

## END RESULT MEASUREMENTS OF QUALITY OF MEDICAL CARE

SAM SHAPIRO

The desirability of determining quality of medical care by its effect on some measurable aspect of health is matched by the pessimism among researchers about the possibility of success in dealing with the issue. A reversal in this outlook seems essential today in view of the need to assess the meaning of the enormous changes in organization, financing and accessibility of medical services that are just beginning to emerge. The forces for change are clear—Medicare; Title 19 and its objectives of assuring comprehensive medical care to all who cannot finance it themselves; demonstration programs of the Public Health Service, Office of Economic Opportunity, Children's Bureau and other governmental agencies that deliberately attempt to modify the nature of medical care that is available and used by selected target groups in the population; and other demonstration programs to broaden the scope, efficiency and effectiveness of medical care available to the population in general.

These programs offer an unprecedented opportunity to investigate the effects of medical care on health status, and evidence indicates that both the administrator and researcher are seeking ways to exploit this opportunity. Hopefully, this paper will help to accelerate the process by reducing to written form some of the thoughts that have been expressed from time to time about the nature of

end-result studies. This will be accomplished by reviewing what has been done and what is now underway and by providing a framework for viewing future research.

## GENERAL CONSIDERATIONS

For purposes of the present discussion, the term "medical care" includes the range of services available, the personnel and facilities for providing them and the conditions that affect their receipt, such as organization, costs and methods of financing them. The term "end result" refers to some measurable aspect of health status which is influenced by a particular element or array of these elements of medical care.

By definition, comparison is an essential element of end-result research and the variable of interest is some identifiable aspect of medical care. Ideally, all other parameters of the end result being measured are to be controlled so that they do not influence the comparison involving medical care differences. Contrasts may be found between aggregates of medical care existing during the same period of time or a modification may occur in one or more features of medical care, the effect of which is to be determined. In both cases, however, the end-result criterion of quality of medical care requires a comparison between two sets of measures. Finally, the measurement of end results requires observing population groups whose characteristics can be established.

Judgments regarding quality of medical care in terms of end results may also be made by determining that medical care associated with a designated end result is being provided in a manner that leads to the known end result. This type of research depends on fairly complete knowledge of the circumstances of the end-result study that demonstrated the end-result effect and their applicability to the situation under scrutiny. For example, assume that a multiphasic screening program leads to earlier diagnosis of conditions A, B and C, and that with appropriate follow-up and treatment, disability from these conditions is reduced. Then, to draw inferences about quality of medical care related to screening in a

particular medical care setting, an examination must be made of the availability of screening, its utilization, and the follow-up and treatment of conditions detected. Each of these components must be looked at critically to arrive at a conclusion.

In the case of utilization, a hard look at "performance" in a medical care setting will go beyond the overall rate of utilization and will examine the extent to which different segments of the population avail themselves of the screening program. The objective of this closer look is to have a basis for estimating the impact on health that might be expected from the program as it is being used. The end result of the program would be quite different if known high-risk groups appeared for examination than if utilization were concentrated among the low-risk groups.

Follow-up is dependent on the behavior of both the patient and the personal physician. As those engaged in screening programs know, one of the more difficult problems is to motivate the patient to seek appropriate follow-up care and to have the physician receiving the results of the screening examination pursue positive findings aggressively. Without knowledge of success in these areas, little can be said about the likely effect of the screening program in a particular setting. Similar types of questions can be structured for "availability," in terms of the organization and conduct of the screening program, and for "treatment" in terms of the methods that are being practiced.

In short, the application of the indirect approach in end-result studies will often not rest on a "presence" or "absence" determination, but will depend on a careful determination of the appropriateness of extending the results from direct studies to other situations and an assessment of the qualifications of doing so in a specific instance. Despite these complications, the indirect method should have a great appeal. It does not require the observation of two groups for later comparison and the study can usually be carried out relatively quickly. Often conclusions could probably be based on the existing information and modest extensions of it. The rub, is that the indirect approach must wait for evidence from the direct method and this has been a long time in coming.

## STUDIES WITH FINDINGS RELATED TO END-RESULT ISSUES

Until recently, one of the more important stimuli to end-result studies of quality of medical care was the desire to find ways to measure the impact of prepaid group practice on the health of its membership. This interest is now subsumed by the broader concern with how new, publicly financed programs of medical care are affecting health. Although the scope has broadened, much can be learned from the research on prepaid group practice.

In 1951, four or five years after the start of service, the Health Insurance Plan of Greater New York (HIP) became part of a comprehensive study which included among its objectives the "attempt to develop broad indexes of quality of medical care received by the general population and by the H.I.P. population."<sup>1</sup> Information was collected on morbidity levels, disability due to illness and medical care practices through a household survey of a random sample of families in HIP and in the city at large. The inquiry was wide ranging. It covered the entire age spectrum; all acute and chronic medical conditions present during an eight-week period prior to the interview; hospital experience over a year's time, dental conditions; use of preventive health services; and physician utilization. Personal characteristics included, in addition to age and sex, labor force status, occupation and industry, education, income, ethnic background and religion. Two general measures of disability were introduced, frequency of disabling cases and days of disability, both expressed as rates per 1,000 persons. A disabling case was defined as one causing inability of the patient to carry out his usual activities for one day or longer, and a distinction was made between disability that confined a patient to bed and other disability. In many critical respects the ideas, procedures and variables incorporated in the survey were similar to those later made part of the National Health Survey.

The research provided a wealth of descriptive information related to the above variables, which had as its end purpose a comparison between the situation in New York City as a whole with that in HIP. Differences were found between these two popula-

tions even after variations in age, sex, labor force status and education were taken into account. A larger proportion of the HIP membership saw a physician during the year; they were more likely to receive what was defined as preventive health services; more of them had family doctors, pediatric care for their children and dental attention than did the general population. Further, HIP members appeared to have a lower threshold for recognizing acute illnesses and they tended to seek medical care earlier in the course of illness than was the case in New York City as a whole. The magnitude of these differentials were in the main not very large, but they were unmistakably present.

Although the investigators were cautious about the potential of the data for "end-result" inferences, they performed the exceedingly useful task of examining these differentials to clarify the problems that arise in attempting to measure effects. The main variable in the study that touched on "effect" was disability. The number of days lost by persons in the labor force because of acute disabling conditions was greater for the HIP sample than for the New York City sample; rates of disabling chronic conditions were very similar in the two samples. In pursuing these observations, particularly with regard to acute disabling conditions, the use of disability (or some other general manifestation of morbidity) as a measure of an end-result is found to be complicated by the influence of the medical care environment itself on the concept of illness and on how the population responds to morbidity. Also, on matters of comparability between HIP, in this instance, and the total population, sources of bias have to be examined in terms of their effect on the specific, rather than the general measures under inquiry. These types of considerations lead away from multi-purpose, generalized research and lend priority to efforts that focus on specific hypotheses, and that provide an opportunity to take into account the more relevant sources of incomparability between comparison groups or to clarify how the results should be qualified.

The household survey concentrated on morbidity, utilization and disability, but the next study involving the HIP population was con-

cerned primarily with mortality. This time the inquiry was narrowed to a specific health issue and most of the above conditions could be satisfied. The investigation was aimed at determining whether women enrolled in HIP had lower prematurity and perinatal mortality rates than women in the general population of New York City.<sup>2</sup> If so, the study was to determine whether this was mainly due to differences between the medical care settings (prepaid group practice with qualified obstetricians in HIP versus fee-for-service, solo practice by physicians ranging from general practitioners to outstanding obstetricians in the city at large), or due to special characteristics of the HIP population. The methodology of the study was straightforward and economical. All information was derived from existing records of live births, fetal deaths and infant deaths from 1955 to 1957, on file in New York City Department of Health for the two comparison groups. Data routinely coded and punched were supplemented by information from the certificates on file concerning occupation and industry of the child's father and the identity of the physician in attendance for private patients. The former was used to obtain a broad measure of socioeconomic status; the latter to distinguish the diplomates in obstetrics-gynecology from other physicians delivering women. These items plus information on age, parity, ethnic composition and whether the woman was a private or general-service patient were expected to provide a firm basis for removing intervening variables, thereby isolating the effects of the medical care setting.

Biases that might not be controlled through available variables were considered. High on the list was the possibility that women in HIP who used obstetricians not associated with the plan did so when faced with difficult obstetrical problems. This type of selectivity would leave women with a favorable prognosis for pregnancy outcome under the care of HIP obstetricians. Prematurity and perinatal mortality data were available for a subsample of women enrolled in HIP, but who went outside of HIP for obstetrical care. The results of this check indicated that no important bias resulted from the missing group. Another point at issue was whether the subscribers in HIP differed substantially in their health

status from those eligible to enroll but did not. This might have operated in either direction, i.e., adverse or favorable selection. No information was available on the point, although the enrollment requirements suggested that the problem was not significant.

The results of this study are well known. A higher proportion of the women in HIP began their prenatal care in the first trimester and both prematurity and perinatal mortality rates were lower in HIP than in New York City in general. When the New York City group was restricted to women delivered in the hospital by private physicians, the differential in the stage at which prenatal care was begun almost disappeared. Differentials in prematurity and perinatal mortality were reduced in magnitude, but not eliminated. After taking into account differences in demographic characteristics between HIP women and other private patients in New York City, the conclusion was that "there is a small but significantly lower prematurity rate in HIP," and that the differential in perinatal mortality rates "is also statistically significant with HIP having a distinctly lower rate." A strong element of conservatism was introduced in the analysis by restricting most of the comparisons to HIP versus private patients in the city as a whole. Certainly, families were found who, if not enrolled in HIP, would have had the same patterns of medical care as the general service patients. No attempt was made to allow for this circumstance which tends to place the HIP members in a less favorable position when comparing them with private patients generally.

Many of the requirements of an end-result study were present in the prematurity and perinatal mortality investigation. These included measures (mortality and low birth weight) that are objective and easily defined; a common source of information for the comparison groups (records on file at the health department); and a sufficiently wide range of correlates of prematurity and perinatal mortality related to personal characteristics to satisfy reasonable criteria in controlling for "extraneous" differentials between the comparison groups. The term "reasonable" is used because, aside from an experimental design in which subjects are randomly assigned to study and control groups undergoing varied "treat-

ments," comparability cannot be measured with a high enough degree of precision to erase all doubts.

An element of judgment shaped by practical considerations will always be required in determining when to stop searching for biases that may explain an observed difference. Consequently, differences of opinion will exist as to whether the investigation has gone far enough. In the case of the above research, any doubts that may still exist about the role of the medical care setting in explaining the differences found arise from the inability to establish conclusively that enrolling in HIP is not in some way associated with more favorable health practices and health status.

Another aspect of research highlighted by the perinatal mortality study is that once the "end result" is accepted other questions are immediately raised. In the final analysis, all of these revolve around the question "why?" What in the prepaid group practice setting leads to improved pregnancy outcome? Is it the qualifications of the obstetricians, the organization of care on a group basis with all medical specialties readily available, or the cumulative effect of the type of medical care the women were receiving before pregnancy? Obviously, the value of the end-result finding of a general association between prepaid group practice and improved prognosis in pregnancy would be vastly increased if these questions could be answered. The difficulties, however, are commensurate with the stakes.

In the HIP-New York City study one abortive attempt was made to deal with a limited part of the problem. Private physicians used by women not in HIP were classified by whether or not they were diplomates in obstetrics-gynecology, the purpose being to provide a basis for investigating whether physician qualifications were responsible for the lower mortality rate in HIP (maternity cases in HIP were all under the care of board-certified or board-eligible obstetricians). This approach was abandoned in the face of evidence suggesting that the diplomates in New York City tended to get poorer risk cases than other private physicians. No practical research since then has studied the "why" question.

A by-product of the HIP-New York City study was the obser-



vation that perinatal mortality among nonwhites in the highest socioeconomic class (as measured by occupation of the child's father) was greater than perinatal mortality among whites in the lowest socioeconomic class. This has been confirmed through more recent and extensive data which explains the necessity for examining care available to nonwhite women, as well as the role of social and economic factors not reflected by a broad classification of father's occupation.<sup>3</sup>

Another research project which set for itself the goal of obtaining an "end result" has recently been completed.<sup>4</sup> This study was ancillary to the HIP-Welfare Demonstration Project, one of a number of experiments the New York City Department of Welfare undertook to improve the quality of medical care for the indigent several years before Title 19 became effective. At the time these experiments began the welfare recipient received his medical care from outpatient departments of local hospitals. Typically, he was seen by a physician who was contributing his services to the clinic or by a physician in training, and no one physician had continuing responsibility for the care of the patient. House calls were obtained from physicians on a panel maintained by welfare, and as a rule these physicians were unfamiliar with the patient's clinic record. If the patient was hospitalized, still a third set of physicians, the house staff, became responsible. Welfare clients in nursing homes received medical care from panel physicians. A nursing home patient was commonly seen by several physicians, none of whom could be considered the patient's regular doctor. Laboratory tests and x-rays were rarely done and medical charts were of the most rudimentary nature.

In September, 1962, the Department of Welfare enrolled over 13,000 public assistance recipients in seven of the medical groups affiliated with HIP in the largest of its experimental efforts to bring welfare clients into the mainstream of medical care. Of the new enrollees, 12,000 were receiving Old Age Assistance and living in their own homes. They represented about 38 per cent of the Old Age Assistance caseload in the city at the time. The other new enrollees were patients in proprietary nursing homes, and made up

about 30 per cent of the welfare clients in such homes. Welfare recipients enrolled in HIP were entitled to the full range of benefits available to non-welfare enrollees, except for a limitation imposed by welfare regulations on the HIP coverage for welfare clients. Hospital admissions were made to general service ward accommodations, and the HIP physician did not have responsibility for the welfare patient's care in the hospital.

Research initiated in conjunction with the demonstration program was aimed at determining whether enrollment in HIP resulted in changes in patterns of use of medical care and in mortality rates. The year beginning March 1, 1963, or six months after the demonstration project itself started, was selected for study because the early months of the project were a period of adjustment for everyone concerned.

Comparisons were based on the medical and hospital care experience of samples of persons receiving Old Age Assistance enrolled in HIP and of those not so enrolled. Another sample compared nursing home patients enrolled and not enrolled in HIP. A major problem affecting the comparison was that enrollment in HIP was determined by where the welfare recipient lived and the location of the nursing home. Administrative considerations precluded random allocation. To overcome difficulties caused by this circumstance in making comparative statements, utilization experience in the year before the demonstration program started and personal characteristics including age, sex, living arrangements and country of birth were ascertained from welfare records. For the study year, March, 1963 to February, 1964, data on medical and hospital services were derived from the records of the Department of Welfare and HIP. Deaths among persons in the HIP and non-HIP samples were identified through the same source.

In the analysis, the more important differences in personal characteristics that became known were taken into account. Physician visit rates were almost identical among the HIP and non-HIP Old Age Assistance recipients; hospital utilization rates were consistent with the difference found earlier, in the year preceding the

demonstration project. However, several changes were noted in the pattern of physician utilization that appeared to be associated with enrollment in HIP. Of the Old Age Assistance recipients, the proportion of HIP enrollees who received no ambulatory care decreased somewhat, and the corresponding proportion among non-enrollees remained unchanged. Where the HIP patient saw the physician shifted from primarily home visits to receiving most outpatient care in the medical group center. This change was partially due to special measures taken by the participating medical groups in HIP to increase the possibility that the Old Age Assistance recipients would obtain their medical care at the group centers where laboratory tests, x-rays and immunizations could be carried out. Medical groups arranged for transportation to facilitate the process for persons with impaired mobility.

Another observation was that patients who tended to be lower utilizers were likely to get more service when they were enrolled in HIP than they did otherwise. For instance, Puerto Ricans, a relatively low utilizing group, saw doctors more often if they were enrolled in HIP than if they were not. Also, the frequency of doctor visits among those who were low utilizers in the pre-demonstration year was raised substantially among Old Age Assistance recipients enrolled in HIP, but remained very low among non-enrollees. On the other hand, patients who, in the pre-demonstration period, used many physician services later continued to obtain large volumes of care, but they averaged fewer doctor visits in HIP than under the traditional system.

Differentials in mortality rates were examined for clues rather than definitive evidence concerning the impact that the change in medical care environment might have on the health of the recipients of Old Age Assistance. Speculation prior to the demonstration program was that in time a small but significant improvement in mortality would be associated with the shift of medical care to HIP. The observations are consistent with this hypothesis. During the study year, the death rates were about the same among the recipients of Old Age Assistance who were enrolled in HIP and

those not enrolled. In the next 18 months mortality among the HIP members was lower than among the other Old Age Assistance recipients.

With regard to the nursing home patients, no changes in physician or hospital utilization occurred in connection with the shift in medical care to HIP. On the other hand, far greater use was being made of laboratory services for HIP patients. Mortality rates were examined only for the study year and no difference was found.

How should the finding that Old Age Assistance recipients enrolled in HIP have a lower mortality rate than non-enrollees be interpreted? To the researcher, a single observation of this type could not be the basis for a conclusion and more evidence would be needed. The additional evidence might be limited to mortality among other samples of Old Age Assistance recipients during different periods than were covered in the previous study, but this would not be completely satisfactory. If a mortality differential does in fact exist, then other manifestations of the disease process, such as disability in its various forms, would be measurable and should be included as a prime concern of future investigations.

This question may soon be pursued on a broader basis. Under the Medicaid program in New York, Old Age Assistance recipients enrolled in HIP will be covered for out-of-hospital and in-hospital medical care from the plan's physicians. This will eliminate the critical break in continuity of care in the demonstration program. Other recipients will be able to receive the full range of medical care from physicians in the community at large. Comparisons of utilization experience and mortality and disability rates would thereby be concerned with a more general question than could be examined in the demonstration program; i.e., from a utilization or health standpoint, what is the difference between being enrolled in a comprehensive prepaid group practice plan and having available the traditional form of medical care in the community at large? Indigent persons under 65 years of age will also have the option of selecting HIP and if the magnitude and nature of the selectivity can be determined, the inquiry might be extended

to an age range in which larger differentials might reasonably be expected than among aged persons.

#### END-RESULT STUDIES IN PROGRESS

To survey the field completely for studies in progress that bear on the end-result question would be a larger task than could be undertaken in this paper. The alternate course is to locate a few studies that have end-result measurements among their stated goals to illustrate the variety of areas in which investigators are probing, without prejudging their chances of success. All of the research programs have multiple purposes, but only those related to end results are considered below.

Of considerable interest is a demonstration of health maintenance services in a prepaid group health program.<sup>5</sup> One of the specific objectives of this project is "to provide an array of protective health services, some well-tried and others newly devised, to a selected chronic disease population with the goal of maintaining optimum levels of function and comfort." Other objectives are concerned with methods of organization and coordination of special services, feasibility, utilization patterns and costs. Selected for the study are patients diagnosed as having one or more preselected specific diseases or disorders which might be expected to show "helpful effects within the time limit of the project," and for which the change in health status could be measured. Provision has been made to assign patients on a stratified random basis to study and control groups. Initial and final medical evaluation forms are to be completed by the project staff for both groups. These are to include several items directed at measuring end results, e.g., major symptoms and signs, severity, disability, number of attacks, complications and work time loss.

Evaluation of functional capacity has been projected on a "before and after" basis for the study patients only, except for a small subset of controls. Measures are to include occupational function (work time loss, work modification, job change, job loss), social function (household and family composition changes, daily activity

changes) and personal function (degree of self-care, sleep and rest patterns, personal hygiene habits, recreational activity). Direct measurement of an effect is not being sought. The emphasis in this area of evaluation is on determining whether the change appears to be influenced by social, medical care, behavioral, attitudinal or other characteristics of the study patients. However, information on the components of functional capacity defined by the research team and evidence on their correlates would be extremely useful to other investigators.

The methodology of one of the more comprehensive efforts to evaluate the effectiveness and efficiency of the service programs authorized under the Social Security Amendments of 1963 and 1965, illustrates how an investigator attempts to approach the problems of observational data.<sup>6</sup> In this project, an evaluation system has been designed "to describe the population of women served by the MIC projects so that the predisposing social, economic and medical factors which may influence outcome of pregnancy can be understood while assessing the adequacy of program services delivered and their effectiveness in influencing outcome." Maternal and Infant Care (MIC) projects are to be initiated in a variety of settings, each of which is directed at high risk groups of women. Comparisons are to be made over time within each project and among the projects to measure effect. The primary measures of outcome are to be birth weight, with various gradients of immaturity, and the Apgar score at five minutes to classify the infant as "essentially normal, moderately distressed and severely distressed."

The principal investigator describes the form of evaluation as "opportunistic" since measurement of the contribution of biases to differences found will be imprecise. However, the expectation is that a sufficient number of parameters of adverse pregnancy outcome will be available for use in the analysis to reach conclusions needed for program assessment and development. The need to seek opportunities to undertake "clinical trial type studies of defined services within the context of MIC projects" is stressed.

### *Other End-Result Studies*

Objectives of two other research programs are of interest. Both are planned by seasoned investigators who fully appreciate the requirements and problems of end-result studies. Both rely primarily on comparisons based on observations of situations as they exist in contrast to experimental approaches involving study and control groups. Reliability of inferences about end results will depend on the extent to which intervening variables can be identified and controlled in the analysis so that the contribution of medical care to differences that are found can be isolated.

One of the programs aims "to describe, relate and evaluate needs, demands, utilization, satisfactions, outcome and organization of personal health services provided in hospitals, clinics, physicians' offices, health departments and elsewhere. Epidemiologic and survey techniques will be employed but considerable emphasis will be placed on the development of new methods or the refinement of present methods for data collection. Wherever possible, representative populations in communities will be studied rather than institutional or selected populations."<sup>7</sup> The other study is an attempt "to determine how the outcomes of medical care relate to patient, doctor and hospital characteristics. This is a study of the quality of medical care that will use end results as a measure." Data are to be collected on "how medical care is used by different parts of the population in relation to disease and disability."<sup>8</sup>

### *Studies Related to Early Diagnosis of Disease*

Probably no more fundamental information would facilitate the conduct of end-result studies than knowledge of the natural history of disease, the physical, social and economic consequences of disease during well-defined intervals following onset, and the role of preventive and therapeutic medical care in altering the course of disease. This idea has been recognized for a long time, but the methodological problems and the personnel and time requirements have proven to be formidable barriers to undertaking studies that deal comprehensively with the issue.

Many useful studies have been conducted on selected aspects of the problem. For example, cancer case registers have permitted survival rates to be obtained at varying intervals following diagnosis, classified by type and source of treatment. Also, the ability of screening and preventive health examinations to detect disease in an early stage has been demonstrated. However, drawing inferences from them about the relative effectiveness of medical care alternatives is difficult because of selectivity factors and the unavailability of suitable comparison groups. This does not preclude the possibility that some of the data may later become valuable for end-result studies by serving as base lines when new treatments are accepted or a basic change occurs in the source of medical care.

During the past few years, two major research projects have been started to measure the effect on health status of detecting disease before it is usually brought to medical attention. One of the projects is being carried out in conjunction with an "automated multiphasic screening" program of the Permanente Medical Group, Oakland, California.<sup>9</sup> The program is designed to demonstrate how automation and computers can be applied to improve "speed, efficiency, and quality control in multiphasic screening techniques so that not only more tests, but more accurate and quantitative measurements can be performed, and at a lower cost." The screening program is integrated into the periodic health evaluation examinations for which several thousand patients enrolled with the Kaiser Foundation Health Plan volunteer each month. Appointments are made for these examinations by the patient's personal physician and he is rapidly informed of the findings. The likelihood of follow-up by the personal physician is high because of the close link that has been established between him and the multiphasic screening center.

The program is heavily engaged in evaluating costs, need for changes in equipment or organization, and the extent to which particular tests contribute to disease detection. This information will unquestionably be of enormous value in shaping the form and content of medical practice of the future. Its importance will be greatly enhanced by the fact that it includes a set of end-result



criteria in the evaluation. Two randomly selected samples of the plan's members have been designated study and control groups. Efforts are made to have the study group appear for the examination; the control group is not approached, but those who request an examination are accommodated. Morbidity, disability and medical utilization patterns are to be determined over a long period of follow-up through periodic questionnaires and medical records. This is an ambitious undertaking, but it has the potential of providing for the first time decisive information on the value of periodic health examinations generally and of selected components of it in particular.

The other study is being conducted by HIP.<sup>10</sup> Its main objective is to establish whether a breast cancer screening program using mammography (soft tissue x-ray) and clinical examinations results in lowering mortality from breast cancer in the female population. Other objectives relate to the epidemiology of breast cancer and the search for high-risk factors that might be useful in future screening programs.

Of the 31 medical groups in HIP, 23 are participating in the study. Within each of the participating medical groups, two systematic random samples of women aged 40 to 64 years with at least one year's membership in HIP have been selected. This was accomplished by first stratifying a file of punched cards for these women by age, size of insured family and employment group through which the family joined HIP. The women were divided into a study group and a control group. The total number of women in each sample is 30,000 with each medical group contributing a share proportionate to its size, adjusted to the availability of physician and technician time in its facility. The sample of study women in a medical group was randomized and women were drawn in sequence from the list as their turn was reached for screening examination. The date they were scheduled for their initial screening examination became their entry date into the study and all observations start from this date. Every study group woman had her matching woman in the control group who was assigned the same entry date. Following the assignment of these dates, pairs of

women were broken apart. The analysis of results of the investigation will be based on data related to characteristics of the total study and control groups and no paired comparisons will be made. Study group women are offered a screening examination in their medical group centers. Every woman who has an initial examination is asked to appear for three annual follow-up examinations even if she is no longer a member of HIP, unless she has a condition that requires earlier follow-up. Women in the control group follow their usual practices in receiving medical care.

A large investment is made to assure a high response rate among the study women. About 65 per cent have appeared for their initial examination. Of these 81 per cent are responding to the request for the first annual reexamination and 78 per cent for the second annual reexamination. As proportions of the total study group, these are 65 per cent (initials), 53 per cent (first annuals), and 50 per cent (second annuals). Third annual reexaminations have not started. The response rates are consistent with projections made before the study started as requirements to reduce the effect of the expected bias among persons who cooperate in a screening program. To measure the magnitude of the bias and the restrictions that it imposes on the study's findings, surveys are conducted to obtain characteristics of the non-respondent group that have been implicated as related to the development of breast cancer. Also, procedures to locate women who have newly diagnosed breast cancer and who die from breast cancer include the non-respondent study women along with the control group and the examined women.

Several overlapping sources of information help to identify women who undergo breast biopsy. They include the patient's medical record in HIP and notice of hospital claims paid by insurance. Surgical and pathological findings in breast cancer cases are obtained from hospital charts. The project's coordinating pathologist reviews slides and conducts special studies of tissue blocks, when available. Each case of microscopically confirmed carcinoma of the breast is investigated to establish the type of surgery performed, histologic type, nodal involvement and size of lesion. The

information on new cases diagnosed will provide an intermediate indication of the possible value of screening.

Effectiveness of the screening program will, in the long run, be judged on the basis of improvement in survival rates among women with breast cancer and on lowered mortality in the female population from breast cancer. For this purpose between five and ten years of follow-up will be needed for both the study and control groups. Deaths are being identified through intensive follow-up of all confirmed breast cancer cases and by matching death records on file in various health departments against the total files of study and control groups to locate deaths attributed to breast cancer. Comparisons of mortality rates will be straightforward and will provide the more definitive decision test of the value of the screening programs. Comparisons of survival rates will, on the other hand, require allowances for the acceleration in the date of diagnosis of breast cancer among screened women. Estimates are to be derived for the amount of time in the natural history of the disease between when breast cancer is detected in a screening program and when it is ordinarily detected.

#### FUTURE PROSPECTS FOR END-RESULT STUDIES

The review of past and current studies in the preceding sections covers methodologies that have varying degrees of practicality for wide application to end-result investigations. They range from comparisons between population subgroups for which some, but not all, of the significant intervening variables can be identified and controlled, to comparisons involving random allocation to study and control groups. Sources of information vary from data retrievable from existing records to observations derived by means of instruments specifically designed to satisfy the study's objectives. One set of conditions leads to results that are open to greater challenges than the other although findings in a study with even the most rigorous methodology require replication to answer questions about generalizability.

Presumably, continued special opportunities will arise allowing

study-control group techniques to be applied to medical care problems, and investigators will be found who are willing to accept the accompanying difficulties. Also, inroads will be made in developing knowledge about natural history of disease and the processes of medical care<sup>11</sup> and their correlates. The quicker this occurs the better will be the position of the investigator in selecting areas of research and in drawing inferences from end-result studies. This may be a rather slow process, however, and responses to the current pressure for evaluation of the changes in medical care cannot wait for the insights these studies can provide.

More aggressive moves are needed to exploit the current openings for end-result studies. "Before and after" studies would seem a natural, for example, in situations where a new health facility is being introduced in a relatively well-demarcated geographic area or where existing facilities such as a municipal hospital system are being reorganized. Such studies would also be useful in programs directed at special risk groups in the population (e.g., obstetrical care for unwed mothers), in programs where a defined segment of the population is likely to experience a change in medical care availability and accessibility (e.g., Title 19 as implemented in the various states), and in demonstration projects where the function of personnel and/or scope of service is being modified (e.g., use of nurses in a program of follow-up for particular conditions).

Another class of observational studies (as distinct from the experimental study control variety) that deserves more complete exploration involves intercomparisons in which effect is measured through an examination of differences in indices of health for populations in two or more medical care environments. The basic weakness of this approach, the problem of selectivity or bias, can never be completely overcome; furthermore, it is not always possible to determine in advance whether the bias can be reduced or sufficiently understood to interpret the differentials found. However, on occasion, even a relatively poor risk is rewarding, as in the case of a study a number of years ago in which attitudes and utilization of services were studied among persons who had been given a choice between HIP and group health insurance.<sup>12</sup> Prior to the

study, selectivity was presumed to be too strong a factor to overcome. This did not prove to be the case. An important reason is that the investigators had built into the research design the means for assessing the magnitude and direction of bias, and sound research principles were followed conducting the study.

The hazards of research based on observational data are not unique to end-result studies. The field of epidemiology is almost completely dependent on this type of information, but the investigators have been willing to accept the attendant risks without overlooking the restrictions that follow. If the approach to end-result studies were imbued with the same venturesome spirit as underlies epidemiology, and if this were supported with enough capital, the number of end-result studies would increase rapidly.

Objective circumstances for developing end-result studies are far more favorable than at any time in the past, even excluding the changing medical care scene. For example, procedures for organizing and conducting household surveys in the health field (sampling, questionnaire construction, field staff management, etc.) have been applied so often by now that they are no longer viewed as complex problems. Also, strengths and weaknesses of the household survey for health study use have been clarified.<sup>13,14</sup> Other information systems have been or are in the process of being developed, nationally and locally. These include the National Center for Health Statistics' health examination surveys of cross-sections of the population and its moves to collect data directly from providers of service using national samples; and the extension of reporting systems of the Social Security Administration and state welfare agencies (under Title 18 and Title 19 respectively) to inquire not only about fiscal issues, but also the nature of services obtained and the reasons for providing them. The range of variables that collection systems can now contemplate has increased in view of the advances in computer technology and its applications to multivariate analysis problems.

The mass of data collected through a National Health Survey or in connection with the national and state medical care programs is not likely to be adequate for an end-result study. However, they

do offer a way to design follow-back studies efficiently. An outstanding example of how a general-purpose survey is used as a first step in conducting an intensive inquiry into a specialized area is found in the Washington Heights master sample survey.<sup>15</sup> The design is applicable not only to small geographic areas or to household surveys, but to any defined segment of the population and to any systematically collected data.

To many who have seriously considered end-result studies, the major deterrent has been the difficulty in defining the measures to be used. The development of new ideas has been slow, but some progress has been made. Mortality is a clear, unequivocal measure which will continue to play a role in many studies. In some investigations it may be extremely useful as a broad measure of quality, e.g., in pregnancy outcome studies or in general comparisons involving different medical care settings or diseases where survival is presumably influenced by early detection and medical management. But measures more specific than mortality are needed. One method that appears to be particularly promising for applicability to end-result studies relates to the functional or social dimensions of morbidity, i.e., measuring morbidity in terms of its effect on the lives of the people concerned.<sup>16,17</sup>

The National Health Survey has already released a vast array of data on disability and its association with a wide range of conditions and personal characteristics.<sup>18</sup> The definition of disability and its components makes the data highly useful as a point of departure for more intensive studies. Disability is defined as "any temporary or long-term reduction of a person's activity as a result of an acute or chronic condition." The components for which information has become available are restricted—activity day, bed-disability day, work-loss day, and school-loss day. These general measures are being extended into areas of occupational and personal function as indicated earlier in the description of the health maintenance service demonstration project. Others have been developing projects suitable for special subgroups of the population, such as an index of self-care functional disability for aged persons in nursing homes<sup>19</sup> and measures of gross behavior changes among

psychiatric patients.<sup>20</sup> How well many of these ideas can be applied to end-result studies is yet unclear, but the increased involvement by investigators with this issue should demonstrate within the next few years what is useful.

## REFERENCES

<sup>1</sup> The Committee for the Special Research Project in the Health Insurance Plan of Greater New York, *HEALTH AND MEDICAL CARE IN NEW YORK CITY*, Cambridge, Massachusetts, Harvard University Press for The Commonwealth Fund, 1957.

<sup>2</sup> Shapiro, Sam, Jacobziner, Harold, Densen, Paul M. and Weiner, Louis, Further Observations on Prematurity and Perinatal Mortality in a General Population and in the Population of a Prepaid Group Practice Medical Care Plan, *American Journal of Public Health*, 50, 1304-1317, September, 1960.

<sup>3</sup> Densen, Paul M., Unpublished data for 1961-63, New York City Department of Health.

<sup>4</sup> Williams, Josephine J., Shapiro, Sam, Yerby, Alonzo S., Densen, Paul M. and Rosner, Henry, *Patterns of Medical Utilization by the Indigent Aged Under Two Systems of Medical Care*, Final Report, Health Research Council, New York City, September, 1966.

<sup>5</sup> Wilson, Glenn, *et al.*, Demonstration of Health Maintenance Services in a Prepaid Group Health Program, Cleveland, Community Health Foundation, in progress.

<sup>6</sup> Tayback, Matthew, Bianco, John J. and Farrell, Eugene V., Evaluation of Maternity and Infant Care Projects, Paper presented to Bi-Regional Institute in Administration of New Programs in Maternal and Child Health, University of California, Berkeley, January, 1966.

<sup>7</sup> White, Kerr L., Medical Care and Patient Care Research Program, Baltimore, Johns Hopkins University, in progress.

<sup>8</sup> Peterson, Osler, *et al.*, International Comparative Study of Medical Care, Boston, Harvard Medical School, in progress.

<sup>9</sup> Collen, Morris F., Automated Multiphasic Screening, Oakland, California, The Permanente Medical Group, in press.

———, Periodic Physical Examinations and Multiphasic Screening, Presented at Group Health Institute, May, 1963.

<sup>10</sup> Shapiro, Sam, Strax, Philip and Venet, Louis, Evaluation of Periodic Breast Cancer Screening with Mammography, *Journal of The American Medical Association*, 195, 731-738, February 28, 1966.

<sup>11</sup> Donabedian, Avedis, Evaluating the Quality of Medical Care, in Mainland, Donald (Editor), *HEALTH SERVICES RESEARCH I*, New York, Milbank Memorial Fund, 1966, pp. 166-207.

<sup>12</sup> Anderson, Odin W. and Sheatsley, Paul B., *Comprehensive Medical Insurance*, Health Information Foundation Research Series No. 9, 1959.

<sup>13</sup> Feldman, Jacob J., The Household Interview Survey as a Technique for the Collection of Morbidity Data, *Journal of Chronic Diseases*, 11, 535-557, May, 1960.

<sup>14</sup> *Health Interview Responses Compared with Medical Records*, Vital and Health Statistics, National Center for Health Statistics, Series 2, No. 7, and *Comparison of Hospitalization Reporting in Three Survey Procedures*, Vital and Health Statistics, National Center for Health Statistics, Series 2, No. 8, Public Health Service, Publication No. 1000, Washington, D. C., July, 1965.

<sup>15</sup> Bailey, Margaret B., Haberman, Paul W. and Sheinberg, Jill, Identifying Alcoholics in Population Surveys, *Quarterly Journal of Studies on Alcohol*, 27, 300-315, June, 1966.

<sup>16</sup> Sanders, Barkev S., Measuring Community Health Levels, *American Journal of Public Health*, 54, 1063-1070, July, 1964.

<sup>17</sup> Sullivan, Donald F., *Conceptual Problems in Developing an Index of Health*, Vital and Health Statistics, National Center for Health Statistics, Series 2, No. 17, Public Health Service Publication No. 1000, May, 1966.

<sup>18</sup> National Center for Health Statistics, Series 10, No. 12, 15, 17, 24.

<sup>19</sup> Muller, Jonas N., Rehabilitation Evaluation—Some Social and Clinical Problems, *American Journal of Public Health*, 51, 403-409, March, 1961.

<sup>20</sup> Gruenberg, Ernest M., Brandon, Sydney and Kasius, Richard V., Identifying Cases of Social Breakdown Syndrome, in Gruenberg, Ernest M. (Editor), *EVALUATING THE EFFECTIVENESS OF MENTAL HEALTH SERVICES*, New York, Milbank Memorial Fund, 1966, pp. 150-155.

## ACKNOWLEDGMENT

*Thanks are due Dr. Mindel Sheps whose comments and suggestions led to a significant change in the paper. This paper was prepared for a Conference on Appraisal of Quality and Utilization of Services, Program Area Committee on Medical Care Administration, American Public Health Association, November, 1966.*