to access a large number of unreliable observations. The error rate in current medical records is extremely high. One study of the reliability of a discharge abstract system showed that 22% of 20,260 items were incorrect. Of 1,829 records, 82% differed from the abstract in at least one respect. Sources of error included physicians (62%), coding (35%), and key-

punch (3%). The average abstract contained 2.14 physician errors and 0.81 coding errors.⁷

What Is the Outcome?

All scientific research reports outcomes ("results") in the plain English sense of the word. But "outcomes research" is indeed different. It primarily concerns "health care delivery systems," rather than specific treatments. And the types of measures tend to be fuzzy, global, and dependent on subjective evaluations.

For example, a study of hospital outcomes may assign a code ranging from "1-6" to every patient at discharge. One can thus count the number of patients who recovered uneventfully, the number of complications, and the number of deaths, but what does this mean? Such codes are useless for comparing treatments or assessing overall quality of care. While they contain little information about the endpoint, they contain no information at all about the starting point (such as the correct diagnosis and the severity of illness). Further, all but the mortality rate are based on reviewer impressions.

Some "outcome" measures may really be "process" measures in disguise: the number of tests ordered, the amount of money expended, or physician compliance with a protocol. They are *system* measures, not *patient* measures.

Indeed, it is the outcomes for the system — and their "relative values" — that are generally of greatest interest to the new researcher. For example, "negative outcomes" include costs and "burdens to the health care system." And the very purpose of the practice guideline (the outcome of which is to be measured) may be to "enable the evaluation of physician practices, (e.g. utilization review, quality assurance), or to set limits on physician choices (e.g. recertification, reimbursement)"⁸ — not to cure patients' illnesses or to relieve their symptoms.

Even studies that do measure patient outcomes in great detail may really be about systems. In drawing conclusions, special atten-

tion should be paid to the length of follow-up. For example, the assessment of the *New York Times*⁹ notwithstanding, one can hardly be confident that a new "delivery system" has no adverse effects, based on a study that follows mild diabetics and hypertensives for only 2-7 years. (The mean baseline blood pressure was about 140/83 and the mean glycosylated hemoglobin was 9.3 to 9.7%.)¹⁰

What About the Controls? (Or, In Science, the Control Group Is Not Armed with Sanctions)

In today's society, we have become so accustomed to thinking of controls in terms of government agents carrying checklists, forms, and guidelines for imposing civil monetary penalties and other sanctions, that it may be useful to review the meaning of controls in the scientific sense.

Analyzing the results of an experiment is always an exercise in comparisons. The goal of the scientist is to disprove the null hypothesis, which states that the outcome in the "experimental" group is no different from that in the "control" group, i.e. that his intervention has had no effect on the outcome of interest. An alternate formulation is that "treatment A" is no different from "treatment B."

If the two outcomes are different, the question remains: was it the intervention that caused the difference? Or something else? The scientist must demonstrate that the control group is like the experimental group in every important respect except the factor being investigated.

Of course, it is immediately obvious in some outcome studies that *no* comparisons are made and *no* control group is present. The idea of practice guidelines is to promote uniformity.

Was the Study Unbiased? (Or, Scientists May Not Play with a Marked Deck or Loaded Dice)

Honesty is the *sine qua non* of science.

Since scientists are human

beings, it is always possible that their own biases can influence the outcome of the experiment, particularly when there is an unavoidable subjectivity in some of the observations. Therefore, the best science uses a randomized, double-blind method.

In clinical medicine, an idealized method is not generally applicable. However, one can and must make every effort to eliminate the most egregious sources of bias. This rules out the use of information that is prepared for the primary purpose of collecting payment from a third party. All hospital records are now contaminated with this type of bias. Choosing one "Diagnosis Related Group" code over another can mean thousands of dollars in profit or loss to the hospital. Including certain statements in the admissions history makes the difference between insurance coverage or denial of payment. Even prescribing certain treatments (an intravenous line, for example) influences the decision of the utilization review department or possibly even the committee that grants or denies staff privileges.

With the proliferation of "managed care" contracts, this type of bias is permeating out-patient care as well. Physicians spend thousands of dollars attending seminars in coding so as to "maximize reimbursement." The medical record is prepared for the benefit of clerks and administrators (not to mention lawyers). Although it does communicate information helpful to those trying to treat the patient, that is not its sole or even primary purpose.

The use of the computerized data on insurance claims forms for any type of "research" on practice guidelines or outcomes assessment makes a mockery of the scientific method.

Was the Sample Size Adequate and the Conclusions Statistically Significant (Or, Sometimes You Get Four Aces Without Cheating)

The perils of drawing conclusions from small samples are well known, at least to statisticians.

Clinical trials must have a large number of patients enrolled — for an adequate length of time — in order to reach statistical significance. The smaller the anticipated differences, the larger the groups must be.

When evaluating physicians for staff privileges on the basis of their complication rates or "efficiency," the small-sample problem must always be considered. People, even doctors, do sometimes experience a "run of bad luck." Clusters of adverse or fortunate events do happen, by chance alone. (And sometimes when you flip a coin you get heads three times in a row.) The probability that a particular cluster has occurred by chance alone can be calculated, as can the probability of observing a given difference between groups — the "P value." In comparisons of physicians' practices that I have seen — for example, the rate of Caesarian sections — such measures of "statistical significance" are not given.

If a difference *is* statistically significant, the reason may be something other than a difference in competence or efficiency. It could be that a highly respected physician receives a lot of referrals of high risk-patients. Conversely, just as the very "lucky" poker player *may* have aces up his sleeve, the most "efficient" doctors and hospitals may be "managing the case mix" to their pecuniary advantage. Thus, one may be dealing with biostatistical entities called the "rancid sample" or the "tilted target."¹¹

Analysis in a Vacuum (Or, the Fishing Expedition)

Once a large data set has been collected, it is highly tempting to mine it for any possible associations or "clusters." The problem is that *all* data sets will be found to have clusters of unusual events, due to chance alone, if one examines enough possibilities. When calculating for statistical significance, one must use a correction factor for making "multiple comparisons."

This fallacy is demonstrable from elementary probability theory, but it is not well recognized. It is the

basis for many outbreaks of public hysteria about alleged environmental risks. If public attention has been focused on a chemical, one can find a "cluster" of adverse events that might be attributed to chemical exposure. Of course, given any population, it is possible to find some disease or outcome or event that is present with a higher than "expected" prevalence. Sometimes, such a fortuitous observation can lead to recognition of an important, previously unsuspected risk. But usually it doesn't. A "fishing expedition" must always be followed up with additional observations; *the hypothesis to be tested must be defined in advance.*

The rationale for collecting computerized data from all patient encounters (now required by law in Maryland), is to enable the third-party payers and regulators to determine "What Works...." The fishing expedition is the only type of maneuver possible with such data.

Conclusions

"Practice guidelines" based on

"outcomes analysis" are not a scientific advance but a scientific veneer for practice based on fashion, peer pressure, and economic coercion. They can only exacerbate the "Doctor's Dilemma" of George Bernard Shaw, causing an even more "intense dread of doing anything that everybody else does not do, or omitting to do anything that everybody else does."

References

1. Forum on Health Care, University of Arizona College of Medicine, January 11, 1992.
2. Terrell HP. Practice guidelines: micromanagement on a broad scale, or a modest proposal for omniscience. *Medical Sentinel* 1996;1(1):11-15.
3. Orient JM. Practice guidelines and outcomes research, Part I; insights from the Clinton health care task force. *Medical Sentinel* 1996;1(1):9-10.
4. Sapira JD. The Art and Science of Bedside Diagnosis, ed. by Jane M. Orient, Urban and Schwarzenberg, 1990, p. 385.
5. Sapira JD. And how big is the spleen? *South Med J* 1981;74:53-60.
6. Debrow RJ, Calatayus JD, Abraham S, Caceres CA: A study of physician variation and heart sound interpretation. *Med Ann District of Columbia* 1964;33:305-308,355-356.
7. Lloyd SS, Rissing, JP. Physician and coding errors in patient records. *JAMA* 1985;254:1330-1336.

8. Wilson MD, Hayward SA, Tunis SR, et al. User's guide to the medical literature. VIII. How to use clinical practice guidelines. *JAMA* 1995;274:1630-1632.
9. Noble HC. HMO quality called equal at less cost. *New York Times*, September 8, 1995, p.10A.
10. Greenfield S, Rogers W, Mangotich M, Carney MF, Tarlov AR. Outcomes of patients with hypertension and non-insulin-dependent diabetes mellitus treated by different systems and specialties: results from the Medical Outcomes Study. *JAMA* 1995;274:1436-1444.
11. Feinstein AR. *Clinical Biostatistics*, St. Louis, Mosby, 1977.

Dr. Orient practices internal medicine in Tucson, Arizona, and is Executive Director of the AAPS. Her address is 1601 N. Tucson Blvd., Suite 9, Tucson, AZ 85716.

A few observations and much reasoning lead to error; many observations and a little reasoning to truth.

Alexis Carrel

THE ADVANTAGES OF AAPS MEMBERSHIP

- Legal Consultation Services • AAPS News • The Medical Sentinel
- Action Alerts on pending Congressional and State Legislation
- Legislative Updates

Yes, I am interested in AAPS.

- Please send more information.
- I am interested in receiving both the AAPS News & The Medical Sentinel for only \$50.
- Enclosed is \$275 for my first year's dues in AAPS.

Name _____
 Address _____
 City, State, Zip _____

800 635-1196

AAPS, 1601 N. Tucson Blvd., Suite 9, Tucson, AZ 85716

