# Foreword

FREDERICK MOSTELLER

*Harvard University*

*When the* Milbank Memorial Fund Quarterly *asked Sonja McKinlay to be guest editor of a special issue devoted to experimentation and social policy, the natural author of the Foreword would have been William G. Cochran. His many contributions to experimental design and analysis and studies of health, plus his growing attention to the policy arena, would have forced him to accept the almost automatic invitation. I can only join with the editors and the readers in regretting that his death has prevented this, and try to emulate his cautious attitude, fairness, and friendliness to all. I believe that he would have advised us not to rush to judgment.*

IN ADDITION TO ORGANIZING THIS ISSUE OF THE *Quarterly*, guest editor Sonja McKinlay also gave herself the specific task of describing the historical development of randomized clinical trials and their place in the growth of statistical methods. Her paper finds that the urge to test innovations combined in the mind of Sir Bradford Hill with the ideas of agricultural field trials to produce today's widely used methods for clinical trials. She explains the roles of various devices for strengthening these comparisons of therapies, preventions, or diagnostics.

McKinlay explains the merits and drawbacks associated with randomization, almost a sine qua non of the definitive clinical trial. Randomization offers control, balance, objectivity, and valid inference to populations. This treatment of randomization deserves special attention because it delineates so clearly and briskly the several uses. She reviews various designs of investigations, including sequential

Milbank Memorial Fund Quarterly/*Health and Society*, Vol. 59, No. 3, 1981

0160/1997/81/5903/0297-11 $01.00/0

methods, and problems of response measurement with a discussion of the placebo effect.

Thus, by describing with care the method of choice for comparing medical treatments, or social reforms, for that matter, McKinlay lays a solid foundation for the reader unfamiliar with randomized clinical trials.

In the social and medical area, we have begun to profit from what Hyman (1972) and others call secondary analysis, or sometimes meta-analysis. The researchers gather data from past investigations, either re-analyzing the data for the same purpose as that of the original study or making original observations for some new purpose, using collections of investigations rather than single ones. For example, using such an approach, we found out that of those subjected to controlled trials, fewer than half of social reforms (Gilbert, Light, and Mosteller, 1975) and about half of surgical innovations (Gilbert, McPeek, and Mosteller, 1977) succeed. Thomas Chalmers and his many colleagues have exploited this method for years. In his paper, Chalmers uses this technique to show us how much the use of controls has increased with the years.

Chalmers, a well-known physician, researcher, and medical educator, brings many specific clinical situations to our attention and introduces some extra information to try to explain the medical profession's response to controlled investigations. His paper elaborates with concrete examples on many issues. Chalmers points out that clinicians need to know about methodology to be able to judge the quality of research. He proceeds to cover important devices for controlling the investigation: blindness of patient and evaluator, care in performing the actual randomization, handling dropouts, and monitoring an ongoing study.

He turns then to the quantitative problems of statistical analysis and design. His and his coworkers' findings on sample size of investigation have shaken the medical world. Inevitably, we want investigations to be small for the sake of economy and speed. Still, we want good answers at the end. Chalmers finds that many investigations have not been large enough to have a reasonable chance of detecting substantial gains when an innovation provides them. This discovery raises serious problems, because a trial that has little chance of discovering what it sets out to determine puts patients at risk for little

reason. Rutstein (1969) says that a trial that cannot, because of its design, determine what it intends to prove is unethical.

"What should be the sample size?" is the first question asked of a statistician and a hard one to answer. The statistical issue we are discussing is technically known as the power of the test. Although the concept is fundamental to the good design of a comparative study, few investigators are aware of the idea. In sixty-seven clinical trials reported in 1979 and 1980 in the *British Medical Journal, Journal of the American Medical Association, Lancet,* and the *New England Journal of Medicine,* only 12 percent contained remarks about the statistical power of the investigation (Dersimonian, Charette, McPeek, and Mosteller, 1981).

Chalmers then addresses ethics, peer review, use of placebos, and informed consent. He is famous for his recommendation that randomization should begin with the first patient treated, and his arguments are carefully stated here. Readers can quickly grasp from his discussion the ethical and practical issues, which repay careful thought. This idea still lies in the realm of controversy. Stopping a trial needs much thought and work and new research that has yet to be carried out. Indeed, only recently did a Food and Drug Administration (FDA) panel of statistical consultants (Lagakos and Mosteller, 1981) recommend that research be conducted on this topic for animal experiments, so we have far to go.

Finally, Chalmers focuses on costs, with some observations on the general social process leading to clinical trials and about the future of clinical trials.

Although, as Chalmers's paper describes, the randomized clinical trial comparing therapies, diagnoses, or preventive measures has wide use, the costs and the difficulties make many yearn for other less demanding methods of appraisal. As a statistician, I have frequently been asked to draft an article comparing the strengths of the various methods of gathering data. For example, are not community-based studies of therapies as used in practice good enough? Will not data banks and registries provide the needed information? If we know the course of the untreated disease, can we not just study the new therapy alone without controls?

The desirability of a definitive article that compares the strengths of such methods for assessing technologies can no more be denied

than the desirability of inventing a perpetual motion machine. The trouble in both instances lies in the execution.

Only recently have statisticians begun to contribute solid ideas on how to use poorly controlled studies for comparing treatments, so we do not yet have a good basis for the proposed work. The question whether methods other than experimentation can offer the desired degree of rigor is wide open, and research on it will probably proceed over the next decade. We have neither adequate theory nor adequate practice for basing our actions on other methods. Yearning for these methods will not take us far; rather, we require massive research on these questions. Chalmers, in a variety of papers, including Grace, Muench, and Chalmers (1966), as well as Gilbert, McPeek, and Mosteller (1977), have given evidence showing that less well-controlled trials ordinarily lead to more enthusiasm for an innovation.

Certainly, we can say that an important field of research remains. One wonders, however, if the most rigorous methods have difficulty getting at the truth, how less rigorous ones can get at it better: presumably, by trading assumptions and experience for tight design.

When we do not have experimentation or other systematic ways of gathering data, evaluating therapies can be a long, slow process. When we do produce timely experimental data directly relevant to a therapy, not all physicians will follow the implied advice, but they do have the opportunity to do so; some may have sound reasons for not following it.

The social scientists, Howard Freeman and Peter Rossi, have years of experience in field work, social surveys, and experiments, much of it with an emphasis on health or medicine. They point out that since World War II social science and health have increasingly turned to experimentation on humans, both as part of ordinary research and development and as ways of evaluating policy reforms. The activity called "evaluation" has become a profession, with journals and societies devoted to it.

In addition to carefully controlled field trials for health care facilities and medical treatments and policies, we also have natural experiments and quasi experiments. A natural experiment might arise from comparing rates of dental caries with the amount of fluorine in the local, untreated water supply. The assumption is that, in deciding to live in the community, people have not been influenced by the amount of fluorine. We have an observational study rather than an experiment,

because we have not intervened to cause variation in fluorine. When more fluorine is associated with less tooth decay, we tend to think fluoridation reduces decay.

A quasi experiment arises when a change in policy occurs—often in government—as when a new law changes the speed limit, possibly reducing the number of fatal accidents. Or when a hospital changes its policy from giving an exploratory laparotomy to all patients with stab wounds to one of merely observing to see whether the progress of the patient shows the need for a laparotomy. The question will arise, "Have the number of infections and the number of days in the hospital gone down?" In quasi experiments, some intervention does occur, and "quasi" warns that full control does not. Donald Campbell once told me of a researcher who wrote in embarrassment to apologize that he had done an experiment instead of a quasi experiment. Campbell hopes that the elegant-sounding "quasi" does not mislead people into thinking it is to be preferred to an unqualified experiment.

Freeman and Rossi illustrate various forms of human experiments, some being large-scale field trials. They discuss new problems posed by such field trials and the difficulties of evaluating health programs already in place.

Even readers familiar with social and health experimentation will want to read the special section that discusses monitoring of interventions. These studies grew from the need to be sure that a program delivers the treatment claimed for it. Failure to implement is one way to make a treatment fail in its mission. After a good discussion, the authors illustrate monitoring in a school health demonstration in Chicago and a health education program, *Feeling Good.*

The historian Arnold Toynbee has given a challenge and response theory for the survival of states and empires. Essentially, states grow until they meet some kind of crisis—the challenge is usually environmental or military—and, if they are up to it, they survive, otherwise they die or are swallowed by others. Since most processes have beginnings, middles, and ends, such a theory is more a framework or mnemonic for describing events than an attempt to provide explanatory causes. In this latter spirit, John McKinlay offers his seven stages in the career of a medical innovation. They make it possible for him to organize the chaos that surrounds the introduction, dissemination, evaluation, and demise of an innovation. He does not pretend, or even suggest, that all medical innovations go through all

of these stages or even that the stages are ordered. Nevertheless, this idealized career helps us follow the conflicts and seeming irrationalities in our present mode of operation.

I might illustrate with his discussion of two forms of double standards. When a procedure has been applied under the mantle of treatment, few restrictions apply to the treating physician, but let the same procedure at the same time be used by the same physician to assess its safety and efficacy, and the physician is surrounded by constraints. No doubt many of these have been wisely developed, but the asymmetry leaves something to be desired. The surgeon William McDermott complains wryly that the most ethical 5 percent of physicians are getting 95 percent of the monitoring.

Similarly, McKinlay notes, when a carefully run randomized clinical trial comes under critical review, as is likely when it opposes standard wisdom, critics seize on the slightest scratch or nick in its methodologic armor to discredit it. Simultaneously, the weak information on the same innovation provided by poorly controlled investigations continues to be accepted without criticism.

Some of the criticisms of trials that McKinlay discusses might be ameliorated by theoretical research of the more global aspects of statistical design. We biostatisticians have much theory about randomization, stratification, controls, and so on. But we have not done much to tackle such questions as: How narrow should the population be in a controlled trial? Or how can we tell which groups might benefit more from one or the other therapy, rather than acting as if one were best for all? A number of such problems have been laid out, inviting prospective researchers to contribute to them (Mosteller, McPeek, and Gilbert, 1980). Such research requires considerable sophistication because, even though it is quantitative and mathematical, nuch of the effort goes into defining and modeling the problem. At first such work will seem oversimplified and subject to easy, but perhaps mistaken, criticism.

After indicating the overall problem area, McKinlay offers a general strategy to improve our process of accepting medical innovations. His bold program invites debate and possibly some reshaping. First, it insists on effectiveness from both old and new therapies. Second, it requires cost efficiency. Third, and this may produce fighting on the ramparts, it asks for social acceptability and equal accessibility for

all subgroups of the society. By proposing this strategy, McKinlay sets the stage for discussions among physicians, social scientists, and decision makers because achieving acceptability and accessibility demands results in activities where we have made only modest progress. Do we have the tools? It may need close examination to make sure we are not asking society to optimize several different outputs at the same time. It may also impinge on freedoms that our citizens do not care to relinquish. Altogether, then, it offers us a stimulating and controversial proposal.

In discussing such problems, the economist Thomas Schelling (1981) argues that to do the most good for people we may not wish to make things comparable for all groups. For example, if we were to decide to raise the safety standards of coal miners to levels comparable to those of workers in other industries, we might price marginal coal mines out of business. The miners, having no other ready source of work, would then be thrown on welfare and might regard their total standard of living as lowered rather than raised by our interference.

Concern has been frequently expressed that information from experiments and other assessments is little used by decision makers. Our difficulty may well be that we do not recognize use when we see it because we tend to think of a use as adopting a recognizable variation of something studied. Instead, the use may be that of understanding what to avoid, and then the onlooker cannot tell that the data have been used, although in a not very recognizable way. Until we have a good way of assessing the details of the political process, this problem of use will continue to plague us. Meanwhile, I continue to assume that policy makers like and use information, and that they have more dimensions of movement than users of other methods appreciate.

For those, like myself, who know little about the workings of the Food and Drug Administration, Stuart Nightingale's brief history provides some pegs for our tent of understanding. The legislative acts in 1906, 1938, and the 1962 amendment brought successively higher standards to the marketing of drugs. I am constantly amazed that a single act and its amendments can have such extensive ramifications and be reasonably responsive to public needs for decades. Nearly all drugs have now been reviewed for safety and efficacy, even very old

ones. This situation contrasts sharply with the lack of assessment in many other areas of medicine, as Banta and Behney point out in their article.

Nightingale leads us through the FDA's comprehensive and systematic arrangements for drug testing, beginning with animal experimentation, clinical protocol development, tests in limited populations, and, finally, general use.

He discusses how FDA confronts such nonroutine problems as: a not-well-tested but apparently marvelous innovation—a breakthrough drug—and postmarketing surveillance. FDA can post restrictions on use of devices (but not drugs) after release and can make requirements for reporting further information on all regulated products after release. He also reviews the responsibilities of the various parties to the testing of new drugs, including the institutional review boards, investigators, and sponsors.

Nightingale guides us to the policy areas where FDA is likely to be active in the near future. He discusses the issue of lag in introducing new drugs and lays out the issues in the distribution and use of an unusual investigational drug, tetrahydrocannabinol—a marihuana constituent. The FDA is studying new questions about measuring effectiveness and postmarket review. Finally, the difficult and time-consuming problem of international cooperation and uniformity in trials and data requirements has become a priority item. Since it took decades to get agreement on a few items for a uniform hospital discharge sheet, we must not be surprised if similar periods pass while we produce internationally comparable trials. Even so, the payoff should be worth the effort. A.L. Cochrane has suggested that two trials of the same therapy, possibly in different countries, would be especially valuable. Without some uniformity, comparability cannot be available, and controversy and confusions destroy consensus.

The Office of Technology Assessment (OTA), an advisory arm of the U.S. Congress, has produced many monographs and reports on health and medical practices since its establishment in 1972. The program on health is now managed by David Banta, with Clyde Behney an important coworker. As they explain, OTA does not directly assess technologies, in spite of its name, but it does, with its own staff and with outside help, review studies of technologies in an attempt to answer questions about the state of a situation. Congressional committees request information, and OTA does studies intended

to respond simultaneously to inquiries from several. In its many reports, OTA describes possible policies and gives the pros and cons of each. Thus, OTA offers a direct attempt to inform the policy process through technological assessments, some of which have been based on experiments.

Banta and Behney offer us a broad perspective. Like Freeman and Rossi, they see experimentation as part of a social movement or, as they would say, an experiment on experiments. They feel it important for us to appreciate this enterprise from a global point of view, even though they, and I, think it too early for a massive evaluation. For one thing, we suffer from believing in instantaneous action following information and we need a time perspective. For example, scientific innovations often take about twenty years to appear in practical applications because we require several distinct scientific breakthroughs to make one technological innovation. So it is early to try to appreciate the experiment on experiments.

After describing early work on technology assessment both outside and inside of health, Banta and Behney describe some important OTA studies of safety and efficacy of medical technologies and of cost-effectiveness. They explain the need for work of, say, the National Center for Health Care Technology. At this time, the Center seems to be discontinued, and, if so, one hopes that its activities will be taken over by some other organization.

Banta and Behney lay out a systematic program for technology assessment requiring us to *identify* technologies needing tests and to set priorities on these tests, to synthesize the information gained, and to disseminate it to decision makers. They explain the need for this program and more detailed steps for carrying it out. Especially noted are weaknesses in establishing priorities and effecting dissemination. They emphasize efficiency and safety as areas needing improvement.

As for the future, Banta and Behney see the problem as one of appreciating the larger picture. Instead of worrying only about the details that make an experiment successful, we need a system that can use information to move us from where we are to some desired state. They would have us identify the current and desired states, use assessment to guide us in handling change mechanisms, and then evaluate the effect of the mechanisms.

A linking theme in this issue of the *Quarterly* is that several authors regard experimentation and evaluation as methods to facilitate social

change. Without despair I cite among the virtues of this issue the many concrete illustrations of complexity, and I appreciate the forward-looking programs outlined by the authors. We readers must be grateful to the editor for having the foresight to persuade Sonja McKinlay to organize the issue, and for her ability to select and persuade authors who have vision beyond the narrow technical side at the same time that they have mastered and appreciated it. This issue will be of special value to a number of organizations that are about to launch studies oriented toward the future of health, such as several seminars at Harvard University under the general direction of David Hamburg, and a new study of medical technology assessment at the Institute of Medicine. It comes during a time of ferment and turmoil with changes wrought by the new administration in Washington, and so is doubly timely. This issue raises our appreciation of the total process of innovation, evaluation, and progress to a new level of self-consciousness. It also offers us a set of plans for discussion, debate, and development. This issue of the *Quarterly* makes a neat contribution by raising our sights and advising us to use missiles of larger caliber and greater range.

## References

Dersimonian, R., Charette, J., McPeek, B., and Mosteller, F. 1981. Statistical Information in Comparative Clinical Trials. Unpublished.

Gilbert, J.P., Light, R.J., and Mosteller, F. 1975. Assessing Social Innovations: An Empirical Base for Policy. In Bennett, C.A., and Lumsdaine, A.A., eds., *Evaluation and Experiment.* New York: Academic Press.

———, McPeek, B., and Mosteller, F. 1977. Progress in Surgery and Anesthesia: Benefits and Risks of Innovative Therapy. In Bunker, J.P., Barnes, A.B., and Mosteller, F., eds., *Costs, Risks, and Benefits of Surgery.* New York: Oxford University Press.

Grace, N.D., Muench, H., and Chalmers, T.C. 1966. The Present Status of Shunts for Portal Hypertension in Cirrhosis. *Gastroenterology* 50:684–691.

Hyman, H. 1972. *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities.* New York: Wiley.

Lagakos, S., and Mosteller, F. 1981. A Case Study of Statistics in the Regulatory Process: The FD&C Red No. 40 Experiments. *Journal of the National Cancer Institute* 66:197–212.

Mosteller, F., Gilbert, J.P., and McPeek, B. 1980. Reporting Standards and Research Strategies for Controlled Trials: Agenda for the Editor. *Controlled Clinical Trials* 1:37–58.

Rutstein, D.D. 1969. The Ethical Design of Human Experiments. *Daedalus* 98:523–541.

Schelling, T.C. 1981. Economic Reasoning and the Ethics of Policy. *The Public Interest* 63:37–61.

---

*Acknowledgments:* The preparation of this foreword was facilitated by National Science Foundation Grant SES 75-15702.

*Address correspondence to:* Professor Frederick Mosteller, Department of Statistics, Harvard University, Science Center, 1 Oxford St., Cambridge, MA 02138.