Technical Reports

The expectation and the reality

Health Program Evaluation in Relation to Health Programing

ROBERT M. THORNER, B.S., M.B.A., M.P.H.

GOVERNMENT at all levels is expanding its contribution to the payment of health services and medical care costs, both in absolute terms and relative to the total investment in these services (*1*). With this rising expenditure has come an increasing demand for the rationalization of the expenditure, manifested by greater emphasis on program planning and evaluation. Organizational entities concerned with planning and evaluation exist at almost every echelon of the Federal health establishment, in most State and local units, and in Regional Medical Programs.

A section on evaluation has become the sine qua non of every grant application for the support of projects for the development of health or medical care services. Frequently, there is a requirement by grant review groups that evaluation must be "end results evaluation" and go beyond mere counts of activities, services rendered, or the achievement of specified goals to measure the effect of the program or project on the improvement of the health in the community. It is implied, also, that in a rational system the planning of health services will be based on sound evaluative data, with priority given to the programs, projects, and techniques shown to be most effective in improving health.

To persons familiar with health projects, the large gap between the statement of intent to evaluate which appears in all grant applications and the delivered evaluations is quite obvious, and the feed-in of evaluations to the planning process is tenuous. It is my intention in this paper to examine some of the difficulties of concept and technique inherent in the evaluative process, to analyze the relationship of evaluation to planning, and to propose a practical policy in regard to evaluation for use by grant review bodies.

The Hypotheses

In concept, most health and medical care programs are predicated on the assumption that the level of health is related to the level of effort expended in providing health services, although the exact nature and strength of this relationship are usually poorly established. It is possible to express the relationship using symbolic language in the following simplified model.

$$H = f\,Hs$$

That is, health (H) is a function (f) of the effort expended on health services (Hs).

In the literature of health program evaluation,

Mr. Thorner is a public health adviser in the National Center for Health Services Research and Development, Health Services and Mental Health Administration, Public Health Service. Tearsheet requests to Robert M. Thorner, 5600 Fishers Lane, Rockville, Md. 20852.

Hs is usually referred to as "program activities," and *H* as "program effectiveness" or "end result" (*2*). Persons familiar with the literature of evaluation of medical care will recognize *Hs* as the "process" of medical care and *H* as the "outcome" (*3, 4*).

Program evaluation usually becomes of interest when a change, either technological, in the level of effort, or in organizational structure, is introduced into the current situation. It is assumed that a change in the level of health (ΔH) will result from a change in the level of effort or kind of health service (ΔHs). The model then may be expressed in the following equation.

$$\triangle H = f \triangle Hs$$

Accepting this rather simplistic model for the moment, evaluation then becomes the process of quantifying the change in the health service ($\triangle Hs$) associated with the change in health ($\triangle H$) and thus developing some quantitative measure of the function (f). Expressed in another way, the health services are viewed as an independent variable, health is perceived as a dependent variable, and it is desired to know whether (*a*) inputs of *Hs* result in increased outputs of *H*, (*b*) the relationship is directly proportional (linear), or (*c*) there is an increasing or decreasing increment of units of *H* with additional (marginal) inputs of units of *Hs*. Presumably, decisions are made on the basis of the strength of the relationship (f). When the evaluator attempts to use the model, he is immediately faced with the practical problems of finding appropriate quantitative measures for the inputs of health services or medical care and for the output of health.

Comparative Methods of Evaluation

Only in rare instances, such as vaccination programs, are truly quantifiable measures available, and the evaluator must usually abandon any attempt to quantify the inputs of health services or medical care, contenting himself with merely knowing that the input has changed. The evaluator concentrates on the problem of measuring health at two points in time, before or early in the process of introducing a change and at some later point when sufficient time has elapsed for the change to be presumed to have had an effect. In essence the evaluator has abandoned any attempt to define the function or to relate marginal quantities of input to marginal quantities of output. He has thus reduced his goal to attempting to see whether the particular program has had any effect on health which may logically be imputed to the "process" or "activity." An exception to this procedure is cost benefit analysis when the evaluator usually attempts to quantify the input (at least in total) in monetary terms.

When the evaluator turns to quantifying *H*, he faces the problem that there are also at present no practical quantitative measures of health, despite numerous attempts to develop such a measure (*5*). In essence, three alternative approaches to this problem have been developed by persons interested in evaluation. These approaches may be summarized as follows.

1. Study only the *Hs* element (process or activities).
2. Substitute intermediate objectives for *H*.
3. Substitute negative measures for *H*.

The first technique is most frequently used in evaluative studies in medical care and is also applied to health programs. Basically, outcome or end result is either assumed to be beneficial, or at least desirable, and assumed to result from appropriate process or activity. Evaluation then becomes a matter of measuring the quality or the appropriateness or both. Such measurements usually entail establishing standards of performance and measuring or judging adherence and conformity to the standard. Donabedian has classified standards as either empirical (derived from actual practice) on normative (derived from sources that set standards of knowledge and practice in the dominant medical care system) (*3*).

One may ask, "How valid is the approach and how acceptable is this procedure to health and medical professions?" Shapiro has pointed out that the validity of this "indirect approach" rests on the degree to which end results or outcome has been linked to process or activity in earlier studies (*4*). Sometimes "clinical" experience serves as a substitute for formal studies. Acceptability rests on the degree to which concerned professionals are willing to accept such studies or experiences. For example, most health professionals would be willing to accept properly conducted programs for environmental sanitation as beneficial to the health of the public. There is less acceptance of the benefits of most attempts to restructure medical care services, such as group practice or regionalization.

It should be pointed out that if one is willing to accept the process or activity as one contributing to health, process evaluation does have a positive

value as a quality control measure, and efficiency can properly be studied in this context. Studies can also be designed to analyze the process to identify elements which appear essential to a successful program.

The second proposed method of handling the problem of quantifying health is to substitute intermediate objectives or goals which are more amenable to measurement. A criterion frequently used in rehabilitation is measurement of ability to function. This method involves use of a scale, profile, or index measurement of the activities of daily living (*6*). A second set of goals frequently substituted for the goal of health is related chiefly to the evaluation of health education efforts. These are improvements of health knowledge, health attitudes, and health behavior. In early disease detection, the intermediate goal of finding asymptomatic, previously unknown cases is a frequently used device.

How valid is this approach and how acceptable is it to the health and medical professions? The problem is identical to that of studying process or activity. When previous research or experience has established a link between the intermediate goal and the broad concept of health, this approach to evaluation has validity and acceptance.

In rehabilitation studies, most professionals are willing to accept the measurement of function as an indicator of health for the type of population being studied. It is apparent that "health" is probably not an absolute value but varies with populations, time, and place.

Health knowledge-attitude-behavior can be viewed as a hierarchial order. That is, it is assumed that a change in health knowledge will lead to a change in health attitude and thus to a change in health behavior. More recently health workers have come to believe less in the logic of this process, based on research results now becoming available, and the acceptability of health knowledge and health attitudes as intermediate measures of health has consequently diminished (*7*).

The acceptability of health behavior as an intermediate goal substitute for "health" again depends on how willing professionals are to accept the particular behavior as a measure of health. Most persons would be willing to accept cessation of smoking as very strongly related to health because of the studies linking smoking to disability and death. A much lower acceptance might be expected for increased physical activity or reduction of an intake of saturated fats.

Much of the current controversy over multiphasic screening and periodic health examinations can be attributed to the lack of complete acceptance by the medical profession of the thesis that a positive contribution to health is made by the detection of early or asymptomatic chronic disease, and an almost total lack of evidence linking this intermediate objective to the "health" of patients (*8, 9*). It is also important to note that acceptance of the health value of many of these activities by the general public is quite different from professional acceptance. This variant is relevant to the relationship of evaluation to the decision process, which will be discussed later.

The third major approach to the problem of quantifying "health" and the concept most commonly used in evaluative studies is to substitute a negative measure for *H*, and to look for an inverse relationship to *Hs*. By far the most common inverse measure is mortality, undoubtedly because death represents a discrete event, easily measured and in our society well recorded. Death is viewed, in general, as the antithesis of health, and therefore the perfect inverse measure. However, when one turns to specific causes, the measure becomes less reliable, and for diseases which usually are not fatal, the measure becomes inappropriate or insensitive.

Another common inverse measure of health is morbidity. Morbidity data is less definitive, usually less readily available, and its reliability varies with the specific disease. Morbidity is, of course, strongly negatively related to health and generally accepted as an inverse measure.

Disability is also a commonly used negative substitute for health. Disability can be measured by such discrete events as confinement to home, to bed, or in hospital; absence from work; or inability to perform usual activities (*10*). It should be noted that disability is similar to the previously mentioned intermediate goal measure, "activities of daily living," disability being the obverse of the same coin. Also, the meaning of disability is not independent of occupational or social group and culture; that is a permanent leg injury is not as disabling to a professor as to a carpenter or steeplejack.

Of the three methods of coping with the problem of measuring "health" in evaluative studies, the substitution of these negative measures is probably the best accepted and certainly the most

widely used. It is worth noting that the "benefit" measure most frequently used in cost-benefit analysis is a measure of the present value of productive years of life "saved" or disability reduced by a particular program activity and derives directly from mortality and disability data (*11*).

The Realities

At this point it would appear worthwhile to consider a more complex evaluative model which more closely approximates conditions of the real world. Obviously the outcome of medical care or the effectiveness of a public health program depends also on factors other than the process of care or the activities of the program.

Among the more easily identifiable factors affecting the outcome of health or medical care programs are the educational level of the population (*Ed*), the economic level (*Ec*), the physical environment (*Ev*), and the political situation (*P*). Undoubtedly there are many other factors, some not identifiable at present. A more complex equation is therefore in order.

$$\Delta H = f(\Delta Hs \pm \Delta Ed \pm \Delta Ec \pm \Delta Ev \pm \Delta P \pm \Delta ?)$$

The problem of evaluating health or medical care programs thus is seen not only as a problem of measuring the input of the health service and the output in terms of health, but somehow isolating the effects of the extraneous terms.

Four methods of approach to this problem will be discussed.

1. Multivariate analysis
2. Controlled experiments
3. Trend analysis
4. Comparisons with appropriate similar experiences or standards.

There is no doubt that multivariate analysis would prove most valuable if it were possible. It would be nice to examine all of the elements of change in a multivariate model and to note the strength of the effect of each element on the level of health.

However, it is difficult enough to define and measure *H* and *Hs*, let alone all of the other factors. For this reason a model of this type is impractical. The approach is analogous to the approach of operations research or systems analysis. Attempts to use the systems or multivariate analysis in a combined approach to evaluation and planning have been undertaken with some success by Emlet, Galliher, and Krystynak (*12–14*). Their models generally have dealt with only cost and health service variables.

Controlled experiment, with random allocation of patients to study and control groups, is a method frequently advocated to eliminate the effects of extraneous variables (*15*). Controlled experiment is the classic method of laboratory experiment and clinical drug trials, and this method has, at times, been proposed for the evaluation of intensive care units, cardiac ambulances, multiphasic screening, and similar activities. However, in actual practice, great difficulties are encountered with this design.

When a program or a mode of treatment appears to be superior to another, it may not be possible to get the persons involved in the program to accept random allocation. Also, if the study involves changing a form of therapy which has been accepted by the medical community and become established practice, withholding treatment or care may be unethical or at least legally dangerous. When patients are entitled to a treatment or service as a matter of law or by virtue of prepayment, it may be impossible to prevent substantial numbers of "crossovers" from the experimental to the control groups.

Experiments, other than drug trials, can seldom be carried out in a double-blind design, and there is a tendency for contamination in handling experimental and control groups and for bias in evaluating results. It is unrealistic to expect that within the same organizational framework patients in one group will be handled differently from patients in another group, without some contamination, and especially if the same persons are handling both the experimental and control groups. Also, since it is generally impossible to keep from the persons assessing final results the knowledge of the group (experimental or control) to which patients belong, it is impossible to avoid bias in assessing final results.

Another difficulty in this sort of experiment, especially in studies of medical care, is that differences in results may be related more to the quality or personality of the personnel involved in the particular program than to actual differences in the method of the program, and results may not have general applicability. Also, study of many problems prospectively involves long periods of observation, and such studies are often plagued by loss of subjects, staff turnover, obsolescence of the technique or program being evaluated, and problems of obtaining large amounts of money for

extended periods. For the aforementioned reasons, studies based on this design have not been used as frequently as would seem desirable in the evaluation of health and medical care programs.

Another method used in evaluating health programs and medical care is trend analysis. This method is especially favored when time series based on routine counts of cases, hospitalization or other activities, or deaths have been produced in a period before a change is introduced, and the effect of the change on the series can be noted. This method has the advantage of being relatively inexpensive, since the data are frequently available, or because creating a reporting system is generally less costly than a clinical trial.

The validity of this approach rests on an assumption of ceteris paribus, and the analysis may not be valid if factors other than the change being studied have affected the trend. This method would probably be more widely applied if the United States had a comprehensive health data system which routinely supplied mortality data, hospital and outpatient utilization data, health facility data, and health manpower data which could be related easily to population data in small areas. The Bureau of the Census is currently experimenting with small area data methodology, including health data (*16, 17*).

Finally, comparisons with data from groups which had experiences similar to the experimental group, except for the change introduced in the program being studied, is a method often used. Thus the experience of the study group may be compared with a life table, against a standardized mortality, with a similar disease, with another geographic location, or another time period. This method will be valid to the extent that the comparison experience is similar to the experimental experience except for the introduced change. Purists will rarely accept this method as a substitute for the controlled experiment.

Assessing Expectations from Evaluation

At this point it may be worthwhile to assess what may be expected from evaluative studies. In my opinion, changes in health or medical care programs, except in some instances where a major technological or organizational change is introduced, are unlikely to produce effects on health which are easily measurable.

Stated in another way, the effects of the other variables in the model and measurement errors are usually stronger or as strong as the effects produced by the change in the health or medical care program, and the difficulty of controlling these variables and errors may vitiate any attempt to measure the program's effects, whatever method is used. On the other hand, when a major technological change is introduced, its effect may (though not necessarily) be detectable by any method, even uncontrolled "clinical" observation.

Consider figure 1 which represents data on infant mortality in Sweden for the years 1751–1960. (These data were supplied in a letter dated September 5, 1968, from Dr. Ragnar Berfenstam, Institute of Social Medicine, Uppsala University, Uppsala, Sweden.) It appears that the reduction in infant mortality during this period is not as strongly related to the expansion of health services as to economic and environmental factors, although the rate of decline is more rapid toward the latter part of the timespan. Some acceleration appears to be attributable to the introduction of compulsory smallpox vaccination in 1816, a major technological change. Similar reductions unrelated to health programs have also been observed for many infectious diseases.

When the Salk poliomyelitis vaccine was to be introduced in the United States, a major clinical trial was undertaken to evaluate its effectiveness. This trial was undertaken because it was believed that if the vaccine were introduced without such a trial, its effectiveness would never be known. The outcome of the trial was favorable and the vaccine was introduced.

In retrospect (fig. 2) it is apparent that case trend data, despite the many weaknesses and problems of morbidity reporting, were sufficient to establish the vaccine as effective because of the strength of this new technology (*18*). (This example is not meant to deprecate the decision to hold this trial which also had as an objective quantification of effectiveness, but to indicate the relationship of strength of change to the feasibility of making an evaluation.) A more current example is that a controlled trial of the effect of kidney dialysis on mortality is hardly necessary.

Evaluation and Decision Making

Finally, let us examine the relationship of program evaluation to the decision making process. Why does the relationship of evaluation to decision making seem to be so tenuous, a situation

which many persons engaged in evaluative studies find extremely exasperating? There are, of course, no universally accepted reasons, but several possibilities appear to be worth examination.

First, and obviously, since problems tend to assume importance and require decisions suddenly and a long time is required for evaluative studies, evaluative data frequently are not available when decisions must be made. The program decisions cannot wait, so the decisions are made on the basis of logic or intuition rather than evidence.

Also, advocates of change, who are missionaries for a program for whatever reasons, have found that an emotional appeal is frequently more effective than one based on facts. La Piere has provided the following discussion of the role of the advocate in effecting acceptance of innovation (*19*).

> Since, as had been observed, the members of a society seldom evaluate their social elements in terms of their functional effectiveness, there is apparently little appeal in the claim that a new element will function more effectively than the old. Specious claims have tended, therefore, to be more impressive than valid ones and have been widely used by advocates in their efforts to secure abandonment of the old and acceptance of the new. Often the claim has been the grandiose one that the innovation will cure all—that the new, whatever it may be, will solve all problems and bring the acceptor health, wealth and happiness, and whatever else it is that he may think he wants and does not have in sufficient degree. Cure-all claims have perhaps been most characteristic in the health field; and at one time or another almost everything has been offered as the certain way to general good health—mineral baths, raw foods, fresh air, nudity, this and that surgical procedure, and, of course, countless medications.

The promotion of health programs has been affected by this same syndrome, and programs are as easily "sold" with great promises as with proof of effectiveness. Most frequently the technique used is to create a demand for the program on the part of the general public, which then forces professional acceptance, often grudgingly.

Figure 1. Infant mortality in Sweden, 1751-1960

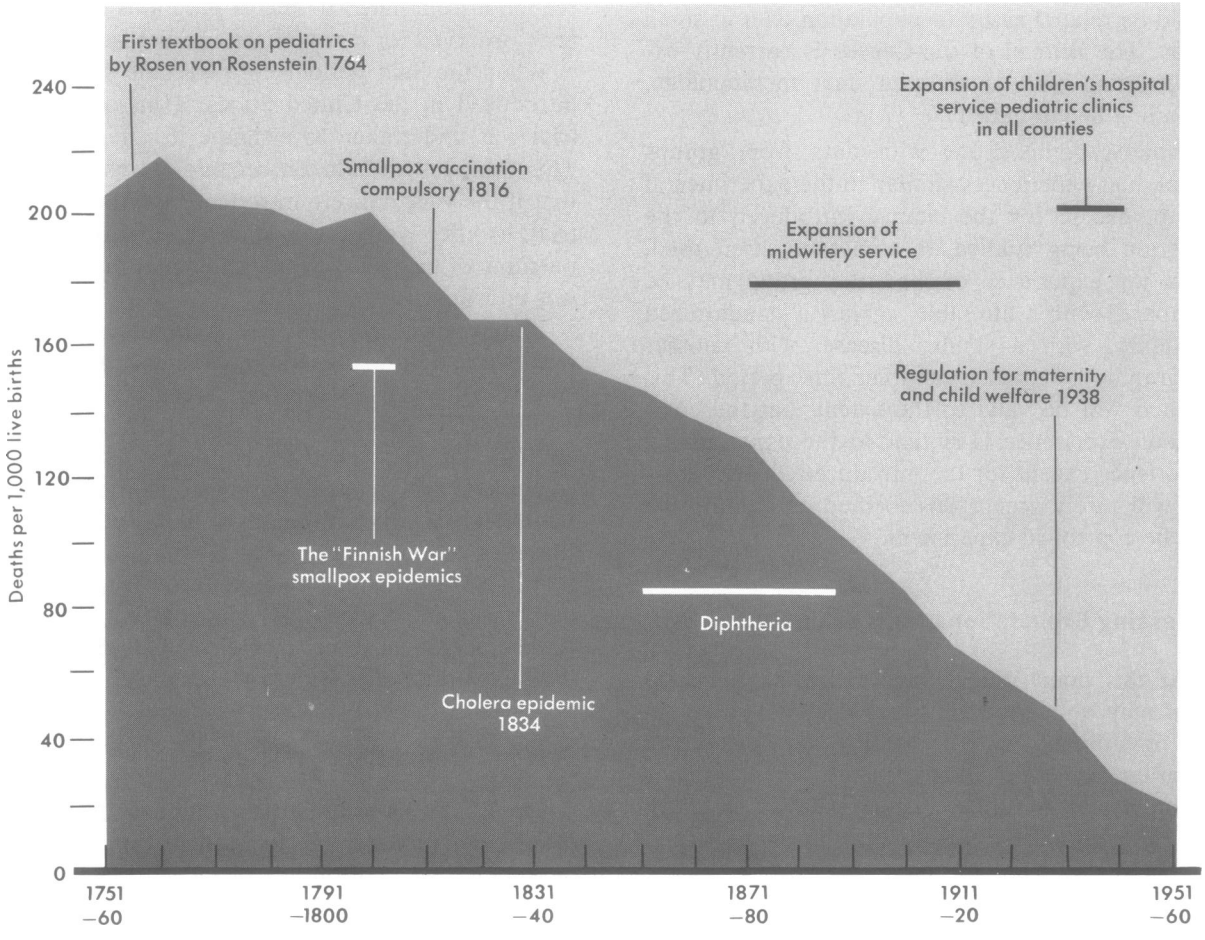
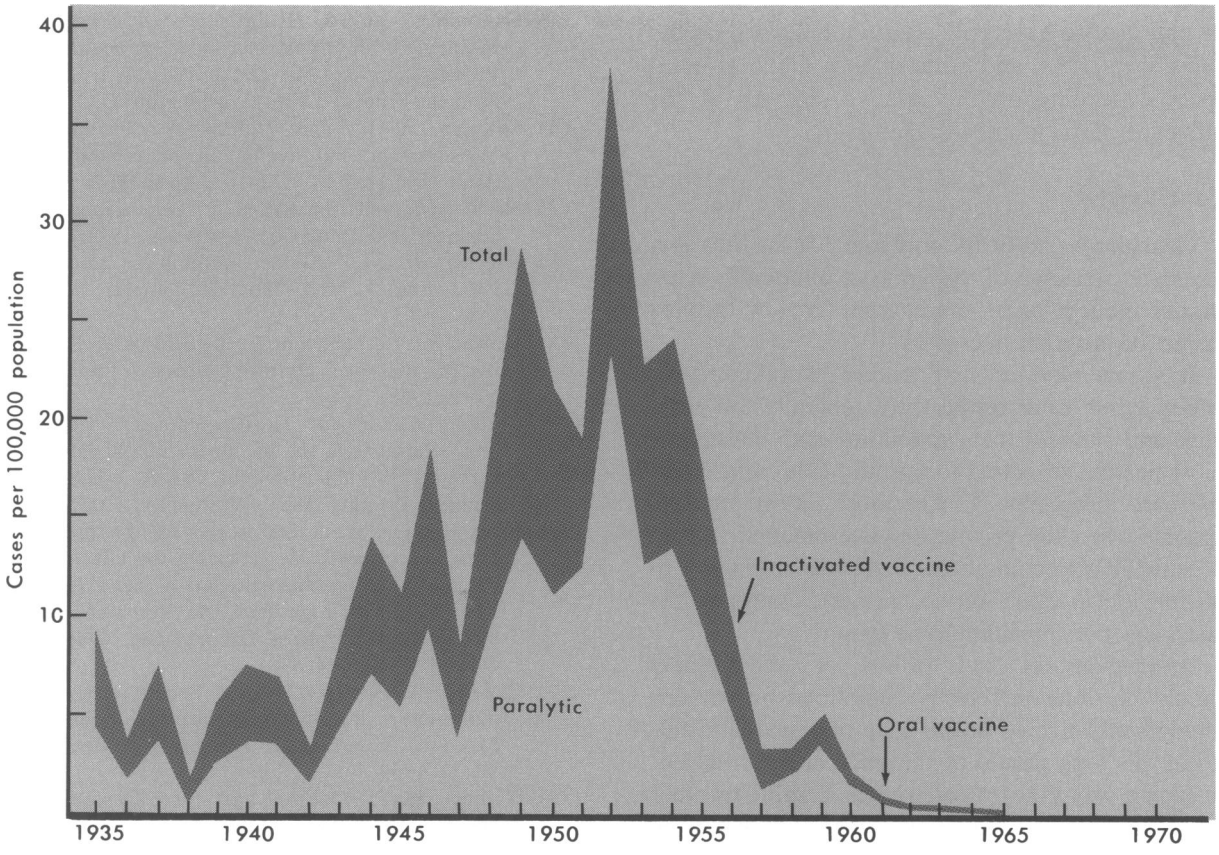


Figure 2. Annual incidence rates of poliomyelitis, United States, 1935-65



Secondary Goals

Still another facet of this problem which warrants exploration is that the true purpose of a health program may be only peripherally related to health. For example, Elinson and Herr have pointed out that the neighborhood health center movement is largely a political and social reform movement (*20*). They list the following among the "latent" objectives of these programs.

1. Improving the image of the black male in poverty communities
2. Stimulating and maintaining solidarity among migrant Chicano farmworkers
3. Pacification of hostile communities by colonial powers
4. Discharging missionary service obligations of the medical-hospital establishment
5. Filling a political void in social and economic action
6. Politicization or radicalization of youth.

Finally, and not necessarily inconsistent with the objective of improving health, a program may be undertaken to save public funds. This is a rather ancient and all pervasive purpose of most publicly financed programs. An early 19th century example from Great Britain will illustrate this point. The following is quoted from the Report of the Poor Law Commissioners of 1838 (*21*).

> All epidemics, and all infectious diseases, are attended with charges, immediate and ultimate, on the poor rates. Laborers are suddenly thrown, by infectious diseases, into a state of destitution, for which immediate relief must be given. In the case of death the widow and the children are thrown as paupers on the parish. The amount of burthens thus produced is frequently so great as to render it good economy on the part of the administrators of the Poor Law to incur the charges for prevailing the evils where they are ascribable to physical causes.

Undoubtedly other objectives not identified in this paper underlie the decisions to undertake health programs, and other factors and tactics used in the decision process are also not identified.

However, it is obvious that the decisions are made in the political arena, in a broad social and economic context, influenced by unstated objectives often unrelated directly to health. These program objectives, if recognized and overtly stated, could be evaluated by appropriate techniques.

However, an evaluation which is constrained solely to the effects of the program on health can be expected to have only a partial influence on the decision process, and quite often a minimal influence, depending on the relative strength of the influence of the nonhealth factors.

Conclusions

Considering all of the foregoing, what then is a practical strategy of program evaluation? What should policymakers and grant review groups expect of program operators?

It seems obvious that studies of outcome or effectiveness (end results) are generally too complex and beyond the capabilities and interest of most persons directing the usual health and medical care programs. Furthermore, most funding agencies, be they private or governmental, would be unwilling to commit the resources necessary to do the job in every service project, assuming the necessary personnel could be found.

Assessment of end results or effectiveness should be done in selected situations by persons highly sophisticated in this type of research. Studies of this type should be designed and carried out in such a way that results are transferable to other places and situations so that end results in service programs can be inferred from the study of process or intermediate goals, in the manner indicated by Shapiro (*4*).

This concept implies that the evaluation in the usual service program or project should be confined to a quality control type of evaluation based on process or intermediate goals. The evaluation should be kept as simple and inexpensive as practical. The service program should not be undertaken unless the decision makers are willing to accept (either on the basis of previous evidence or faith) the premise that a properly conducted program of that type does ultimately have a beneficial effect on health.

REFERENCES

(*1*) Rice, D. P., and Cooper, B. S.: National health expenditures, 1929–68. Soc Security Bull 33: 3–20, January 1970.

(*2*) Deniston, O. L., Rosenstock, I. M., and Getting, V. S.: Evaluation of program effectiveness. Public Health Rep. 83: 323–335, April 1968.

(*3*) Donabedian, A.: Evaluating the quality of medical care. Milbank Mem Fund Quart 44: 166–206, July 1966, pt. 2.

(*4*) Shapiro, S.: End result measurement of quality medical care. Milbank Mem Fund Quart 45: 7–30, April 1967, pt. 1.

(*5*) National Center for Health Statistics: Conceptual problems in developing an index of health. PHS Publication No. 1000, Ser. 2, No. 17. U.S. Government Printing Office, Washington, D.C., 1966.

(*6*) Kellman, A. R., and Willner, A.: Problems in measurement and evaluation of rehabilitation. Arch Phys Med 43: 172–181, April 1962.

(*7*) Karlins, M., and Abelson, H. I.: Persuasion. Ed. 2. Springer Publishing Co., New York, 1970.

(*8*) Thorner, R. M.: Whither multiphasic screening? New Eng J Med 280: 1037–1042, May 8, 1969.

(*9*) McKeown, T.: Screening in medical care: Reviewing the evidence. Oxford University Press, New York, 1968.

(*10*) National Center for Health Statistics: Concepts and definitions in the health household interview survey. PHS Publication No. 584–A–3. U.S. Government Printing Office, Washington, D.C., 1958.

(*11*) Fein, R.: Definition and scope of the problem: Economic aspects: *In* Assessing the effectiveness of child health services. Report of the Fifty-sixth Ross Conference on Pediatric Research, edited by A. Bergman. Ross Laboratories, Columbus, Ohio, 1967, pp. 44–53.

(*12*) Emlet, H.: Progress in cost benefit analysis of multiphasic screening. Engineering Foundation Research Conference, Deerfield, Mass., Aug. 18–22, 1969.

(*13*) Galliher, H. P.: Optimal ages for Pap smears: A preliminary report. Interim report No. 4, Contract PH 43–67–737, Regional Medical Programs Service, June 1969.

(*14*) Krystynak, L. F.: Cost effective regional control of rheumatic fever. Doctoral dissertation, University of Michigan, Ann Arbor, 1970.

(*15*) Greenberg, B. G., and Mattison, B. F.: The whys and wherefores of program evaluation. Canad J Public Health 46: 293–299, July 1955.

(*16*) U.S. Department of Commerce: Census use study. General description. Report No. 1. Bureau of the Census, Washington, D.C., 1970.

(*17*) U.S. Department of Commerce: Census use study. Health information system. Report No. 7. Bureau of the Census, Washington, D.C., 1969.

(*18*) National Communicable Disease Center: Reported incidence of notifiable diseases in the United States, 1965. Morbidity and Mortality Weekly Report, vol. 14, No. 53, Annual supplement, Summary 1965, released Oct. 14, 1966, Atlanta, Ga.

(*19*) LaPiere, R. T.: Social change. McGraw-Hill Book Company, New York, 1965, p. 159.

(*20*) Elinson, J., and Herr, C. E. A.: A sociomedical view of neighborhood health centers. Med Care 8: 97–103, March-April 1970.

(*21*) Fourth annual report of the Poor Law Commissioners, 1838, *quoted in* H. Eckstein: The English Health Service: Its origin, structure and achievements. Harvard University Press, Cambridge, Mass., 1958.