Experimentation in Human Populations

SONJA M. McKINLAY

The Memorial Hospital,
Pawtucket, Rhode Island;
American Institutes for Research,
Cambridge, Massachusetts

EXPERIMENTATION IN HUMAN POPULATIONS IS NOT a new concept but has existed wherever and whenever man has experienced curiosity concerning his environment. It has always been one of the primary approaches in the search for knowledge. However, as with all concepts, there are shifts in emphasis in usage and meaning, between different research fields and between different periods (Kuhn, 1964). Words associated with the concept also change in usage, meaning, and fashion. Indeed, the papers in this issue of the *Quarterly* provide immediate illustrations of such words, some of which are used consistently, some of which appear to have different meanings in different contexts.

This paper presents one attempt to define and describe the concept "experiment" in the context of health-related research on human populations in the latter half of the twentieth century. A basic, somewhat general definition is first proposed and the major historical developments that have influenced current use of the concept (within this broad definition) are reviewed. The discussion then focuses on issues that are particularly relevant to health-related experimentation, such as the role of randomization, the sequential availability of subjects, and other frequently raised ethical issues. The purpose of this dis-

Milbank Memorial Fund Quarterly/*Health and Society*, Vol. 59, No. 3, 1981
© 1981 Milbank Memorial Fund and Massachusetts Institute of Technology
0160/1997/81/5903/0308-16 $01.00/0

cussion is to identify and distinguish those problems currently encountered in the experimental approach that are contrived, possibly for other purposes, those that are real with available solutions, and those that are real but for which no good solutions currently exist. This paper is not intended as an exhaustive treatise, but rather serves as an overview of the current state of experimentation, while attempting to refocus discussion and highlight the major issues.

A Definition

The Oxford English Dictionary (1971:930) offers as a definition of the noun "experiment" the following rather general meanings:

1. The action of trying anything, or putting it to proof; a test, trial. . . .
2. A tentative procedure; a method, system of things, or course of action, adopted in uncertainty whether it will answer the purpose. . . .
3. An action or operation undertaken in order to discover something unknown, to test a hypothesis, or establish or illustrate some known truth. . . .

Rather than choose the definition that seems to fit most aptly the type of experiment being considered here, it would seem preferable to offer a more focused definition that embraces the various approaches to be discussed. It is clear, from the types of research usually labeled in the literature as "experimental," that the two essential ingredients of any definition should be 1) the *deliberate manipulation* of material (subjects); and 2) the *careful observation* of responses to this manipulation.

The second requirement of "careful observation" is, indeed, common to all forms of empirical investigation, whether a sample survey, a census, a comparative or case-control study. All researchers would agree that observation must be as detailed, consistent, reliable, and as unbiased as is possible, within the constraints of the investigation.

The first requirement is the one that clearly distinguishes experiments from all other forms of empirical investigation. Many attempts have been made to do this. Moser and Kalton (1971:224) imply the same distinguishing characteristic in their text on social surveys, as does Kempthorne (1977:4), who defines a study as "an experiment

only if certain forces are varied at the will, whim or choice of the investigator." However, the importance of this distinguishing factor has recently been blurred by the introduction of such terms as "quasi experiment," "natural experiment" (both of which represent internal contradictions), "randomized controlled trials" (RCTs), and the parallel increase in the use of the word "experiment" as a synonym for "randomization" (Campbell and Stanley, 1966; Cochrane, 1972).

A third component of the basic definition is frequently included and is essentially implied by the first; namely, that the purpose of an experiment is to establish a cause-effect relation or, more generally, an empirical model. Logically, one can establish such a relation as cause-effect *only* by deliberate manipulation. It is *never* sufficient to observe a relation in order to infer that one factor causes another, or that observations always fit a particular model. Observations may *suggest* a possible cause-effect relation or a possible model (rule) that a set of observations will fit under defined conditions. In accord with the empirical approach first expounded by Descartes (1637) and developed particularly in the nineteenth century, however, these observations must be followed by tests or experiments in order to establish the reality of the cause-effect relation or model. Again, this purpose of experimentation has been blurred by recent developments in multivariate forms of analysis, facilitated by developments in computer science. It is now so easy to fit observations to a variety of complex models that one tends to overlook the data source and its implication for inference. The fact remains that, unless the observations are the result of planned manipulation, any cause-effect relation or model identified in the analysis (however sophisticated) remains speculative at best.

The definition of "experiment" that is therefore offered as a basis for discussion in this paper is: *The planned manipulation of material, subjects, or processes, by the experimenter, and the careful observation of responses to this manipulation, in order to establish a cause-effect relation or a rule (model) for the variation of observations.*

A Brief History

The present role of experimentation in research on human beings, as defined above, has evolved from at least two distinguishable approaches—the continuing need to test all innovations in the treatment

or care of human beings, and the modern empiricism (evolving in the latter part of last century) that pervades basic scientific research. The distinction between these approaches, their purpose and development, may partially explain some of the tensions that permeate the current use of experimentation on humans, particularly in health-related fields.

The first, humanistic approach encompasses a common-sense recognition that potential improvements to the human condition must be tested. Daniel, in King Nebuchadnezzar's court, recognized this need in his concern to provide the best possible diet for the youths in his care (Daniel, chapter 1, verses 1–15). This early description of a test between two diets meets the basic definition of an experiment. Two competing diets were deliberately assigned to the youths in such a way that the superior diet could be established. One could certainly argue the quality of the observations made, but this was a function of sophistication in observational techniques, rather than a violation of the definition. The well-known test of the efficacy of lemons in preventing scurvy, conducted by Lind in the eighteenth century, is another excellent example of such an experiment (Lind, 1753). The closing of contaminated wells by Snow (1936) was strictly an experiment, as the subsequent incidence of cholera was carefully monitored to establish a cause-effect link. The initial work on smallpox vaccination can be similarly defined as experimental.

In other words, there is a well-established but not always recognized tradition of testing or experimenting in all fields relating to improvements in health—if we accept the basic definition of an experiment given above. The emphasis in this approach has been manipulation. The ensuing observation has not always been consistent or as carefully made as it might have been.

The second approach had its foundation in the positivism of the nineteenth century, exemplified in the work of investigators and philosophers such as Comte (1975), Booth (1889–1902), and Gauss (1889), among others. The emphasis was on the careful, methodical recording of observations from which knowledge could be developed and is exemplified in the classic genetics experiments of Mendel (1965). These experiments involved the manipulated growth of pea varieties in order to verify, from observed proportions, a postulated model for the inheritance of characteristics. It is perhaps not coincidental that Fisher (1935), to whom modern experimental design approaches are generally attributed, began his career as a geneticist.

Although positivism with its imperative of exact observation influenced all fields of enquiry to varying extents, it was inevitably expressed most clearly in those sciences in which control of units and observations was facilitated—such as chemistry, physics, genetics, biology. Such treatises as Popper's (1959) or Nagel's (1961) on the logic of scientific inquiry reflect this bias. It was within the highly controlled, manipulable environment of agricultural research at Rothamstead Experimental Station that the modern theory of experimental design was established and continues to develop. A similar environment at Iowa State University provided the impetus for experimental research in the United States under the leadership of George Snedecor.

In such highly controlled environments, with homogeneity among experimental units the rule rather than the exception, the emphasis is on optimizing precision with the fewest possible observations (minimum cost) through complex designs and careful specification of the analytic model. Issues relating to the assignment of units and unbiased observation of responses are relatively unimportant in such situations; this is reflected in the standard statistical texts, which accord cursory or no attention to assignment of units or response observation. This omission is also noted by Kempthorne (1977).

These two divergent approaches, with their distinctive histories, were finally joined by Bradford Hill (1962) in his pioneering clinical trials in Britain, to be quickly followed by the field trial of poliomyelitis vaccine in the United States in 1957, designed by W.G. Cochran (Francis, 1955). The adoption of this modern experimental approach in social research outside the health field followed in the 1960s and now encompasses a wide range of examples (Boruch et al., 1978).

The remainder of this paper focuses on four related areas in which the feasibility of experimentation on human populations is frequently challenged, especially in health-related research: randomization, cost, population (definition and availability), and observation of response.

Randomization

The procedure known as "randomization" involves the random assignment of available experimental units to "treatments" (manipulated

protocols). This method of unit assignment was first proposed by Fisher (1926) and generated a well-known debate with Gosset (1938), which was apparently resolved in favor of randomization at the expense of systematic control strategies. The basic motivation for a randomized procedure was recognition that some material—even in the basic sciences—was not sufficiently controllable. The prevailing notion, particularly in a physics or chemistry laboratory, was that external variation could be well controlled by sufficient attention to homogeneity of units, consistency of manipulation, and reliability of response measurement. Variability in response was attributed to poor control of one or more of these facets. Experiments on more variable material, such as large fields and genetically diverse animals or plants outside a laboratory environment, forced Fisher and his colleagues to confront the issue of uncontrollable variation and to devise methods of handling it. The result was randomization, which was designed to serve the following major function: *that the error component in any observation would be additive and independent* of any treatment or other manipulated effect (such as blocking) *and therefore estimable.* This property would be reflected in a model that consists of added components.

Although the reason for randomization was initially limited to this issue of the separation of response variability in estimation, the power of the procedure in other respects began to be appreciated as it was applied to increasingly variable material. In particular, it was realized that randomization of a sufficient number of experimental units was likely *to equalize sources of variability that could not be controlled in the design.* With highly variable human subjects, this is a powerful argument, as the degree of design control in most human experiments (whether a clinical, institutional, or "field" setting) is severely limited. Part of the controversy over the results of the University Group Diabetes Program trial, for example, centered on this use of randomization (Cornfield, 1971).

A further advantage of this procedure was the fact that *valid statistical inference did not require additional random sampling* of experimental units from a population. The randomization alone was sufficient. This important property is discussed at length by Kempthorne (1952, 1974) and Cox and Kempthorne (1963). Unfortunately, this same property is not always appreciated by researchers who may still confuse the two quite distinct concepts of randomization and random sampling. Moreover, because the results of most human experiments are

intended for application to the most general populations possible, there is an independent and real concern that the experimental units be representative of such populations. Because the established means of ensuring such representation is, in fact, random sampling, the two procedures are frequently confused and considered erroneously as part of one process with the same purpose.

The final property of randomization that, in many social or clinical experiments with considerable room for human error, tends to outweigh any other advantage *is its assurance of objectivity in unit assignment.* When carefully applied, this extraordinary procedure ensures that there is no subjective bias—intended or unintended—in the assignment of units to treatments. Early trials that used alternate assignment procedures were shown to have important differences in patient groups not attributable to treatment, which may have occurred because of physician bias in patient referral; for examples, see Snow (1965) and Wright (1948). More recently, suspected cheating on random assignment was considered a major weakness in the results of the recent British trial of coronary care units versus home care for myocardial infarctions (Mather et al., 1976).

This last mentioned property, and the second advantage concerning equalization of variation sources, now dominate as the reasons for randomization in clinical and field trials. Certainly, given the need to avoid any subjective bias in unit assignment and the relatively high, uncontrollable variation among the units (human subjects, singly or in groups), randomization has become an essential feature of well-controlled human experiments. Without this procedure there can be no assurance that the experiment was indeed sufficiently well controlled.

Despite this imperative, randomization is not always accepted by researchers as necessary or even feasible. Particularly in clinical experimentation, physicians who collaborate as researchers, by accepting randomization, publicly acknowledge that they are uncertain as to the relative effectiveness of treatments under comparison. This uncertainty conflicts with the widely accepted image of confidence expected of physicians in practice. Similar conflict can occur wherever professional reputations are involved (for example, asking a judge to randomly assign sentences in court). In such situations, the potential for sabotaging the randomization procedure is increased and/or the

recruitment of subjects for the experiment may be so difficult and selective that the study becomes impracticable.

Therefore, not only must the researchers be convinced of the need for randomization, but they must also themselves be convinced that there is not clear evidence on which to base a preference for one treatment protocol above another and be prepared to recruit subjects on this basis. Essentially, the researcher must *unlearn* the traditional dictum: "It is unethical not to provide the best service or treatment thought to be available," and learn a new one, namely: "It is unethical not to randomize when the best treatment or service is in doubt." This need for re-education is not yet generally appreciated, although Chalmers (a contributor to this issue) has been persistently proposing it in the medical field in the United States for at least the last decade—see for example Chalmers (1975)—while Cochrane has been doing likewise on the other side of the Atlantic (Cochrane, 1972).

The resistance to randomization may appear somewhat contradictory to the tradition, alluded to above, of testing innovations in the health field. However, it should be understood that, although such testing was definable as experimentation according to the definition employed in this paper, it was not necessarily the *well-controlled* experimentation typically found in, for example, a chemistry laboratory. Rather, such testing was perceived as an addition to clinical or field experience, performed on an ad hoc basis (depending on the facilities and subjects available) and generally using subjects as their own controls. Even Lind was constrained by the availability of ships.

Cost

The dollar cost of human experimentation, especially in the United States, has become a major issue and is frequently a primary reason given for not embarking on a well-designed trial with sufficient numbers of subjects. In the clinical field this problem is exacerbated by 1) the requirement that the cost of all treatments be included in the direct cost of the trial, even though such treatments would otherwise be covered by third-party payments; and 2) the growing use of multicenter trials in recognition of the need for large numbers. This is discussed elsewhere in this issue by J.B. McKinlay (1981), who points

out the false economy of this attitude in terms of knowledge accumulation. The fact remains that one hundred observational studies cannot produce the confidence in results that one well-designed experiment provides. And the cost of that experiment, however high, probably would not exceed the combined costs of the observational studies.

Apart from the financial cost of conducting well-controlled human experiments, there is the cost of the treatments themselves, both in dollars and in risk. To date, there appears to be no example of an experiment on human beings designed to optimize precision within cost constraints. This issue is discussed at length in S.M. McKinlay (1981) and in relation to work on sequential experimental design; see for example Colton (1963). It would seem logical, if one or more treatments under consideration are costly, to minimize the number assigned to these treatments without sacrificing precision. Similarly, if one or more treatments carry higher risks of side effects (or risks of more serious side effects) the number exposed to such protocols should also be minimized.

With the complex treatment protocols being considered, particularly in clinical trials, it would seem timely to introduce these considerations into the design rather than continuing blindly with traditional equal assignment. Nam (1973) and S.M. McKinlay (1981) provide useful results concerning the effect of disparate costs on the distribution of numbers between treatments, which indicate that random assignment in ratios other than 1:1 may be more efficient.

The major problem that has yet to be resolved (if it is indeed resolvable) is how to combine elements of financial cost with the risks of adverse effects. As yet no feasible method of assigning dollar costs to risks has been devised to permit the two types of cost to be combined.

Population Definition and Availability

The wide variability in human subjects has generated a set of problems not found in a laboratory situation and not solved by randomization. First, it is clearly difficult to include in any experiment a representative sample of the population in which the results of the experiment are to be applied. In some large experiments such as the field trial of the

Salk vaccine, or some large multicenter clinical trials such as that of Mather et al. (1976), it is possible because of numbers and the ubiquity of the treatment. In other situations, perceived ethical considerations may limit the experimental population to the less sick or to the more seriously ill even though the treatment is applicable to a wider group of patients. The restricted inference offered by these experiments may detract from any results obtained, especially if linearity (a straight-line relationship) cannot be assumed in extrapolation. Further, from those trials that have included patients with varying disease severity, it is clear that linearity of effect can seldom be assumed. Unanticipated interactions may be identified, such as the success of coronary artery by-pass grafting for patients with severe left ventricular disfunction only (Takaro et al., 1976), and the apparent superiority of home care for older patients with myocardial infarctions (Mather et al., 1976).

This inferential difficulty is frequently exacerbated by the rigidity of experimental protocols that could not be easily transferred to routine care settings. The experiment may demonstrate effectiveness of a protocol under ideal settings, with no assurance that this level of effectiveness would be maintained under conditions of routine use. This discrepancy remains a major issue in contraceptive research that distinguishes "theoretical" from "use" effectiveness of contraceptive methods. To date, most experiments in this field have addressed highly controlled theoretical effectiveness, but most data on use effectiveness are observational.

Clearly there is a need to establish a potential sequence of experiments in some situations, progressing from limited to wider populations, from rigid to more flexible protocols. This has been addressed with respect to drug evaluations by the Food and Drug Administration (see Nightingale, 1981), but has not yet been systematized in other health-related fields (this subject is discussed by other papers in this issue).

The numbers required in many well-designed experiments and the *sequential* availability of subjects are also forcing researchers to focus on a related inferential problem. Human populations are highly variable, not only in cross-section but also longitudinally. Subjects, programs, diseases, diagnoses, treatments are all in a state of continual change. It is true that most such changes occur relatively slowly, taking five or more years to become detectable (for example, the

unaided decline in chickenpox incidence). Provided experiments are conducted within relatively short periods, one can assume that the experimental and target populations remain reasonably comparable. However, with the requirements of large numbers of subjects and the length of time required to admit sufficient subjects and observe responses, many major experiments may span five to ten years or more. In this same period, the population and/or alternative treatment protocols may change sufficiently that either the experimental results are outdated or they are no longer applicable, given current population needs.

Complementing this problem is the need to reach quick decisions so that the minimum number of experimental subjects are exposed to the inferior treatment. Fully sequential designs, as first proposed by Armitage (1960), have had limited use, mostly in the 1960s. The major restriction on the use of such designs has been the need for immediate response measures, the waiting time for which is either less than or equal to the waiting time between admissions of subjects to the trial. As is noted by Peto et al. (1976), the current trend in medical and other health-related research is toward the use of variable, time-dependent responses (such as mortality, length of remission) that do not adapt easily to fully sequential designs.

An alternative that is receiving increasing attention is the possibility of interim analyses, scheduled and adjusted in such a way that the experiment is not prematurely terminated. Early terminations in, for example, the Coronary Drug Project (Canner, 1977) and the University Group Diabetes Program (see Cornfield, 1971) have left equivocal results that are still debated. There is clearly a need for statistical methods that encompass an "early stopping rule"—a partially sequential design.

Response Measurement

The importance of objective, replicable, and unbiased measurement has always been recognized in basic scientific research. In many laboratory experiments, the major part of the experimenter's energy and resources is usually spent in devising the optimal means of measuring a response. This focus is consistent with the emphasis on observation typical of positivist movements in scientific enquiry.

At the same time, clinical observation has typically consisted of signs and symptoms gathered and interpreted by the physician or other health worker. The need for reasonably objective observation has always been offset by the dynamics of the patient-physician relationship, including the physician's exploitation of the "bedside manner" and patient suggestibility. The power of what is now termed the placebo effect has always been recognized by the medical profession and, when there were few reliable treatments available, was frequently used as a necessary supplement to drugs or leeches (see, for example, Wolf, 1950). The rapid spread of sophisticated combination drugs and medical technology in recent years has detracted from this important phenomenon, although there are signs that it may again assume a position of prominence. Certainly there is evidence that the placebo effect may account for one-third to one-half of a treatment's effectiveness for some conditions (Hubbe, 1975).

When reliance must be placed on a subject's self-report to measure responses, it becomes essential to separate placebo from treatment effects in an experiment. Blind and double-blind techniques were devised to accomplish this by keeping the subject and possibly the observer in ignorance of the actual treatment administered. This is most easily accomplished in drug trials, although it has also been used successfully for surgical procedures (Beecher, 1961; Ruffin et al., 1969), and has even been proposed with respect to acupuncture (correct versus incorrect placement of needles)! However, the maintenance of "blindness" in the design is not always easy. Texture or side effects may permit subjects or observers to identify the placebo drug (Blumenthal et al., 1974). Moreover, with the increasingly stringent requirements for informed consent, the possibilities for placebo treatments are severely curtailed. Not only are sham surgical procedures out of the question (who would consent to the risks of anesthesia and the discomfort and disfigurement of a surgical wound with a 50 percent chance of a dummy operation?), but the validity of an observed placebo effect is questionable when subjects know that a placebo treatment is a possible alternative. The whole point of a placebo treatment is the patient's ignorance of its existence. The patient must believe that the treatment is real.

Even apparently objective responses can be subject to considerable bias. Blood pressure measurement is an obvious example, observers tending to underreport when the boundary that generally defines

hypertension is reached. The use of a "random zero" sphygmomanometer is one remedy, a randomly determined amount of mercury disguising the true values until after the measurements are taken and the "random zero" is then determined for subtraction. Variability in diagnosing cause of death and in reading X-rays provides further examples of supposedly objective responses that are affected by human judgments and biases (Garland, 1949).

Conclusion

This overview of experimentation in human populations has highlighted aspects of its development as well as some major problems and issues surrounding its current application. Although early use of this approach appeared to be innovative and successful, especially in the 1950s and 1960s, recently experiments appear to have become more difficult both to initiate and to complete.

There are several issues that offer partial explanations for this apparent lessening of interest. *First,* most social experiments are now focusing on comparisons for which differences tend to be relatively small or subtle. Such comparisons require large numbers in order to reveal the differences in question and this requirement is costly, both in financial and recruitment resources. Large, multicenter clinical trials and multicommunity social experiments are becoming the rule rather than the exception, but logistics and cost limit their introduction. *Second,* informed consent and other legal requirements for experimentation involving humans have limited, to some extent, the types of experiments that can be performed. It is now more difficult to experiment on children and to use placebos, for example. (Sham operations are now out of the question.) *Third,* the time required to complete many modern experiments—five years is now becoming standard and even longer may be required—is a contraindication for all but the major, relatively stable treatments or programs. There is no point embarking on a five-year experiment to evaluate a program that is likely to be replaced or radically changed in three years.

These apparent obstacles to human experiments unfortunately can be all too readily used as legitimate excuses not to perform them. It is, after all, so much easier to complete a cheaper, shorter retrospective observational study that will provide quicker results.

We must not lose sight of the fact that, however difficult they may be to perform, well-designed and controlled experiments provide the only sure means of detecting cause-effect relationships. Despite the obstacles, experiments still provide what is ultimately the most cost-efficient method for evaluating definitively treatments and programs in human populations.

References

Armitage, P. 1960. *Sequential Medical Trials.* Oxford: Blackwell.

Beecher, H.K. 1961. Surgery as Placebo. *Journal of the American Medical Association* (July 1):1102–1107.

Blumenthal, D.S., Burke, R., and Shapiro, A.K. 1974. The Validity of "Identical Matching Placebos." *Archives of General Psychiatry* 31(2):214–215.

Booth, C., ed. 1889–1902. *Labour and Life of the People of London* (17 volumes). London:Macmillan.

Boruch, R.F., McSweeny, A.J., and Soderstrom, E.J. 1978. Randomized Field Experiments for Program Planning, Development and Evaluation. *Evaluation Quarterly* 2:655–695.

Campbell, D.T., and Stanley, J.C. 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Canner, P. 1977. Practical Aspects of Decision-Making in Clinical Trials: The Coronary Drug Project as a Case Study. *Biometric Society (ENAR) Meeting* (April). Chapel Hill, N.C.

Chalmers, T. 1975. Randomization of the First Patient. *Medical Clinics of North America* 59:1035–1038.

Cochrane, A.L. 1972. *Effectiveness and Efficiency.* London: Nuffield Provincial Hospitals Trust.

Colton, T. 1963. A Model for Selecting One of Two Medical Treatments. *Journal of the American Statistical Association* 58:338–400.

Comte, A. 1975. *Cours de Philosophie Positive.* Paris: Librairie Larousse.

Cornfield, J. 1971. The University Group Diabetes Program: A Further Statistical Analysis of the Mortality Findings. *Journal of the American Medical Association* 217:1676–1687.

Cox, D.F., and Kempthorne, O. 1963. Randomization Tests for Comparing Survival Curves. *Biometrics* 19:307–317.

Descartes, R. 1637. *Discours de la Méthode.* Paris: Librairie Larousse.

Fisher, R.A. 1926. The Arrangement of Field Experiments. *Journal of Ministry of Agriculture, Great Britain* 33:503–513.

———. 1935. *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Francis, T. 1955. An Evaluation of the 1954 Poliomyelitis Vaccine Trials: Summary Report. *American Journal of Public Health* 45:1–63.

Garland, L.H. 1949. On the Scientific Evaluation of Diagnostic Procedures. *Radiology* 52:309–327.

Gauss, C.F. 1889. Allgemeine Lehrsaetze in Beziehung auf die im Verkehrten Verhaeltnisse des Quadrats der Entfernung Wirkenden Anziehungs- und Abstossungs-Kraefte. In *Ostwalds Klassiker der exacten Wissenschaften, 2.* Leipzig: Wilhelm Engelmann.

Gosset, W.S. 1938. Comparison between Balanced and Random Arrangements of Field Plots. *Biometrika* 29:363–379.

Hill, A.B. 1962. *Statistical Methods in Clinical and Preventive Medicine.* Edinburgh: Livingstone.

Hubbe, P. 1975. Controlled Clinical Trials of Drugs for Use in the Prophylaxis of Migraine. *Danish Medical Bulletin* 22:92–96.

Kempthorne, O. 1952. *Design and Analysis of Experiments.* New York: John Wiley and Sons.

———. 1974. Sampling Inference, Experimental Inference and Observational Inference. *Proceedings of the Mahalanobis International Symposium on Recent Trends of Research in Statistics.*

———. 1977. Why Randomize? *Journal of Statistical Planning and Inference* 1:1–25.

Kuhn, T. 1964. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Lind, J. 1753. *A Treatise of the Scurvey.* Edinburgh: Sands, Murray and Cochran.

Mather, H.G., Morgan, D.C., Pearson, N.G., et al. 1976. Myocardial Infarction: A Comparison between Home and Hospital Care for Patients. *British Medical Journal* 1:925–929.

McKinlay, J.B. 1981. From "Promising Report" to "Standard Procedure": Seven Stages in the Career of a Medical Innovation. *Milbank Memorial Fund Quarterly/Health and Society* 59 (Summer):374–411.

McKinlay, S.M. 1981. A General Approach to the Design of Clinical Trials with a Pre-specified Termination Point. *Biometrics* 37. Forthcoming.

Mendel, G. 1965. *Experiments in Plant Hybridization* (Reprint in English). Edinburgh: Oliver and Boyd.

Moser, C.A., and Kalton, G. 1971. *Survey Methods in Social Investigations.* London: Heinemann Educational Books.

Nagel, E. 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation.* New York: Harcourt, Brace and World.

Nam, J. 1973. Optimum Sample Sizes for the Comparison of the Control and Treatment. *Biometrics* 29:101–108.

Nightingale, S.L. 1981. Drug Regulation and Policy Formulation. *Milbank Memorial Fund Quarterly/Health and Society* 59 (Summer):412–444.

Peto, R., Pike, M.C., Armitage, P., et al. 1976. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient: I. Introduction and Design. *British Journal of Cancer* 34:585–612.

Popper, K.R. 1959. *The Logic of Scientific Discovery.* London: Hutchinson.

Ruffin, J.N., Grizzle, J.E., Hightower, N.C., McHardy, G., Shull, H., and Kirshner, J.B. 1969. A Cooperative Double-Blind Evaluation of Gastric "Freezing" in the Treatment of Duodenal Ulcer. *New England Journal of Medicine* 281:16–19.

Snow, J. 1936. On the Mode of Communication of Cholera. In *Snow on Cholera,* 1–175. New York: The Commonwealth Fund.

Snow, P.J.D. 1965. Effect of Propanolol in Myocardial Infarction. *Lancet* 2:551–553.

Takaro, T., Hultgren, H.N., Lipton, J.J., and Detre, K.M. 1976. The VA Cooperative Randomized Study of Surgery for Coronary Arterial Occlusive Disease. II. Subgroup with Significant Left-Main Lesions. *Circulation* (Supplement 3) 54: III-107–III-117.

Wolf, S. 1950. Effects of Suggestion and Conditioning on the Action of Chemical Agents in Human Subjects: The Pharmacology of Placebos. *Journal of Clinical Investigations* 29:100–109.

Wright, I.S. 1948. Report of the Committee for the Evaluation of Anticoagulants in Treatment of Myocardial Infarction. *American Heart Journal* 26:801.

Address correspondence to: S.M. McKinlay, Pawtucket Heart Health Program, Memorial Hospital, Pawtucket, R.I. 02860.