From "Promising Report" to "Standard Procedure": Seven Stages in the Career of a Medical Innovation

JOHN B. McKINLAY

Boston University

> If men could learn from history, what lessons it might teach us! But passion and party blind our eyes, and the light which experience gives us is a lantern on the stern, which shines only on the waves behind us!
>
> S.T. Coleridge, 18 December 1831, in T. Alsop, *Recollections*, 1836.

IN 1972, IN AN EDITORIAL FOR THE *Annals of Thoracic Surgery*, Tom Chalmers (1972:323–327), a highly respected contributor to this special issue, provided examples of widely advocated and commonly used therapies that either had never been established to be effective by adequate clinical trials, or in a number of instances had actually been shown to be without merit (Chalmers, 1972: 326). The editorial concluded with the question, "Can we learn from our mistakes of the past?" That question is as poignant today (perhaps even more so) as it was when posed by Chalmers nearly a decade ago. This paper is a response (admittedly belated) to this question and is divided into two main sections. The *first* describes some types of mistakes by outlining the typical career of a medical innovation. How do medical innovations (diagnostic techniques, surgical procedures, and medical interventions) become part of established medical practice? What stages do they typically pass through? What

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

0160/1997/81/5903/0374-38 $01.00/0

admission requirements (if any) must be satisfied before innovations gain privileged access to the House of Medicine? What types of evidence are employed, when, and by whom, in support of which innovations? Having outlined the typical career of an innovation, the *second* section describes an alternative approach, based upon evidence derived from randomized controlled trials (RCTs), to the allocation of ever scarcer public funds to health care. This alternative approach takes account of some past mistakes and the wastefulness and irrationality associated with the ways that things are done at present.

On the basis of a review of many studies of diffusion and the careful examination of the careers of several different types of innovation, it is considered useful to break the diffusion process into the seven distinct stages outlined below. Several aspects of the career approach to the diffusion of medical innovations should be highlighted here in order to minimize misunderstandings. There is, *first,* no suggestion here that every innovation passes through each of the seven stages in the exact order in which they are discussed. The concept of a career, with its seven stages, is simply a heuristic device employed to highlight particular issues and possible points of intervention. The following discussion is intended to identify the typical or usual stages in the career of an innovation—the pattern that is followed more often than not. There are situations in which, as in the case of tragedies like thalidomide, some of the typical stages are circumnavigated, or telescoped, but the career of an innovation has at least a beginning point (the promising report) and an ending (established procedure, or erosion and discreditation). The heuristic advantages of this notion of a typical career are fairly obvious: it enables one to break fairly complex social behavior and political processes into a manageable form, identifies possible points of intervention, and bestows a semblance of order on the present chaotic state of diffusion studies. Therein, however, lies a major disadvantage: it suggests that more order and coherence exist than is actually the case.

Second, every effort is made to avoid the suggestion that the medical establishment is only self-interestedly involved in the diffusion of innovations. The ideological basis of much recent research on medical technology appears to require a villain to whom responsibility can be ascribed: hospitals, and more particularly physicians, are highly visible targets. This paper represents an attempt to move beyond this superficial level to a consideration of the activities of hospitals, phy-

sicians, and other interest groups in relation to more basic structural processes that impinge upon them.

Third, there is obviously variation within, and between, medical specialties in the sponsorship of innovations, and the willingness to undertake proper evaluations. Spodick (1973) argues that there is a double standard in the acceptance of reports of surgical versus medical treatments, and that this arises from professional and lay attitudes. After reviewing some 70 reports in specialty journals appearing during 1971, to compare the methods used for evaluating medical versus surgical treatments for cardiovascular disease, he found that 9 of 16 qualifying studies evaluating medical treatments were controlled. None of the 49 studies of surgical intervention involved a controlled study. Cochrane (1979) considers it not unreasonable to judge the medical profession, and its specialties, by the use they have made of the RCT technique. He rather humorously awards first prize (the "Bradford" in praise of Sir Austin Bradford-Hill) to the tuberculosis chest physician, with almost unlimited praise. After considering psychiatry, surgery, and cardiology, he awarded the wooden spoon to obstetrics and gynecology, with the following explanation:

> The specialty missed its first opportunity in the sixties when it failed to randomize the confinement of low risk pregnant women at home and in the hospital. This was followed by a determined refusal to allow "Pap smears" to be randomized, with disastrous results for the whole world. Then having filled the emptying beds by getting nearly all pregnant women into the hospital, the obstetricians started to introduce a whole series of expensive innovations into the routines of pre- and post-natal care and delivery, without any rigorous evaluation. The list is long but the most important were induction, ultrasound, fetal monitoring and placental function tests. The specialty reached its apogee in 1976 when they produced 20 percent fewer babies at 20 percent more cost. G and O stands for gynecologists and obstetricians but it could also stand for GO ahead without evaluation! (Cochrane, 1979:11)

1. The Stage of the "Promising Report"

Many studies of the diffusion of innovations resemble those frustrating occasions when one arrives late at the theater, after a performance has commenced, and is forced to leave before the final curtain. One is

never really on top of what occurs, finds it difficult to unravel the relationships between the various actors, and is left wondering how the whole thing ended anyway. Diffusion studies tend to cover what can be termed the midpoints in the career of some new procedure, drug, machine, or whatever (Gordon and Fisher, 1975; Russell, 1978). They provide useful information on the most publicly visible stages, but give inadequate attention to either their points of origin, or where they eventually end. By failing to trace innovations back to their points of origin in order to identify the interests and processes that launched them, such studies have limited utility for social policy (McKinlay, 1977a, 1977b). And without some understanding of the final stages, social policy upon the allocation of resources cannot be properly informed by the successes and failures of the past.

The careers of most medical innovations seem to be launched with the appearance of an enthusiastic report on some promising performance, increasingly in the mass media. One reads, almost daily, of startling "scientific" breakthroughs and new weaponry for the battle against illness and death (Sontag, 1978). A few examples will vividly illustrate the point. The *Boston Globe* (1980) recently reported that a doctor at Johns Hopkins University *is* extending the lives of patients suffering from inoperable liver cancer by making them temporarily radioactive. In 7 of 8 patients treated, tumor size has been drastically reduced from 70 to 18 percent of the liver in one case, and from 50 to 5 percent in another. To quote this report:

> The new technique floods the liver with continuous radiation for days or weeks, rather than brief bursts, which is the usual treatment. To do this, a protein made by liver cancers is extracted, purified and injected into rabbits. The rabbits then make antibodies—disease fighting proteins—against the cancer protein. The rabbit antibodies are then heavily dosed with radioactive iodine and injected into the patient. The antibodies then attach themselves to the liver cancer, and the radioactive material begins its destructive work.

A similarly optimistic report of a different type of innovation recently appeared in the *New York Times* (1980a):

> Researchers at the Mayo clinic have developed an advanced X-ray machine that displays organs in three-dimensional moving images that may then be electronically dissected without ever touching the

> patient with a scalpel. The device, known as the Dynamic Spatial Reconstructor or DSR, has been built at a cost of approximately $5.2 million with grants from the National Institutes of Health. The DSR is regarded by many scientists as the most significant technological advance in diagnostic medicine since the invention of computerized axial tomography, known as CAT. . . . The main component of the DSR is a 15 foot circular gantry that supports 14 X-ray tubes (the final model will have 28 tubes) aimed at the gantry's center. To obtain a DSR scan, a prone patient is inserted into the gantry, which spins around his body once every four seconds. The X-ray tubes take 60 X-ray photographs a second, with each pulse of radiation lasting 350 millionths of a second. . . . [the DSR] stands 15 feet high and weighs 17 tons.

While on acronyms, one must not overlook the *Newsweek* (1980:63) report on the PET, or positron-emission tomography—a machine that detects changes in chemical physiology:

> In a PET scan, the patient first receives an injection of deoxyglucose, a compound chemically related to glucose. The compound contains radioactive fluorine. Then the patient lies with his head inside the scanning device. The radioactive deoxyglucose—which has been absorbed by the brain cells—emits positively charged particles called positrons. These immediately collide with negatively charged electrons normally present in the cells. Each positron-electron collision produces high energy particles called photons, and PET's detector records their path and speed. This information is processed by a computer that runs out a composite color picture on a display screen.
>
> Various shades of color in the PET image indicate levels of glucose metabolism. A region of high activity might show up as light beige, while a low area would be deep russet. In their most significant preliminary studies, researchers have found altered glucose metabolism in schizophrenia and manic-depressive psychosis.

It is understandable that the media should seize upon these promising reports, since they are highly newsworthy for a public conditioned to expect miracle cures for the diseases that plague it. Many magazines and newspapers regularly devote a section or column to promising reports, and have a staff of medical reporters constantly looking for the latest newsworthy innovation in professional journals and through contact with the researchers directly. These promising media stories usually report activities that meet no methodological

criteria whatsoever. Moreover, it is not always possible to obtain information on the specifics of the unpublished "preliminary studies," the results of which frequently form the basis of the promising report. Certainly, the association of a promising report with "respected scientists," and with prestigious institutions, serves to arouse public interest (especially if, as is usually the case, it concerns a life-threatening problem such as heart disease or cancer).

A second, but not exclusive, source of promising reports may be found in major medical journals, some of which set aside space for accounts of the management of a few, or even solitary, cases. Dr. A, from a respected institution, may describe how he successfully treated Mrs. B, who was suffering from X, by employing Y. Medical journals, along with those in other fields, tend to publish only reports of "successful" interventions. One seldom reads of *un*successful interventions, even though their frequency may be equal to, and probably greater than, those purported to be successful. The fact that much can be learned, and many mistakes avoided, from the reporting of negative findings is persistently overlooked. Although perhaps reporting careful and sophisticated measurements involving intricate apparatus, these case reports, which usually discuss just a few subjects (and often only one), are inferentially worthless. Despite their appearance in professional journals, they are no more reliable than the sensational media stories already discussed, and certainly have no value as a basis for social policy.

Sometimes, after reading these isolated "promising reports" in either the media or in professional journals (or both), but usually after word-of-mouth endorsements among colleagues (one of the principal means by which physicians and researchers exchange information on innovations), someone may decide to check out the promising report on a series of cases—say, 25, or even 100. Some medical researchers seem to believe that the soundness of an inference is related to the number of observations from which it is derived, or perhaps the length of period of follow-up (see, for example, Vineberg, 1975; Sheldon et al., 1975). This paper will argue that such reasoning is seriously flawed because, while numbers are certainly important, other equally essential methodological criteria are overlooked, and such studies still constitute only promising reports.

Clinicians sometimes do attempt exploratory or pilot studies of the *effectiveness* of an innovation at this stage. Although well intentioned,

these studies are usually also severely limited and can most appropriately be included in the category of uncontrolled observational reports. Chalmers discusses these studies and suggests that they may actually observe the conduct of properly designed and conducted evaluations:

> The so-called "pilot study" is a trap that may be the major factor leading to the lack of decent randomized controlled trials in the evaluation of new therapies. Unfortunately, when a physician decides he has an exciting new therapy, he usually feels he cannot start a controlled trial immediately because he is not sure of the dose and the patients to select; so he does a pilot study of consecutive patients. That prevents him from ever doing a randomized controlled trial for one of three reasons: He is so impressed with the efficacy of the drug in the uncontrolled trial that he cannot do a study for ethical reasons, and he publishes his "excellent results" in a preliminary paper. He concluded that a controlled trial should be done, but he does not do it because he is convinced that the drug works. It is often ten years before other clinical investigators, stimulated by a lack of success in less well patients, report equally uncontrolled negative series or finally do a controlled trial. A *second* possibility is that the originator of the therapy cannot do a controlled trial because the treatment seems so ineffective that he decides he cannot subject more people to it; yet it is entirely possible that it is the selection of patients who receive the treatment rather than the treatment itself that is at fault. This is especially true in the case of drastic "last resort" therapies. . . . At any rate, the therapy appears so unfavorable that a controlled trial would be unethical. The *third* possibility is that the therapy appears similar to other therapies, and the investigator has no incentive to spend his time doing a controlled trial to prove that a suggested therapy is no different from the standard. . . . *The only way to avoid this trap is to randomize the first patients.* (Chalmers, 1975a:755; emphasis added)

2. The Stage of Professional and Organizational Adoption

Up to this point support for the innovation appears somewhat disparate, coming from the manufacturers of the technology and/or from enthusiastic medical researchers. During this second stage more widespread and influential support is mobilized. Here we see movement from scattered support for the innovation to its *adoption* by powerful

interest groups, involving professional associations, institutional structures, and other resources. The term "adoption" is employed to distinguish the scattered support evident during the first stage from the organized commitment of institutional resources during this second stage. It denotes a unique relationship with the innovation: the act of formally accepting and taking up some activity and using it as one's own, without the idea of its having been another's—to embrace or espouse it (Webster's Dictionary, 1976; Oxford English Dictionary, 1971). Such a special relationship must also be distinguished from "acceptance" of an innovation by the public, which denotes a more passive consent or approval of something, to regard it as proper, normal, or inevitable. Whereas professional *adoption* usually involves an investment (resources, time, reputation), public *acceptance,* as we shall see, suggests that people favorably receive or simply approve of something.

With regard to professionals and their associations, there are obviously many reasons for the adoption of innovations. Some physicians may be simply responding to peer pressure, known to be very strong in medicine (Coleman, Katz, and Menzel, 1966). Others may view it as an opportunity to deliver improved care, to be seen to be up-to-date, scientific, more professional (also strong motivations in medicine), or just helpful. Perhaps because of what Freidson (1971) terms their "clinical mentality," physicians may be precipitately eager to adopt innovations through their sincere desire to respond to the problems of disease and suffering that their patients may present. Warner (1977) has shown that rapid diffusion of new treatments occurred as a "desperate reaction" in the case of leukemia patients, when limited treatment alternatives were available.

Some commentators, who regard the professions as a conspiracy against the laity, consider personal financial gain as the motivation for physicians' adoption of innovations. Although this explanation is appealing in some circles, it is too simple an explanation of the phenomenon being discussed. Undoubtedly, some physicians, in common with workers in just about every occupation, may adopt an innovation because of a personal financial interest. This is just a fact of life and is certainly not unique to the occupation of doctoring (McKinlay, 1977b). But it is doubtful whether this is a primary motivation for the majority of physicians. In adopting innovations, physicians and their associations believe that they are being more

effective, humane, scientific, or whatever, and *secondarily* derive financial benefit.

With regard to the adoption of innovations by medical organizations (hospitals, medical centers, health clinics), it appears that organizational and economic considerations may be more important than they are among physicians (Kaluzny, Veney, and Gentry, 1974). In adopting an innovation, administrators may, of course, be simply responding to intense pressure from their medical staff. Such a response, although consistent with the requests of physicians, may be motivated by a different set of reasons. For example, the adoption of certain innovations may be viewed as an attempt to enhance a hospital's reputation in the community, thereby improving its competitive position with respect to other hospitals. Additionally, adoption may be affected by the nature of a hospital's relation with other organizations and interests (insurance companies, banks, construction companies). One study reveals that expensive technologies diffuse more rapidly as the percentage of hospital resources derived from third parties increases (Cromwell et al., 1975; Kaluzny and Veney, 1977). According to Russell (1976:559) the paths of diffusion for hospitals have been quite similar for most of the recent major innovations: "In general, the facility in question first gained a foothold in large hospitals. Then, as large hospitals adopted it in increasing numbers, it began to spread down through the size distribution to smaller and smaller hospitals, with the smallest generally the last, and the slowest, to pick it up."

In a study of the diffusion of computed tomography (CT) scanners, Banta (1980) found that they followed the pattern of other expensive medical technologies in that the largest hospitals were also the first to adopt cobalt therapy, electroencephalographs, and intensive care units. The way in which hospitals at present adopt innovations not surprisingly results in some distributional inequalities. With respect to CT scanners, for example (Banta, 1980:261): "Only one of New York City's city hospitals, Bronx Municipal, has a scanner. Many large public hospitals whose main clientele are the the poor are without CT scanners; e.g., Bellevue and Kings County hospitals (New York), Charity Hospital (New Orleans), Cook County Hospital (Chicago), and Cleveland Memorial Hospital."

A hospital's affiliation with a medical school has been shown, in studies of intensive care units (Russell and Burke, 1975) and nuclear

medicine facilities (Rappaport, 1978), to be related to early adoption of an innovation. Other studies (Cromwell et al., 1975; Banta, 1980) did not uncover this relationship.

Medical education gives the career of many innovations an early and influential boost and creates formidable impediments to the removal of those that eventually prove worthless or dangerous. All professionals are reluctant to alter practices that they have been taught. Innovations gain added legitimacy once they find their way into the medical curriculum and receive endorsement from influential educators in distinguished medical institutions. All students receive their first exposure to norms of practice from respected clinicians, who can recount first-hand experiences with a condition and its associated therapies. These experiences are essentially of the same quality as the promising reports (media stories and individual case histories) already discussed and should never be confused with reliable scientific evidence. Although they may suggest areas that are worthy of properly designed studies that may eventually yield such evidence, clinical experience or opinions can never be a substitute for scientific evidence, no matter how distinguished the observer or how numerous the observations. Unfortunately, this same clinical experience (frequently unsystematic observation) is the basis upon which most medical students learn. No wonder such a high proportion of what medical students are taught has a half-life of only a few years and comes to be viewed as worthless or actually harmful (e.g., radical mastectomy). The careers of many eventually discredited practices are temporarily prolonged through the support received from medical education, and by professional reluctance to relinquish what has been taught. With increasing specialization, students and practitioners may be trained to be dependent on certain practices or technologies; hence their continuing livelihood is to some extent contingent on the perpetuation of them. It is understandable, therefore, that some specialties are uncritically committed to particular interventions, and that they vigorously resist attempts to displace and sometimes even to evaluate them (surgical specialties are obvious examples).

It is essential that we be clear on the sources and quality of the information employed by physicians, hospitals, and medical educators as a basis for decision-making during this second stage in the career of an innovation. If this early information is defective, then subsequent action may be useless or harmful, and the resources devoted to it

totally wasted. With regard to physicians, several studies have shown that a physician's adoption of a drug and his or her subsequent prescribing behavior is largely determined by drug industry sources (Silverman and Lee, 1974; Subcommittee on Health, 1974). One study on the adoption of a new drug found that detail men (the sales representatives of pharmaceutical companies) were the first source of information for about half the physicians, and drug-house mail periodicals for about a quarter. The final source of information, before adoption, was drug-house mail and periodicals for a third of the physicians, colleagues for a quarter, and professional journals for only a fifth (Miller, 1975; Stross and Harlan, 1979). Virtually all physicians use the *Physicians' Desk Reference* (PDR) for information on drugs; two-thirds use it four or more times a week (Subcommittee on Health, 1974). The PDR contains listings that are essentially the package inserts prepared and paid for by the manufacturers and subject to approval by the Food and Drug Administration (FDA). Although a minimal standard of accuracy in this information is ensured by FDA regulation, biases in reporting are to be expected from such interested sources.

An excellent paper by Waldron (1977:45) contrasts evaluations of Valium and Librium in two industry-sponsored sources (advertisements and the PDR) with what she considers to be two more independent sources (the *Medical Letter* and articles in the *New England Journal of Medicine*). She concludes that

> the two industry-sponsored sources (the advertisements and the PDR) recommend these drugs for substantially more uses than the two independent sources (the *Medical Letter* and *The New England Journal of Medicine*). Aside from the obvious factor of motivation to sell, this reflects the difference between reliance on uncontrolled studies, in which apparent efficacy is inflated by placebo effects and spontaneous recovery, as compared to reliance on controlled studies. This discrepancy between drug industry sources and independent medical sources is greater in the earlier years when a higher proportion of the available studies were not properly controlled studies. An additional misleading aspect of advertising information for these products is the citation of 177 references in one ad of which 160 had nothing to do with the use recommended in the ad. . . . All four sources gave rather similar lists of adverse side effects, except in the earliest years when several important adverse effects were omitted in the advertisements and the PDR. . . . Industry sources

> tended to evaluate these drugs more favorably when they were newer. . . . Finally, the most striking time trend in these sources of drug information is that before 1970, virtually all the information available on Valium and Librium in *The New England Journal of Medicine* was in the form of advertisements rather than articles.

Waldron's paper deserves careful attention by anyone interested in the empirical basis (or lack of it) upon which innovations were adopted by physicians, hospitals, and medical educators during this second early stage in their career. Moreover, her analysis refers to the quality of information available for the class of medical innovation (drugs) for which there is the best quality control.

3. The Stage of Public Acceptance and State (Third-Party) Endorsement

Partly because of exposure to promising reports (Stage 1), but mainly as a result of professional and organizational adoption (Stage 2), a general approval (acceptance) of the innovation emerges among the public. This approval broadens its base of support and creates a constituency that can be appealed to in further advancing the career of the innovation. Public acceptance takes the form of a generalized belief that the innovation is a "good thing" and ought to be available. It is usually less well formulated and organized than the boost received from its adoption by professional and organizational interests. Having once fostered acceptance and even a demand among the public, these interests are in a position to satisfy it, while appealing to a demand that they may have created as justification for their activities with respect to the innovation.

In the typical career of an innovation being described, public acceptance is placed *after* the stage of professional and organizational adoption, despite the frequent assertion by these interests that, in adopting an innovation, they are simply responding to public demand. Generally speaking, public demand or community enthusiasm for an innovation must receive impetus and direction from professional interests that are already committed to the innovation. Demand does not usually occur in a vacuum. At the same time, manufacturers, professionals, and/or medical organizations may use public acceptance

or demand to legitimate their association with an innovation, and as a major reason for its expansion.

An innovation can be said to have "made it" when it eventually receives endorsement or support from the state and/or is underwritten by third parties. Two mutually supportive activities on the part of professional and organizational interests may induce the state and/or third-party insurance to underwrite certain innovations. The first activity, which can be termed the "indirect method," involves intensive lobbying with public officials, so-called expert testimony before legislative committees, campaign contributions to potentially supportive individuals and parties, and so forth. Certain important interests in the medical care field (American Medical Association, American Hospital Association, among others) have full-time lobbyists in Washington and strongly influence decision-making upon the allocation of state funds to medical research and practice. The second activity, or "direct method," involves obtaining support for the innovation among community groups or interests, which then pressure the state to support the particular innovation. The state's response to professional or organizational interests and public demand should not be viewed as a response to two separate constituencies, as they are sometimes depicted in the literature. Rather, they can be viewed as two methods—one direct and one indirect—by which the state is induced to underwrite a promising but yet-to-be-tested practice. There is nothing new in the suggestion that, generally speaking, the state and other third parties act *not* on the reasonable basis of reliable evidence, but on the basis of some combination of professional, organizational, and public pressure.

Again, it must be emphasized that despite the good intentions of well-motivated legislators and decision makers, the "authoritative" reports from state agencies, so-called expert testimony from authorities in the field, and the magnitude of the resources involved, *an innovation at this highly public third stage in its career usually remains without formal evaluation.* Usually, it still awaits a study that meets even minimally acceptable methodological criteria. The claims of manufacturers, the opinions of enthusiastic researchers, the well-intentioned adoption by experienced clinicians and educators, as well as public demand, are no substitute for a proper evaluation, and do not provide a rational, scientific basis for policy decisions. Once the state acts to support an

innovation and social policy is implemented, the career of an innovation can be viewed as having passed the point of no return.

The defective empirical foundation of the state's endorsement of an innovation is manifest in the research and development that the state subsequently funds. Having taken the step of endorsing an innovation (through, for example, reimbursement mechanisms and outright grants), the state, often through the funding of research and development activities regarding effectiveness, cost efficiency, and so forth, seeks to determine whether in fact it was the correct step to take. A careful review of the early careers of many different innovations reveals that, more often than not, the step was in quite the wrong direction and wasted resources, diverted professional and organizational resources to unproductive activities, and misled the public (Bunker, Barnes, and Mosteller, 1977).

Hemminki and her colleagues (Hemminki, 1980; Hemminki and Falkum, 1980) recently studied the number and quality of clinical trials in the applications submitted by the drug industry to licensing authorities in Finland, Norway, and Sweden. Many clinical trials were included but most of them were uncontrolled, otherwise deficient, or concerned only some of the indications applied for. Many of the reports cited had not been published and, since the submissions are secret, were not available to physicians or other researchers for evaluation. New drugs were sometimes registered for indications for which there were no good controlled trials. Thus, efficacy could not have been proved by the documentation included in the application. One suspects that a similar situation exists in most other countries.

4. The Stage of "Standard Procedure" and Observational Reports

There follows a period during which the innovation (having received professional and public support and legitimation through state endorsement and third-party coverage) achieves the privileged status of a "standard procedure." For a period of time it becomes generally accepted by interested parties as the most appropriate way of proceeding with a particular problem or situation. It is probably incorrect to refer here to the activity as an "innovation" (although we shall

continue to do so in this paper), since at this stage it has graduated from being just another promising performance (something new with great potential) to the position of being an established and respected activity. Although there is a bias against reporting unsuccessful or untoward performances, they certainly occur but are usually dismissed as infrequent, the result of having poor material to work with, public misunderstanding, and so forth. So entrenched has the activity become that it takes rare courage for any individual or group even to question its effectiveness or desirability. To do so, as we shall see, is to invite retaliation from professional organizational interests, public indignation, and even in rare cases sanctions from the state.

Again, it should be recalled that, despite the resources and institutions already committed to the innovation at this midpoint in its career, it still remains without any formal evaluation. A clear example is provided by computer axial tomography (CT scanning), which was alluded to earlier. Banta (1980:263), after discussing the diffusion of this technology, suggests that:

> Despite more than five years of experience with CT scanning, its usefulness and ultimate place in medical care are largely unknown. The development and diffusion of CT scanners took place without formal and detailed proof of their safety and efficacy. The evidence existing today did not come from well-designed, prospective clinical trials, but from analyses of clinical experience. However, this evidence is restricted almost entirely to assessing diagnostic accuracy and usefulness, and gives little indication of the effects of CT scanning on therapy planning or on patient outcome. (See also Fineberg, Bauman, and Sosman, 1977.)

Virtually the same situation occurred with respect to electronic fetal monitoring (Banta and Thacker, 1979). Other examples are cited in companion papers in this issue.

The position of the innovation in any medical care system is further secured by the many comparative observational studies that are conducted as its career develops. These "studies" are not dissimilar to critics' reports in the theater, the not-always-independent opinions of supposedly knowledgeable individuals, who may have witnessed a performance, or even participated in its production. Several features of these studies should be highlighted: First, many are underwritten by the state as part of the effort already discussed to ascertain the

effectiveness of an innovation *after* it has received general endorsement, and they usually take the form of retrospective studies, case reports, or follow-up investigations of an arbitrarily selected series of patients who have already been subjected to the innovation. Second, they are frequently initiated and conducted by the constituencies identified with the preceding stages (manufacturers, professional interests, and hospitals) who may have a vested interest in uncovering a beneficial outcome for the innovation. Thus, the objectivity of such observational studies could be seriously compromised. Third, these observational studies usually suffer crippling methodological limitations, such as inadequate sample size, restriction to a highly selective group of patients or problems, lack of an appropriate comparison group, use of subjective outcomes only (see Gore, Jones, and Rytter, 1977). Schneiderman (1975) provides examples (methyl GAG for myelogenous leukemia and 5-FU for certain forms of cancer) of how early uncontrolled results were actually misleading in evaluating potential cancer treatments. Another example is provided by clofibrate (Atromid-S), which was supposed to lower cholesterol in the blood and thereby prevent heart attacks, the leading cause of death in the United States. After thirteen years of widespread use and a proper evaluation (by the World Health Organization), it was found that users who took the drug regularly were 25 percent more likely to die of a broad range of disorders, including cancer, stroke, respiratory disease, and, ironically, heart attack, than those who got a placebo capsule! Fourth, although comparative observational studies sometimes provide useful information relating to the cost efficiency and social acceptability of an innovation, they seldom add much to knowledge concerning the *effectiveness* of the innovation in relation to the problem that it is designed to assist. Indeed, the term "evaluation" is becoming synonymous with studies of aspects of efficiency (Cochrane, 1972), and whether or not the recipients of an innovation (and sometimes even those who employ it) are "satisfied" (McKinlay and Dutton, 1974). Decisions concerning allocation of resources to or continuation of particular activities are increasingly influenced by whether such client or patient satisfaction is manifest.

In view of these and other limitations, it is difficult to determine from most observational reports whether the innovation is actually effective and whether some observed outcome may with certainty be attributed to it; for a review of the literature on observational studies

see S.M. McKinlay (1975). For every one controlled trial that provides evidence against an innovation's effectiveness, there are sometimes literally hundreds of observational studies that produce support for it, at considerable public expense. This will be recalled at a later point when we consider the cost of properly controlled trials.

Chalmers (1975) reports that in 1970 there were 77 papers in professional journals reporting results on coronary artery surgery on over 5,000 patients. Only 2 of these studies were controlled and, unlike most of the remaining observational studies, both provided evidence against the effectiveness of this surgery. Elsewhere (Chalmers, Sebestyen, and Lee, 1970) he describes how 61 uncontrolled studies of emergency surgery for bleeding peptic ulcer resulted in enthusiastic support for the procedure, although the only 3 controlled trials available failed to demonstrate the superiority of this surgical intervention. Chalmers has also shown, in a review of clinical trials of anticancer agents, that only a very small percentage of trials are in any way controlled (Chalmers, Block, and Lee, 1971). In 1972 there were 152 studies reported in the English language medical literature of internal mammary artery ligation, but only 2 were properly controlled (made use of randomization), and both found the procedure to be of no value (Chalmers, 1972). One recent study critically reviewed 134 different articles published in English between 1965 and 1979 that compared ambulatory and inpatient care with regard to clinical outcome and economic cost. Only 4 of the 134 reports provided sufficient data to allow statistically valid conclusions (Berk and Chalmers, 1981). If it could be done, the costs of the hundreds of observational studies supporting the use of coronary care units (CCUs) (not to mention the costs of the units themselves) should be compared with, say, the costs of the Mather and/or Hill trials, which showed that most patients with myocardial infarction did as well at home as they did in the CCUs (Mather et al., 1971, 1976; Hill et al., 1978).

Coronary artery bypass surgery today (1981) is at about this fourth stage in its career. Having moved from the status of a promising report, through the stages of professional adoption and public acceptance, it now enjoys a favored position as a standard or conventional surgical means of handling heart disease for more than 100,000 patients annually (*New York Times*, 1980b). Medical workers and organizations are honestly committed to the procedure, believe it is an effective and ethical approach, and also derive considerable financial

benefit from it. Public support for the procedure remains at a high level, and state and third-party arrangements underwrite most of its costs. To suggest that the procedure is ineffective, or even undesirable, is to court organized hostility and even ridicule. Despite the claims made for it, and the phenomenal investment of resources (estimated to be now around $2 billion a year in the United States), only one properly designed and conducted objective evaluation has been initiated and is still in progress (Murphy et al., 1977).

Mundth and Austen, after an exhaustive three-part review of around 150 different reports on surgical measures for coronary heart disease, were unable to turn up a single randomized controlled trial. At the time of writing they concluded that

> the duration of effectiveness of symptomatic relief, the effect of the functional status of the left ventricle and the effect on longevity have not yet been documented in toto. . . .
>
> *Controlled clinical trial is the ideal scientific solution for answering the questions raised concerning the effectiveness of coronary-artery surgical treatment.* . . . Whether randomized clinical trials comparing results of surgical versus medical therapy can be undertaken "ethically" has been a source of considerable debate. . . . However, when medical therapy has not been used or when it has been successful to a large extent in controlling symptoms, as in mild stable angina, *clinical trial with a prospective randomized study is ethically not only feasible but also advisable.* Similarly, meaningful data comparing the effect on longevity of surgical versus medical therapy can only be obtained with a randomized clinical trial. (Mundth and Austen, 1975:129; emphasis added)

Some features of this extensive report by two respected surgeons should be highlighted: first, the medical "evidence" for coronary artery bypass grafting (CABG) consisted entirely of observational reports (such as retrospective studies, case reports, clinical experience, follow-up studies); second, although not one randomized controlled trial (RCT) had been conducted among the more than 100 reports cited, the superiority of evidence derived from RCTs was recognized; third, proposals for prospective RCTs were alluded to with the hope that such studies could provide answers to the many questions that are still not completely resolved; fourth, it is reasonable to assume, on the basis of their commitment to the superiority of evidence derived from RCTs,

that their practice would be influenced by the results (either positive or negative) of such studies.

Two years after Mundth and Austen's review, Murphy and his colleagues published preliminary results of the cooperative trial sponsored by the Veterans Administration (Murphy et al., 1977). This report confirmed what many people suspected, that CABG reduces the incidence and severity of angina but results in no difference in the survival of almost 600 patients with chronic stable angina (excluding those with obstructive disease of the left main coronary artery), randomized into medically and surgically treated groups, and followed for 21 to 36 months. The nature of the response to this and other trials will be discussed in detail in the following two stages (five and six) which are more directly concerned with RCTs. Braunwald (1977:663) in considering some of the criticisms of the Veterans Administration study suggests:

> An even more insidious problem is that what might be considered an "industry" is being built around this operation; the creation of the facilities for open-heart operations in community hospitals in which no other cardiac procedures are performed and the enlargement of surgical facilities in teaching hospitals; the proliferation of catheterization and angiography suites as well as facilities for performing screening exercise electrocardiograms; and the expansion and development of training opportunities in clinical cardiology, cardiovascular surgery and cardiovascular radiology. *This rapidly growing enterprise is developing a momentum and constituency of its own, and as time passes, it will be progressively more difficult and costly to curtail it materially, if the results of carefully designed studies prove this step to be necessary.* (Emphasis added)

5. The Stage of the Randomized Controlled Trial (RCT)

Whether an innovation is worthwhile or not is an issue that, for many, appears to be satisfactorily settled by the sheer volume of observational reports. However, such reports never really place the issue of the *effectiveness* of an innovation beyond dispute. Byar and his colleagues (1976:79) have discussed the relative merits of different ways of assessing various medical treatments, and they conclude that

"randomized clinical trials remain the most valuable method of evaluating the efficacy of therapies." Ten British and American statisticians have also reached the same conclusion (Peto et al., 1976, 1977). Although there are many understandable questions concerning RCTs, this author has yet to meet an informed researcher who doubts the inferential superiority of experimentally derived evidence (Guttentag, 1971; Riecken and Boruch, 1974). The ratio of observational reports to RCTs may be the order of 100 to 1 (as we have seen was the case with aortocoronary bypass for arteriosclerotic heart disease), but there are few who would not prefer experimental results, given a choice. The imbalance between the two types of studies results in large part from the difficulty of designing, and then implementing, an RCT in situations where an innovation is already standard procedure, and powerful interests and reputations are invested in its continuing success. It should be recalled at this fifth stage of a career that it is not really an innovation that is being considered for evaluation, but an activity that has become a norm in the field. In general, the more advanced the career of an innovation, the more difficult it is to undertake an RCT.

Anyone who has designed and implemented an RCT, or had one sabotaged, will be aware of the formidable obstacles that are placed in the path ahead and of the power of the interests that must be accommodated (Conner, 1977). Such obstacles tend to be glossed over in published reports. It is interesting to speculate on how many RCTs are initiated for every one that is successfully completed and eventually finds its way into print. The obstacles and objections to RCTs take many different forms (McKinlay, 1973). Some are legitimate and reflect genuine concern from various quarters. Others, probably the majority, are not legitimate objections and are designed to make proper evaluations virtually impossible, thereby protecting the innovation from potentially incriminating results. Ethical issues provide one excellent example. Some people, particularly clinicians, do have honest ethical and legal qualms concerning randomization (Kempthorne, 1977) and the withholding from a group of cases of some standard procedure that they consider worthwhile. Chalmers (1975a, 1975b) considers this issue and concludes that most ethical quandaries would be removed by randomizing the first patient before observations or reports could intervene and prejudice the physician or researcher.

Double standards exist over whether a study is considered ethical or not. On the one hand, it is ethical to subject all patients to an innovation, despite the absence of reliable evidence concerning its effectiveness, or its potential for harm. But, on the other hand, it is unethical to withhold the as yet unevaluated innovation from certain patients in order to ascertain its effectiveness and potential for harm. So-called ethical and legal objections are a major obstacle to the conduct of randomized controlled trials, although there are other problems such as their high cost, who should sponsor and conduct them, whether they are inappropriate to some interventions and situations, whether they can alter individual clinicians' behavior, and so forth.

Recognizing the legitimacy of certain objections, researchers often attempt to accommodate them in the design of an RCT—for example, by randomizing the first patient, by use of sequential techniques, permitting cases to be their own controls, etc. In making these accommodations and implementing a study in the real (sometimes hostile) world, certain methodological allowances must be made and certain categories of patients or conditions must perhaps be excluded. The researcher here has been forced by circumstances to depart from the ideal textbook design, and this obviously affects the generalizability and reliability of any inferences made. But without these methodological accommodations, the RCT would never have been permitted in the first place. These allowances, which are forced on researchers by practical considerations, are seized upon by critics to discredit the entire RCT. For example, some of the criticism of the eventual design of the first Mather trial of coronary care units versus home care for myocardial infarction came from the very interest groups whose initial objections determined what the original research design could be (Mather et al., 1971). It is analogous to someone saying they will not attend a party unless they can decide who is to be invited, and then complaining after the party that the company left much to be desired!

Sometimes the results of an RCT do show an innovation to be effective and these are immediately seized upon by its proponents and used to advance its career. More often, however, RCTs show innovations to be either ineffective or, at best, no more effective than existing and often much cheaper alternatives. Under these circumstances it is wasteful and perhaps unethical that proper evaluations

of innovations should be postponed until the penultimate stages in their career. Imagine the potential for harm that could be avoided, and the resources saved, if innovations were routinely evaluated during earlier stages in their careers—certainly well before they become "standard procedure."

It is remarkable that the results of sometimes only one RCT so frequently create problems for the medical establishment and elicit defensive responses. Some of these reactions are discussed in the next section. Perhaps they are a measure of the flimsiness of the evidence employed up to this point by certain groups and organizations in support of, and as justification for, "standard procedure."

Up to this point we have distinguished several different types of "data" (but have employed the term very loosely). Stage 1 (promising reports) includes media stories that originate from manufacturing interests, and/or pilot studies on case histories that are essentially worthless but suggestive. A second type of data (clinical experience) was discussed in relation to Stage 2 (professional or organizational adoption). Again, such data do not constitute evidence of an innovation's effectiveness. While describing the elevation of innovations to "standard procedure" in Stage 4 we discussed the proliferation of a third kind of data—those derived from uncontrolled, comparative observational studies. Although these were considered more useful than either of the first two forms, they are still less reliable than a fourth form—the results from RCTs. One ought to be mindful of the different stages in the career of an innovation, and just how advanced that career usually is before data suitable for the formulation of social policies involving the allocation of public resources make their entrance.

6. The Stage of Professional Denunciation

We are concerned here with the defensive reactions that RCTs often elicit from the medical establishment when their findings appear to question what has become standard procedure. Although some of this reaction certainly constitutes legitimate criticism (concerning, for example, methodological, statistical, and ethical issues), much of it is simply a hostile response from a group self-interestedly protecting a domain of activity. Although often disguised in scientific and ethical

jargon, such a response is not unique to medicine, but quite like the response of other groups and interests who perceive their livelihood and/or status threatened. It is understandable that a clinician would be defensive if a lifetime of "clinical experience" (one of the most valuable commodities in medicine) has been found to be simply wrong. Similarly, any researcher can be expected to vigorously defend the observational studies upon which his/her reputation has been built. I know few social scientists who would not be threatened by findings that run contrary to what they and their closest colleagues have espoused over many years. It should surprise no one that, in protecting their own interests and reputation, physicians respond like other human beings!

There are, of course, many ways of discrediting the results of an RCT that challenges some standard procedure. One favorite technique is to severely restrict the application of the results by depicting RCTs as impractical, ivory-tower activities, which seldom help in the handling of everyday problems in the real world. While perhaps conceding that disquieting findings may be applicable to experimental or research situations, clinicians may depict them as inapplicable to the everyday practice of medicine, or to what Lasagna (1974) refers to as "naturalistic" circumstances.

Professional denunciation often takes the form of letters to the editors of professional journals, particularly those that have the temerity to first publish the results of an RCT that question the effectiveness of some standard procedure. By mounting a write-in campaign, some interests can create the impression that there is more opposition than there actually is. At the same time, disquieting findings are often "put into context" by special editorials from "invited experts," who may attempt to reconcile the contradictory findings with their own clinical experience (Sanders, 1973). These experts may also be called upon to constitute a special committee, or working party, to evaluate the results of an RCT that pose enough of a challenge to standard procedure (Joint Working Party, 1975). Attempts are made to shore up an activity, the general standing of which is being threatened by an RCT, by appointing a panel of experts, or an "advisory group," which is expected to review any available evidence and make some recommendation (McKinlay, 1980). Such a committee was appointed in Great Britain after the publication of the first Mather (1971) trial. A federal advisory panel in the United States, headed

by the chairman of the Division of Cardiovascular Diseases at the Mayo Clinic, recently considered the status of coronary artery by-pass surgery and enthusiastically endorsed it (National Institutes of Health, 1981). The claim that such special committees represent a genuine attempt to ascertain the proper place of certain innovations is strained if one analyzes the composition of these groups, and the interests they represent. Furthermore, it must be recalled that, to the extent that they review "data," these groups usually review (with the exception of the RCT) only the (methodologically defective) observational reports.

The issue of double standards is perhaps most evident during this sixth stage involving professional denunciation. The many defective observational studies conducted up to this point seldom receive adequate methodological and statistical scrutiny, whereas RCTs are subjected to the most stringent criticism, employing standards that are almost never invoked during the earlier stages. Questions are raised and motivations challenged that, again, are seldom raised during the earlier stages. There is absolutely *no* objection here to RCTs being subjected to the most gruelling methodological and statistical criticism. In an area such as medicine, with such a clearly documented potential for harm, there cannot be criticism enough. And if an RCT fails to meet even minimal standards of adequacy, it ought to be improved upon or disregarded. What is being questioned here is the near absence of any such criticism during the formative and strategically more important stages in an innovation's career. The nature of the scientific standards invoked, and the force with which they are applied, seem to vary depending on whether or not the results are supportive of what has become standard procedure.

Occasionally, the medical establishment seems simply to disregard the overwhelming evidence against some standard procedure. For example, in 1971, after many cases of vaginal adenosis and some carcinoma among young women whose mothers had received stilbesterol to prevent spontaneous abortion, the FDA issued a directive against the use of the drug for this purpose. Chalmers discusses this in a paper on "The Impact of Controlled Trials on the Practice of Medicine" and suggests:

> Obstetricians cannot be faulted for not anticipating such a long-term side-effect of the therapy. At the time of its use there was

> no known toxicity, but was there any evidence of its efficacy? Between 1946 and 1955 thirteen evaluations of the efficacy of stilbesterol in preventing abortions in women with histories of habitual abortion and in diabetic women were carried out. Seven of the studies resulted in enthusiastic conclusions. None of these seven were controlled trials. Three had no controls, two employed historical controls, and in two studies the controls can be best called contrived, because the data were gathered after the results had originally been reported without any control patients. During this same period of time six studies revealed no evidence for efficacy of stilbesterol. All six of these had simultaneous control patients and three were double blind. In the 1950's randomization was not a commonly used procedure, and only one of the six was a randomized controlled trial.
>
> These data should have been most impressive to practicing physicians. The evidence was overwhelming that stilbesterol had no effect in preventing spontaneous abortion in any group of patients, and six of seven textbooks of obstetrics came to this conclusion during the 1960's. Yet several studies of the marketing of stilbesterol in the late 1960's revealed that roughly 50,000 women received the drug during pregnancy per year, 15 years after six reasonably well-controlled studies showed it to be totally ineffective. The late appearing toxicity was irrelevant. (Chalmers, 1974:754)

It is reasonable then to argue that the success of an innovation has little to do with its intrinsic worth (whether it is measurably effective, as determined by controlled experimentation) but is dependent upon the power of the interests that sponsor and maintain it, despite the absence or inadequacy of empirical support. The power of such interests is also evident in their ability to impede the development of alternative practices (for which there may also be considerable observational support) that could conceivably threaten an activity in which there is already considerable investment.

7. The Stage of Erosion and Discreditation

One often reads reports of or, better still, witnesses an actual performance by some exciting new artist. For a while this artist is a major topic of conversation, receives rave reviews, much public recognition and support. Sometimes the artist is able, largely through good management and regulated exposure, to stay in the public eye

for a period of time, and may eventually be more or less taken for granted. More often, however, he or she quietly disappears into obscurity, leaving the public wondering whatever happened to so-and-so. After such a promising beginning, what did the artist end up doing, and where is he or she now? Many promising careers are embarked upon, but few seem to survive and be established in the big time. The same phenomenon appears to be associated with the diffusion of medical innovations—whatever happened to that much heralded wonder drug, the machine that was supposed to revolutionize the practice of medicine, or the surgical procedure that offered new hope? Schneiderman (1975:68) urges us to "remember the number of drugs that have been Roman candles, making a bright and beautiful flash for a short time and then burning out."

After some period of time (often more than a decade), and ever so gradually, an erosion of support begins to set in. The enthusiastic claims made for the innovation during its earlier stages are modified somewhat; it is not as universally applicable as once thought; it is useful only for certain groups of the population, or particular types of or stages in an illness. Its career can sometimes be propped up for a while if it is combined with some other innovation or it may be relegated to the position of therapy of last resort—something to be employed after all other efforts are exhausted. Sometimes there is a scandal (e.g., thalidomide, dystilbesterol) and the innovation's career is abruptly terminated. And then, of course, people emerge who now claim to have had doubts about it all along! In these relatively rare situations the innovation may become so discredited that it is viewed as unethical to continue with it (e.g., radical mastectomy); it may eventually even be ridiculed. But scandal is not the only reason for a promising career's demise. More often it is simply eclipsed by some other rising star, and just drops out of public view. The innovation no longer enjoys public attention, little prestige is derived from association with it, cheaper alternatives become available, and so forth. Finally, it is relegated to that great dust heap called History. The slow demise of innovations is, indeed, consistent with the medical profession's resistance to change and preference for accepted procedures (McKinlay, 1973). Discreditation or discard usually occurs only when a *replacement* procedure becomes available.

From the above discussion it is evident that the typical career of most medical innovations is extraordinarily wasteful of scarce resources. To paraphrase Hegel, "What experience and the history of

medical innovation teach is this—that the professions, hospitals, and the state never have learned anything from history, or acted on principles deduced from it." Innovation after innovation begins its slow, costly journey through the stages described only to end up—in the overwhelming majority of cases—either discarded or discredited. Is there no way of ensuring that more reliable information is available during earlier formative stages when the innovation is not yet firmly ensconced? Is there no way of temporarily curtailing the understandable, but so often premature, enthusiasm of the professions and medical organizations? Can the public and the state be encouraged to withhold their endorsement until some more reliable objective evaluation has been undertaken? How can some of the resources now consumed by wasteful observational reports be diverted into more reliable controlled experimentation? Is there some way of moving the evaluation of an innovation from the end of its career to some earlier beginning stage? Must we always remain uninformed by earlier experience and end up going through all these stages with every new innovation? Given our present knowledge of the disturbing ratio of careers embarked upon to those that are successful, is there no way of avoiding the endless repetition of our wasteful past (Fineberg and Hiatt, 1979)?

The particular innovations discussed illustrate a more general problem besetting most medical care systems: the repeated adoption of unevaluated innovations. Positron emission tomography, zeumatography, dynamic spatial reconstruction, and the numerous other costly innovations that are now lined up and seeking admission to the House of Medicine are of little intrinsic concern to this author. What is worrisome is the way in which just about all innovations, with active support of the state, slip into the medical care system of most countries without any proper evaluation either before or during most of their careers. The legal maxim that a person is presumed innocent until proven guilty appears also to apply to most medical innovations. They are assumed to be effective until they are shown *ad nauseum* to be ineffective. And on those occasions when something is shown to be ineffective, it is difficult to remove because of the pressure groups associated with, and even dependent upon, its survival. Unlike therapeutic nihilists, this author is certainly not opposed to any and all surgical, diagnostic, or pharmaceutical interventions (McKinlay, 1978). To be anti-innovation is obviously to be antiprogress. At the

same time, we must oppose premature state support for the introduction and proliferation of just about every promising innovation, without any requirement that their effectiveness be properly demonstrated either before or at a very early stage in their careers.

A Strategy for the Future

Present patterns of social expenditure are clearly unrelated to any widely accepted outcome measures, and many practices exist, or are proposed, in the knowledge that they are either ineffective or questionable, or have never been evaluated on proper scientific grounds (McKinlay, 1980). Such a situation was perhaps tolerable, although not acceptable, during the 1960s, when there was considerable economic growth, and when the irrationality and waste described in the first part of this paper did not require subtractions from existing allocations. Most countries are now beset with structural payments imbalances, uncontrollable inflation, deepening recession, ever-increasing unemployment, and negative or negligible economic growth (McKinlay, 1980), which ought to preclude continuation of this wasteful course (McKinlay, 1977a; O'Connor, 1973). In the United States, for example, there are public movements to limit taxation, cities increasingly cannot meet necessary expenses, and major industries are facing doubtful survival, despite heavy government support (Tamaskovic-Devey and McKinlay, 1981).

Since many view the state as responsible for providing the greatest possible benefit to the greatest number of people (utilitarianism), and because the allocation of ever scarcer resources cannot continue as has been described for the past, some clear criteria must be invoked to inform social policy. Any policy based on ad hoc responses to defective data and particular interest groups structurally precludes the state from allocating resources in accordance with utilitarian principles. Although the present irrational distribution of resources does occasionally help meet some social needs, there is no structural mechanism for ensuring that it do so. It is therefore essential that some objective (i.e., interest-free) criterion inform the allocation of at least public resources to medical innovations. And the requirement that there be proper *demonstrations of effectiveness* (preferably before their introduction

and diffusion) is such a criterion (Light, Mosteller, and Winokur, 1971; Gilbert, Light, and Mosteller, 1975). On the basis of this argument, the following premise should, therefore, inform all social policy:

> *Government Should Not Support through Public Funding for General Public Use Any Service, Procedure, or Technology, the Effectiveness of Which Has Not Been, or Cannot Be, Demonstrated.*

This is not an unreasonable premise. The *public* ought to be confident that the services they receive and pay for are effective. *Workers* involved in the production and distribution of human services should wish to know that they are beneficially altering some condition or problem. The *state* ought to be concerned that public funds are not devoted to ineffective or questionably effective activities. Surely, it is irrational and nonutilitarian for the state, as at present, to estimate the possible costs and survey the social acceptability of some service without some previous demonstration of its effectiveness—i.e., its measurable contribution to some beneficial alteration in the natural course of some problem (Cochrane, 1972). This is indisputable.

It is no part of the present argument that ineffective services should necessarily be declared illegal or removed from the human service marketplace (McKinlay, 1978). People should be free to purchase just about any human service desired, no matter how ineffective. (If people want to undergo a coronary artery bypass graft—at a cost of around $20,000—then they should be free to do so). What is questioned here is whether, through public expenditures and/or third-party payments, the rest of society should be required to pay for such prodigal purchases.

There are a number of things that we need to consider urgently if we are to develop a human or social service system that can deal effectively with the problems that now plague our society:

—First, *we must determine, in some objective and independent fashion, the nature and magnitude of the human needs in our society.*

—Second, *all new services must be evaluated objectively (perferably and where appropriate by RCTs) before they are introduced.* There ought to be a moratorium on the widespread adoption and diffusion of any unevaluated innovations.

—Third, and in order to separate the grain from the chaff, *we should begin a systematic evaluation of the technologies that are already ensconced and widely accepted* but have never been properly evaluated.

—Fourth, if one subscribes to the view that society should be responsible for those who are generally recognized to be in need—that human services are an inalienable right—then *the state should be responsible for, and encourage the use of, only the services that have been shown to be effective*—that is, services that are able to beneficially alter the natural course of some recognized problem.

Now if *demonstrated effectiveness* can be generally accepted as the primary criterion for the allocation of resources (a necessary but *not* sufficient condition for the public funding of an intervention), then the methodology for the determination of effectiveness becomes a critical issue in social policy. It has been argued that one can determine the effectiveness of some program, or intervention, only through scientifically sound comparisons. For most people, RCTs represent the methodology of choice because there is no way, other than through experimentation (involving the objective—preferably random—assignment of cases to the intervention being evaluated, or to some appropriate alternative), that the matter of effectiveness can be placed beyond dispute. Consequently, *any social policy that seeks some distribution of public resources on the reasonable basis of effectiveness must also be concerned with the optimal methodology by which effectiveness can be demonstrated.*

Three principal criteria that may be employed to determine whether or not some proposed procedure, service, or technology (hereinafter referred to as an intervention) should be publicly funded, are presented *in order of logical importance.*

1. *Effectiveness. Whatever the intervention, it must first demonstrate some ability to beneficially alter the natural course of a clearly defined condition or set of conditions.*

- It is particularly important to demonstrate a benefit *over and above any possible placebo effect.*
- The demonstration of benefit must be as free as possible from important sources of *bias,* care being taken to minimize or adjust for, and document, any remaining biases.
- In general, the demonstration must involve some *comparison.* The proposed intervention must be demonstrably better than existing

procedures or services (if any) designed for the same purpose, and this improvement must be real, *not* a placebo effect.

- In determining any benefit of some proposed interventions, due account must be taken of any accompanying negative *side effects or added risks.* These must be incorporated in any outcome measure.
- The benefit should be applicable to as *wide a section of the population* subject to the condition as possible. This requires that the demonstration be carried out on as representative a group as is feasible.

2. *Cost Efficiency. Where two or more proposed interventions of approximately equivalent effectiveness are available, that one should be preferred that involves the least cost.*

- The *cost* of some proposed intervention must be *compared* with the cost of any existing procedures.
- The demonstrated effectiveness of the intervention, compared with standard or existing procedures, must be *related to the cost* of the intervention. In general, the costlier the intervention, the greater its demonstrated effectiveness must be *before* it can be considered an acceptable alternative.

3. *Acceptability and Equity. A proposed intervention that is both effective and cost efficient is of no value unless it is socially acceptable and equally accessible to all the relevant subgroups of the society into which it is being introduced.*

- A proposed intervention should be in a form that ensures its utilization by those groups whose needs it is designed to beneficially alter. If, for example, a condition is prevalent among the elderly and the proposed intervention requires a physical fitness usually associated with youth, then such an intervention is clearly inappropriate.
- When two or more interventions appear to be of equivalent effectiveness and cost efficiency, and are both of an appropriate form, some evidence of public preference may be incorporated in the decision-making process.

As stated, these three criteria (effectiveness, cost efficiency, and social acceptability and equality of access) are presented in order of logical importance. Each criterion is considered to be a necessary but not sufficient condition for the inclusion of the next criterion. For example, an intervention must demonstrate some effectiveness that is clearly attributable to the intervention, and *not* to a placebo, before any considerations of cost can be logically included in the decision procedure. At the same time, if effectiveness is demonstrated, it still may not be of sufficient dimension to justify possible increased cost of the proposed intervention.

In the context of the above formulation, the role of user preference or client satisfaction is clearly a minor one, and is important only if two or more interventions display equivalent effectiveness, cost efficiency, and appropriateness of form. One could argue that, given two equally effective interventions, one of which is more costly but preferable, the more costly should be chosen, as the less costly alternative will not be used (it is not preferred). However, it may be more efficient to render the less costly alternative more socially acceptable than to publicly fund the preferred, costlier intervention. One simply cannot evaluate the effectiveness (as defined in this paper) of an intervention by surveying the extent of consumer satisfaction. The public may be dissatisfied with services that are effective and of technically high quality, and satisfied with services that are ineffective and of poor technical quality.

Because *effectiveness* must be the key criterion in any consideration of a proposed innovation for public funding, the nature of the evidence for such effectiveness is of prime importance. The evidence submitted should be of such a form that: a) the worth of the evidence can itself be evaluated; and b) inferences and generalizations can be made from this evidence to the population to be exposed to the innovation. This evidence should consist, minimally, of carefully designed and executed studies, in clearly defined populations, using objective or reproducible outcome measures. Most studies should be of the form of experiments used to compare types of interventions, or surveys to estimate population characteristics.

With few exceptions, acceptable evidence of an intervention's effectiveness can be established only through comparative experiments (RCTs) that are as free as possible of sources of bias. Then and only then can there be confidence that the observed effect (if there is one)

is *actually the result of the intervention.* Such experiments (trials) must minimally include the characteristics described by McKinlay (1981) (see also Chalmers et al., 1981; Nyberg, 1974; Wulff, 1977; Lionel and Hexelheimer, 1970).

Two caveats must be emphasized regarding any implementation of the decision procedure that is proposed above. First, there is *no* suggestion that the criteria of effectiveness, cost efficiency, social acceptability, and equality of access should be applied only to innovations newly proposed for public funding. Clearly, innovations already ensconced in our publicly funded human services system must be subjected to the same scrutiny. Moreover, it is likely that a large proportion of standard procedures or services would not meet even the minimally sufficient criteria and should, therefore, be excluded from further public funding if we are ever to receive value for money in human services.

Second, there is *no* suggestion that the criteria proposed should be applied only to particular innovations, or to those proposed by particular groups. Any intervention proposed for public support or funding (whether acupuncture, cardiothoracic surgery, chiropractic, income maintenance, psychiatry, social work, or transcendental meditation) should be subject to the same basic criteria. A situation must be avoided where, as at present, double standards exist regarding the criteria to be met, depending on the relative power of interested groups proposing or supporting some promising innovation.

References

Banta, H.D. 1980. The Diffusion of the Computed Tomography (CT) Scanner in the United States. *International Journal of Health Services* 10:251–269.

———, and Thacker, S.B. 1979. *Costs and Benefits of Electronic Fetal Monitoring: A Review of the Literature.* DHEW Publication No. (PHS) 79-3245.

Berk, A.A., and Chalmers, T.C. 1981. Cost and Efficiency of the Substitution of Ambulatory for Inpatient Care. *New England Journal of Medicine* 304:393–397.

Bloom, B.S., and Peterson, O. 1973. End Results, Costs and Productivity of Coronary-Care Units. *New England Journal of Medicine* 288:72–78.

Boston Globe, 1980. Doctor Cites Gains against Liver Cancers. October 7, 1980.

Braunwald, E. 1977. Coronary Artery Surgery at the Crossroads. *New England Journal of Medicine* 276:661–663.

Bunker, J.P., Barnes, B.A., and Mosteller, F., eds. 1977. *Costs, Risks and Benefits of Surgery.* New York: Oxford University Press.

Byar, D.P., et al. 1976. Randomized Clinical Trials. *New England Journal of Medicine* 295:74–80.

Chalmers, T.C. 1972. Randomization and Coronary Artery Surgery. *Annals of Thoracic Surgery* 14:323–327.

———. 1974. The Impact of Controlled Trials on the Practice of Medicine. *Mount Sinai Journal of Medicine* 41:753–759.

———. 1975a. Ethical Aspects of Clinical Trials. *American Journal of Ophthalmology* 79:753–758.

———. 1975b. Randomization of the First Patient. *Medical Clinics of North America* 59:1035–1038.

———, Block, J.B., and Lee, S. 1972. Controlled Studies in Clinical Cancer Research. *New England Journal of Medicine* 287:75–78.

———, Sebestyen, C.S., and Lee, S. 1970. Emergency Surgical Treatment of Bleeding Peptic Ulcer: An Analysis of the Published Data on 21,130 Patients. *Trans American Clinical Climatology Association* 82:188.

———, Smith, H., Blackburn, B., et al. 1981. A Method for Assessing the Quality of a Randomized Control Trial. *Controlled Clinical Trials.* In press.

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

———. 1979. 1931–1971: A Critical Review with Particular Reference to the Medical Profession. *Journal of the Royal College of Physicians of London* 14:2–12.

Coleman, J.S., Katz, E., and Menzel, H. 1966. *Medical Innovation: A Diffusion Study.* Indianapolis: Bobbs-Merrill.

Conner, R.F. 1977. Selecting a Control Group: An Analysis of the Randomization Process in Twelve Social Reform Programs. *Evaluation Quarterly* 1:195–243.

Cromwell, J., Ginsberg, P., Hamilton, D., and Summer, M. 1975. *Incentives and Decisions underlying Hospitals' Adoption of Major Capital Equipment.* Cambridge, Mass.: Abt Associates.

Fineberg, H.V., and Hiatt, H.H. 1979. Evaluation of Medical Practices: The Case for Technology Assessment. *New England Journal of Medicine* 301:1086–1091.

———, Bauman, R., and Sosman, M. 1977. Computerized Cranial

Tomography: Effect on Diagnostic and Therapeutic Plane. *Journal of the American Medical Association* 238:224–227.

Freidson, E. 1971. *Profession of Medicine.* New York: Dodd, Mead.

Gilbert, J.P., Light, R.J., and Mosteller, F. 1975. Pressing Social Innovations: An Empirical Base for Policy. In Bennett, C.A., and Lumsdaine, A., eds., *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* 39–193. New York: Academic Press.

Gordon, G., and Fisher, G.L., 1975. *The Diffusion of Medical Technology.* Cambridge, Mass.: Ballinger.

Gore, S.M., Jones, I.G., and Rutter, E.C. 1977. Misuse of Statistical Methods: Critical Assessment of Articles in B.M.J. from January to March 1976. *British Medical Journal* 1:85–87.

Guttentag, M. 1971. Subjectivity and Its Use in Evaluation Research. *Evaluation* 1:60–65.

Hemminki, E. 1980. Study of Information Submitted by Drug Companies to Licensing Authorities. *British Medical Journal* 280:833–841.

———, and Falkum, E. 1980. Psychotropic Drug Registration in the Scandinavian Countries: The Role of Clinical Trials. *Social Science and Medicine* 14:547–559.

Hill, J.D., Hampton, J.R., and Mitchell, J.R.A. 1978. A Randomized Trial of Home-versus-Hospital Management for Patients with Suspected Myocardial Infarction. *Lancet* 1:837–841.

Joint Working Party of the Royal College of Physicians of London and the British Cardiac Society. 1975. *Journal of the Royal College of Physicians of London* 10:5.

Kaluzny, A.D., and Veney, J.E. 1973. Attributes of Health Services as Factors in Program Implementation. *Journal of Health and Social Behavior* 14:124–133.

———, Veney, J.E., and Gentry, J.T. 1974. Innovation in Health Services: A Comparative Study of Hospitals and Health Departments. *Milbank Memorial Fund Quarterly/Health and Society* 52 (Winter):51–82.

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

Lasagna, L. 1974. A Plea for the "Naturalistic" Study of Medicine. *European Journal of Clinical Pharmacology* 7:153.

Light, R.J., Mosteller, F., and Winokur, H.S. 1971. Using Controlled Field Studies to Improve Public Policy. *Federal Statistics* 2:367–402.

Lionel, N.D.W., and Hexelheimer, A. 1970. Assessing Reports of Therapeutic Trials. *British Medical Journal* 3:637–640.

Mather, H.G., et al. 1971. Acute Myocardial Infarction: Home and Hospital Treatment. *British Medical Journal* 3:334–335.

———. 1976. Myocardial Infarction: A Comparison between Home and Hospital Care for Patients. *British Medical Journal* 1:925–929.

Mathur, V.S., et al. 1975. Surgical Treatment for Stable Angina Pectoris: Prospective Randomized Study. *New England Journal of Medicine* 292:709–713.

McKinlay, J.B. 1973. On the Professional Regulation of Change. In Halmos, P., ed., *Professionalization and Social Change,* 61–84. Sociological Review Monograph No. 20.

———. 1977a. On the Medical-Industrial Complex. *American Medical News* (April 11):26.

———. 1977b. The Business of Good Doctoring, or Doctoring as Good Business: Reflections on Freidson's View of the Medical Game. *International Journal of Health Services* 7:459–488.

———. 1978. The Limits of Human Services. *Social Policy* (Jan.–Feb.):29–36.

———. 1980. Evaluating Medical Technology in the Context of a Fiscal Crisis: The Case of New Zealand. *Milbank Memorial Fund Quarterly/Health and Society* 58(Spring):217–267.

———, and Dutton, D.B. 1974. Social-Psychological Factors Affecting Health Service Utilization. In Mushkin, S.J., ed., *Consumer Incentives for Health Care,* 251–303. New York: Prodist.

McKinlay, S.M. 1975. The Design and Analysis of the Observational Study: A Review. *Journal of the American Statistical Association* 70:351, 503–523.

———. 1981. Experimentation in Human Populations. *Milbank Memorial Fund Quarterly/Health and Society* 59 (Summer):308–323.

Miller, R.R. 1975. Prescribing Habits of Physicians. *Drug Intelligence and Clinical Pharmacology* 7:492–500, 557–564.

———. 1976. Prescribing Habits of Physicians. *Drug Intelligence and Clinical Pharmacology* 8:81–91.

Mundth, E.D., and Austen, W.G. 1975. Surgical Measures for Coronary Heart Disease. *New England Journal of Medicine* 293:13–19, 75–80, 124–130.

Murphy, M.L., et al. 1977. Treatment of Chronic Stable Angina. *New England Journal of Medicine* 297:12.

National Institutes of Health, 1981. Coronary-Artery Bypass Surgery: Scientific and Clinical Aspects. *New England Journal of Medicine* 304:680–684.

Newsweek. 1980. Scanning the Human Mind. September 29.

New York Times. 1980a. 3rd Dimension of Organs Seen on New X-Ray. October 12.

———. 1980b. Findings Are in Conflict on Value of Coronary Bypass Operations. November 18.

———. 1980c. U.S. Study Backs Bypass Surgery. December 6.

Nyberg, G. 1974. Assessment of Papers of Clinical Trials. *The Medical Journal of Australia:*381.

O'Conner, J. 1973. *The Fiscal Crisis of the State.* New York: St. Martin's Press.

Peto, R., 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.

———. 1977. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. II. Analysis and Examples. *British Journal of Cancer* 35:1–39.

Rappaport, J. 1978. Diffusion of Technological Innovation among Non-profit Firms: A Case Study of Radioisotopes in U.S. Hospitals. *Journal of Economic Business* 30:108–118.

Riecken, H.S., and Boruch, R.F., eds. 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* New York: Seminar Press.

Russell, L.B. 1976. The Diffusion of New Hospital Technologies in the United States. *International Journal of Health Services* 6:557–580.

———. 1978. *Technology in Hospitals: Medical Advances and Their Diffusion.* Washington, D.C.: Brookings Institution.

———, and Burke, U.S. 1975. *Technological Diffusion in the Hospital Sector.* Washington, D.C.: National Planning Association.

Sanders, C.A. 1973. The Coronary-Care Unit: Necessity or Luxury? *New England Journal of Medicine* 288:101–102.

Schneiderman, M.A. 1975. How Do You Know You've Done Any Better? *Cancer* 35:64–69.

Sheldon, W.C., et al. 1975. Surgical Treatment of Coronary Artery Disease: Pure Graft Operations, with a Study of 741 Patients Followed 3–7 Years. *Progress in Cardiovascular Diseases* 18:237–253.

Silverman, M., and Lee, P.R. 1974. *Pills, Profits and Politics.* Los Angeles: University of California Press.

Sontag, S. 1978. *Illness as Metaphor.* New York: Farrar, Straus and Giroux.

Spodick, D.H. 1973. The Surgical Mystique and the Double Standard. *American Heart Journal* 85:579–583.

Stross, J.K., and Harlan, W.R. 1979. The Dissemination of New Medical Information. *Journal of the American Medical Association* 241:2622–2624.

Strupp, H.H., Hadley, S.W., and Gomes-Schwartz, B. 1977. *Psychotherapy for Better or Worse.* New York: Jason Aronson, Inc.

The Compact Edition of the Oxford English Dictionary. 1971. New York: Oxford University Press.

Tomaskovic-Devey, D., and McKinlay, J.B. 1981. Bailing Out the Banks: The United States and Private International Debt. *Social Policy* 11:8–17.

U.S. Senate. 1974. Subcommittee on Health, Committee on Labor and Public Welfare. *Examination of the Pharmaceutical Industry, 1973–74.* Washington, D.C.: Government Printing Office.

Vineberg, A. 1975. Evidence That Revascularization by Ventricular-Internal Mammary Artery Implants Increases Longevity: Twenty-four Year, Nine Month Follow-up. *Journal of Thoracic and Cardiovascular Surgery* 70:381–394.

Waldron, I. 1977. Increased Prescribing of Valium, Librium, and Other Drugs: An Example of the Influence of Economic and Social Factors on the Practice of Medicine. *International Journal of Health Services* 7:37–62.

Warner, K.E. 1977. Treatment Decision Making in Catastrophic Illness. *Medical Care* 15:19–33.

Webster's New Collegiate Dictionary. 1976. Springfield, Mass.: Merriam.

Wulff, H.R. 1977. Check List for Assessment of Controlled Therapeutic Trials. *Acta Neurologica Scandinavia*, Supplement, 60:79–80.

Acknowledgments: Much of the background reading for this paper was facilitated by a 1976 grant from the Milbank Memorial Fund. The paper has benefited from discussions with my colleagues Professors Mark Field, Sol Levine, S.M. Miller, and Dr. Sonja McKinlay.

Address correspondence to: Dr. John B. McKinlay, 59 Princeton Road, Chestnut Hill, MA 02167.