At a time when the problem of quality medical care is becoming of increasing concern to the health profession this statement prepared for the Program Area Committee on Medical Care Administration is highly relevant.

THE MEDICAL AUDIT AS AN OPERATIONAL TOOL

Mildred A. Morehead, M.D., M.P.H., F.A.P.H.A.

THE two major questions that arise when studies of the quality of medical care are considered are "What is quality medical care?" and "How can it be adequately measured?" Both definitions and measures, as well as study objectives, are legion. One excellent definition of care has been given by Esselstyn:

> Standards of quality of care should be based on the degree to which this care is available, acceptable, comprehensive, continuous, and documented, as well as on the extent to which adequate therapy is based on an accurate diagnosis and not on symptomatology.[1]

Within the boundaries of this statement lie at least seven areas to evaluate, and areas which cannot be readily measured by the same instrument. Quantitative measures of brick and mortar, surveys of patient attitudes, components of adequate documentation, to say nothing of defining accurate diagnosis and adequate therapy, are all areas which would profit by established measuring tools.

The need to develop a method of study that is efficient, reliable, objective, reproducible, and universally accepted has been recognized by all those working in this field. Yet almost every worker in this area has also come to the conclusion that the clinical judgment of the attending physician is also a major factor and that this aspect cannot readily be measured with precise tools.

Lembcke, who in his development of a *scientific method*[2] attempted to minimize subjective judgments, allowed for a *standard degree of compliance* which could vary from 50 to 100 per cent according to the specific disease selected. This wide margin was felt to be necessary in order to allow for a place for clinical judgment in those disease categories where detailed standards or criteria for optimal care would not have encompassed all the variations of the pathology manifested in different individuals.

Among the substantial efforts made to assess quality was a method developed in the studies of medical care delivered by the Health Insurance Plan of Greater New York which, while failing to a considerable degree to contain components of an ideal measurement, was yet very effective in relation to its objectives.

The Health Insurance Plan (HIP) Studies

The level of medical care provided by the 32 medical groups affiliated with HIP and the adequacy of the financing of care have long been matters of discussion between the Board of Directors, the Central Office, and the medical groups.

In 1948 Makover reviewed clinical records of specified types from each medical group.[3] The results of this review, combined with selective attributes

of the group facilities and administrative policies, led to a rating for each medical group. The findings and recommendations of this study were very useful in specifying areas in need of improvement.

In 1953, during contract negotiations between the medical groups and the Plan, the subject of quality again came to the forefront. The medical groups maintained that additional financing was essential to provide the highest level of care; the Board of Directors responded by asking what the existing level of care was.

Immediately following the negotiations a study team was organized by the Central Office to conduct studies of the quality of care and to make recommendations for improvement, should deficiencies be found.[4]

An outstanding clinician, Dr. I. O. Woodruff, was selected to head the study team. An associate medical director (the author) from the Central Office staff was assigned full time to this project. Discussions regarding the scope and methods to be used were held with a wide range of physicians and health workers. During this period alternate ways of evaluating care were discussed. The use of indexes ranging from death and morbidity rates to the more sophisticated technics developed by the Commission on Professional and Hospital Activities (PAS) was accepted as a valid and useful way of demonstrating differences in the various components of health care; e.g., hemoglobins, urines, length of stay. It was felt, however, that such measurements alone were not sensitive enough to enable judgments to be made of an individual physician's performance, for there was early agreement that the essential element in the provision of medical care is the professional judgment of the attending physician trained to perform in accordance with the accepted present-day standards of treatment and knowledge of disease.

Prior to developing the evaluation method, it was also agreed that:

(a) Determining performance levels of individual physicians, as opposed to an over-all group profile, would enable more rapid strides to be made in the improvement of care.

(b) The study would deliberately not cover the highly important areas of patient attitudes or their degree of satisfaction. The study would be directed toward the level of professional quality; for, despite the importance of a satisfactory patient-doctor relationship, it cannot be accepted as a substitute for poor professional work.

(c) The study would initially encompass clinical fields providing the greatest volume of service—namely, medicine, pediatrics, surgery, obstetrics and gynecology, pathology, and radiology.

Outstanding clinicians with no prior association with the Plan, were selected for each of the fields to be studied. All of the team members were affiliated with teaching institutions and, to avoid the problem of *ivory tower* standards, were also engaged in clinical practice. The decision was made that the clinical judgment of the surveyor was to be the yardstick against which individual performances would be rated.

The original study design envisioned a rating derived from three components for each group physician:

(a) Assessment of the performance of preventive health measures (for pediatrics, medicine, and obstetrics only) from a review of records.

(b) Assessment of the management of ten cases of specified illness from a review of records and discussion with the physician.

(c) Responses to a questionnaire regarding administrative and professional relationships between the physician and his medical group.

As the study progressed it became apparent that the review of the ten cases of illness was most useful in the delineation of problem areas and in identifying physicians who were in need of greater supervision.

The rating of quality in the preventive health field consisted of numerical weights being given for the presence of

specified items in medicine, pediatrics, and obstetrics. The only rating in this area that was of a subjective nature was the judgment pertaining to the quality of the recorded history and physical examination. The results of this aspect of care showed uniformly good results in the fields of obstetrics and pediatrics and uniformly poor ratings for the family physicians in the departments of medicine. Although the findings from the latter group could be translated into recommendations for the group as a whole, it did not serve to distinguish levels of performance between individual physicians.

The questionnaire, relating to professional policies and practices such as the use of the Visiting Nurse Service, indications for referral, and postgraduate educational courses, failed to be a useful tool. Far too frequently the *correct* answers were supplied, but the indicated action failed to take place in patient management.

Obviously, the way in which the ten cases of illness were selected for review was a crucial point in the study design. It was agreed that no purpose would be served by a random selection of cases, because either the majority of patients coming to the family physician had conditions that were of a self-limiting nature, or sufficient documentary evidence for supporting the diagnosis would not be available. The decision was made that cases would be restricted to specific major illnesses where confirmatory evidence would be expected.

If the same number and type of cases in the same stage of disease could have been selected for each physician, then in all probability more serious efforts would have been made to provide detailed criteria for each condition to be studied. For the more than 400 family physicians in the Plan, this was clearly not possible. Diabetes, hypertension, coronary artery disease, peptic ulcer, anemia, and kidney disease were the conditions selected for review in medicine. Cases of carcinoma and liver disease were included when encountered. It was felt that these cases would need fairly extensive diagnostic procedures, involve the group specialists and facilities, and would require the physician to exercise more skill and acumen than are required by more common conditions such as respiratory infections and minor trauma. An attempt was made to select two cases from each disease category mentioned in order to obtain a broad picture of physician performance. When this was not possible the list was supplemented with other chronic diseases.

The source for selection of cases was a form which each physician submitted monthly to the Plan's headquarters and which included the patient's name, identifying data, and a tentative diagnosis of the condition for which the patient sought care. A list of cases was developed from a review of three months' submission of these forms from each physician for a period six to nine months before the date of the interview.

The internist on the survey team interviewed each family physician at the office where he saw the majority of his HIP patients. The average interview lasted about two hours; the range was from an hour and a quarter to three hours. Payment to the surveyor was initially made on the basis of an hourly rate. For this field, it was later changed to a monthly retainer which was based on an expected number of weekly interviews. (Surveyors in other fields were paid a set amount for each group studied.)

The surveyor asked for the first ten available charts from a list of 15 cases of illness. He reviewed each case record and entered on an evaluation card (see Appendix A) a rating of *good*, *fair*, or *poor* for all items listed. The cases were then summarized on the back of the card, where all laboratory work, x-ray studies, and consultations were noted.

The dates of the first and last visits and the total number of visits were recorded. The charts of patients who had made only one visit, or who had been completely worked up before being seen by the family physician, were omitted. The surveyor also enumerated those items which he felt were indicated but not done. The cases were discussed with the physician who was given the opportunity to comment on each case and, in theory, supply missing information. This aspect of the study method was not felt by any of the surveyors to be of value. It invariably placed the physician being reviewed on the defensive and became an apology for his record keeping. It did not serve to elucidate further understanding of the clinical handling of the case—a very sensitive area when compared to that of record keeping. On the other hand, the group physicians frequently did not feel these discussions were thorough enough, particularly when they felt, as was generally the case, that a lowered score resulted from failure to give credit for items said to have been done whose results were remembered.

Items Evaluated

Each evaluation card contained the items listed below:

I. Records
- A. History
- B. Physical examination
- C. Progress notes
- D. Organization of the medical record
- E. Justification of the recorded tentative diagnosis

II. Diagnostic Management
- A. Time involved in obtaining indicated procedures
- B. Indicated laboratory studies, with a minimum of hemoglobin, urinalysis, and serology required in every case
- C. X-ray examinations, with a minimum of chest film required in every case
- D. Indicated consultations
- E. Summary of over-all diagnostic handling

III. Treatment and Follow-up
- A. Therapy
- B. Follow-up laboratory and x-ray studies
- C. Adequacy of follow-up visits
- D. Over-all management

As previously mentioned, because of the variety of conditions studied and the varying degrees of clinical severity, no attempt was made to have specific criteria for each disease studied. However, general criteria were available for each item; for example:

History

Rating

Good—History includes present illness, family, and past history. If a complete history is present somewhere within the chart, an interim history will be sufficient.

NOTE: If history is adequate in all respects except that of family history, please note.

Fair—Record includes chief complaint and history of the present illness only.

Poor—Record includes chief complaint only, or nothing.

Each case studied had a potential value of 100 points. Each item had an arbitrary weight which was fixed for designations of *Good* or *Fair*. Items rated *Poor* received no credit. The general categories had the following weights:

Weights

	%
Records	30
Diagnostic management	40
Treatment and follow-up	30

The final score given to a physician consisted of the averaged scores of the cases studied. To assure uniformity, editing and scoring were done by the associate study director.

In order to relate a physician's score to a level of acceptable performance, a conference was held with the interviewing internists as well as with others who had been interested in the project. Fortunately, from a methodological standpoint, the first medical group studied had 16 family physicians whose final scores ranged from 19 to 97. After reviewing the over-all clinical picture of the cases handled by each physician, it was felt that the physicians could be grouped into four classes, each representing a different level of medical practice. Arbitrary lines were drawn at 45,

60, and 75 per cent. It was felt that any physician receiving a score of 61 or higher was practicing an acceptable quality of medical care. Those scoring between 46 and 60 were considered to have below-average performance levels; and those with scores of 45 or less were judged to be rendering a poor quality of medical care. Considerable consistency was noted in the ratings of individual cases of physicians in the various categories; e.g., a physician in the second rating class usually had the majority of his cases scored between 61 and 75.

The results of the study were summarized in clinical terms and presented to each medical group with the case examples demonstrating areas of weakness and strength of both specific physicians and the group as a whole. Individual physicians were given a code number which was released only to the group medical director. For physicians who had received a score of 60 or less, corrective measures were recommended —either more intensive supervision by the chief of the department or termination of the physician's affiliation with the group.

An indication of the validity of at least one extreme of the scoring system was the fact that with few exceptions the medical groups concurred with the study team's evaluation of physicians in the lowest scoring group, and by and large these physicians were dropped from the medical groups.

As individual scores came closer to the dividing line of 60, greater resistance was encountered in the groups' acceptance of the ratings as valid. All physicians with scores lower than 60 were resurveyed within six months. Several years after the completion of the study, one-third of the physicians whose scores had been between 45 and 60 had ratings above 60, one-third had left the medical group of their own volition, and the remaining third were still a point of disagreement between the group and the headquarters office.

The impact of these studies on the medical groups was felt to be considerable. Many improvements were made in the record systems, administrative policies relating to patient care, and in the organization of the various clinical departments. The position of the "chief" of each clinical department was strengthened, as more emphasis was placed on his role in the supervision and responsibility for care provided by members of his department.

Of equal importance were the measures undertaken by the Central Office of the Plan as a result of the study findings. When characteristics of the family physicians were examined in relation to their performance ratings, the relationship that was the most outstanding was the number of years of approved hospital training after graduation from medical school.[5] The standards for new family physicians were raised to require that all applicants must have had three or more years of such training before becoming eligible to join a medical group.

Another characteristic that was found to be associated with those physicians having higher scores was the amount of total practice time devoted to HIP patients. This was one added source of data which helped the Central Office develop plans to achieve a greater number of full-time physicians within each group.

During this period, Peterson was conducting his studies of general practice in North Carolina.[6] By having an internist spend several days observing the performance of a general practitioner, a *score* was given to each of the 94 physicians reviewed. The proportion of physicians found to be providing medical care of *below average* quality was higher than that for the *below 60* category of physicians in the HIP studies. Certainly, however, no conclusion could be reached

that general practice in North Carolina was below the level of the New York City HIP physicians. It was our impression that the standards in the HIP studies were considerably more lenient, and perhaps unjustifiably so. In the analysis of HIP family doctors and their characteristics, three classes were used: Class I—those physicians with scores over 74; Class II—those with ratings between 60 and 74; and Class III—those scoring below 60. There was evidence that the Class II physicians did have different characteristics than those in Class I, and that the speculation could be made that a considerable portion of these physicians might have been placed in Peterson's lower rating classes. The idea of using his direct observation method, particularly for the "follow-up studies," was considered at length but finally abandoned, primarily because of expense and difficulty in obtaining surveyors with sufficient time available. There was also the feeling that by concentrating on the immediate doctor-patient contact much valuable information relating to over-all patient management and follow-up was lost.

The first quality study at HIP took four years to complete. Following this, reviews were undertaken of other fields —allergy, otolaryngology, ophthalmology, orthopedics and physical therapy, psychiatry, neurology, and urology. This comprised all of the specialties provided by the medical groups with the exception of dermatology, which for some inexplicable reason was omitted. The findings leading to recommendations in these fields, as well as in most instances of the specialty fields reviewed earlier, were with rare exception concentrated in the areas of professional and administrative policies. Clinical judgment and performance were the issues only in isolated instances. This was primarily due to the fact that the Plan had maintained high standards for specialists since its inception, while in the early years standards for family practitioners had been given less consideration.

Problems with the Method

Weighting

The arbitrary assignment of weights to different components of care was one of the most difficult areas to defend against the criticism of the group physicians. In large part, it was dictated by the volume of material to be studied, e.g., 4,070 cases for the family physicians alone. In presenting the findings to the medical groups, every effort was made to retranslate into clinical terms all of the lower ratings; this had been the prime reason for requiring the surveyor to summarize each case. Administratively, the weighting system was invaluable; but it continued to present problems, not all of which were ever resolved.

During the pretesting of the evaluation card and its weights, the therapy component equaled the diagnostic area. This occasionally resulted in a judgment of *fair* or *good* for the therapy itself with incomplete information as to what condition was being treated. For this reason, the diagnostic management was subsequently given heavier weighting. It is of interest that at about this same time Butler and Quinlan were independently developing a rating system for auditing care for hospitalized patients.[7] They had encountered similar difficulties with therapy rating and had adjusted their weights accordingly. There was marked similarity between the two methods.

During pretesting a summary item related to patient outcome was included. The motivation behind this item was admirable. It was designed to answer the question "Did the physician's management of this patient result in amelioration of the patient's disease process?" However, there were too many disease entities under consideration for such a question to apply equally to all condi-

tions. As a compromise an over-all patient care rating, allowed only five points out of the 100, was substituted.

Another weakness in the scoring technic, which was never completely overcome, was the failure to allow for interdependence among the three areas of *Records, Diagnosis,* and *Therapy* as, in theory, each was scored independently. Consideration was given to the development of interlocking scores; i.e., if the diagnostic area failed to rate *X* number of points, then the therapy area could not exceed *X* number of points. It was felt, however, that this would be an additional artifact to explain to the clinicians. To a certain extent the 20 points allocated for a summary of diagnostic handling were used to serve this purpose.

The weighting method created problems in the surgical fields. The list of cases selected for the review of each surgeon was divided into *Major Surgical Procedures, Minor Surgical Procedures,* and *Consultations.* The difficulty resulted from the allotment of the same weights to each type of case. It did not seem logical that a varicose vein stripping should be equated with a gastric resection, when both procedures were handled in a creditable fashion. It was even less logical when the score of a minor procedure could mask inadequate handling of a major procedure.

The scoring also presented some problems when consultant-specialists were reviewed, whether they be the supervising internists, surgeons, or ophthalmologists. Unsatisfactory ratings for these specialists in general did not have the same connotation regarding performance as they did in the case of family physicians.

Frequently, the way the consultant viewed his role in the medical group influenced his action. Some specialists considered themselves purely consultants, with their responsibility limited to a review of the presenting problem and to making recommendations to the referring physician for further action. In groups where the family physicians were all certified internists this approach did not adversely affect patient care. Where this was not the case—as in the great majority of the medical groups—patient care frequently suffered. Therefore, for the basic specialties of *Medicine, Surgery,* and *Gynecology,* and particularly for *Pediatrics* (in groups where family physicians provided pediatric care), the study team felt that the consultant should assume over-all responsibility up to and including periodic supervision of the case. In other fields this was not always rational. At the inception of the second study the surveyors were asked to give a rating on the total care of the patient. Scoring of this area was abandoned, however, after the plaintive complaint of the ophthalmological surveyor that it had been "a long time since being involved with the ordering and judging of babies' formulas."

The methods described were recognized to be a crude index at best. They served their purpose well in delineating those physicians providing medical care of an unacceptable level of quality, but were not so efficient in separating the *average* from the *good* or *excellent* performances. For example, a physician who had given a patient with a peptic ulcer an adequate basic work-up, repeat hemoglobin levels, and x-rays when indicated, would have received a score of 100. But another physician who, in addition to all the services mentioned, had arranged for the patient to discuss his diet with the consultant-nutritionist, had involved the social worker in possible problems existing in the patient's home, or had referred the patient to a psychiatrist if indicated, would have received the same score.

The method also lacked effectiveness in determining overuse of diagnostic aids. One of the medical groups, affiliated with a strong teaching hospital, specifically requested the study team to

examine this area, and careful tabulations were made of all tests and their frequency. There was no question that the use of laboratory and x-ray facilities was considerably higher for this group than for other groups studied. However, it was found to be much more difficult to justify criticism of overuse than it was to condemn absence. This same question arose later in studies of hospital care and once again was not resolved to the point where a negative rating resulted from overutilization.

The Teamster Studies

The motivation for the studies undertaken for the Teamsters[8,9] was basically the same as for the studies described earlier—namely, the cost of medical care. Both the management and labor Trustees of the various Welfare Funds providing health benefits were concerned about the increasing expenditures required for hospital and medical care. A special Trust Fund* was established to explore solutions to this problem. At the request of this fund, the Columbia University School of Public Health and Administrative Medicine undertook a series of projects in this area. One phase was the provision of a course on hospital operation for the Trustees.[10] Every two weeks an afternoon or evening was spent visiting various departments in a hospital, hearing lectures from leaders in this field, and observing such activities as the midnight change-over of the nursing staff or the functioning of an emergency room. At the same time, a study on the costs and quality of medical care received by a sample of Teamsters and their families was undertaken. As the Trustees' interest was primarily in the hospital costs, the decision was made to focus the studies on *inpatient admissions*, the *quality of care received*, and *whether or not the hospitalization was indicated* as important components of studies of costs.

* Teamsters Joint Council No. 16 and Management Hospitalization Trust Fund.

Several factors from the earlier HIP studies influenced the design of this study. One was case selection. The earlier studies had shown that it was advisable to select conditions where confirmatory evidence of the diagnosis would be present in the physical findings, diagnostic work-up, or course of illness. Therefore, the sampling frame consisted of all hospitalized patients with specified diagnoses in the fields of medicine, surgery, and obstetrics and gynecology during a six-month period. The data were supplied through the cooperation of the local Blue Cross Plan, which provided coverage to all Welfare Funds participating in the Trust Fund. (The three-digit International Classification of Disease code, plus a surgical code, was used by Blue Cross to identify diagnoses.) A random sampling was made in the different disease categories (with the exception of cesarean sections, where all 13 cases were reviewed). To protect the confidential nature of the diagnosis, the final names supplied by Blue Cross were not identified by diagnosis, but the sample was drawn according to study design specifications. Considerations such as the cost of photostatic copies of records and the amount of case material one could expect an outstanding clinician to be able to review within a reasonable amount of time also influenced the total number of cases to be selected. (Three hundred and three admissions were in the final sample selected.)

The importance of ambulatory care and management to the total course of disease had been evident from the HIP studies. In fact, except for the surgical specialties, little attention was given to the hospital period *per se*, the assumption being that care had been adequate. Therefore, in the first Teamster survey the household interview was designed to obtain information about the patient's medical usage and symptomatology prior and subsequent to the hospital admission, about his out-of-pocket costs, and

to secure his authorization for obtaining copies of his hospital records. However, the data were of little value in identifying pertinent facts in relation to the hospitalization. This was due primarily to lack of specificity of the questions (the diagnosis of the particular patient being interviewed was unknown). Furthermore, the lay person's unfamiliarity with medical practice on the part of both the respondent and the interviewer resulted in a great deal of irrelevant information.

As it was realized that the findings would not be used to focus attention on an individual physician or hospital, it was not felt necessary to use the scoring mechanism, its primary function being to isolate physicians and specific areas of patient care in need of improvement. Furthermore, it was believed that the arbitrary nature of any weighting system only added to the controversy over methodology.

The question of developing criteria arose again for discussion, but again it was decided to use the professional judgment of the interviewing physician as a standard. This was felt to be particularly necessary for cases in the field of internal medicine. Quality studies in this field have lagged far behind those in surgery, being hampered in large part by the variations occurring in any specific disease and by the absence of such definitive action as a surgical incision with its usual by-product of tissue to confirm or deny the rationale of the procedure.

The surveyors in this study were essentially the same as those participating in the HIP studies. Remuneration was made on the basis of $10 for each patient whose record or records were reviewed.

The second Teamster audit was based on hospital admissions that occurred three years after the first study. Because the first audit had been written for lay Trustees and had little methodology described, considerable criticism was directed toward various aspects of it. For this reason certain modifications were made in the approach to the second audit, but none which might impinge upon the study team's fundamental philosophy—namely, the importance of professional judgment in case management and the artificiality of the limits imposed by detailed criteria of case management. This last factor became rather academic, as it was decided that a completely random selection of admissions would be made among patients hospitalized during a one-month period, with the exception of normal deliveries and tonsillectomies and adenoidectomies. This decision resulted from experience from the first study, where many of the illness records obtained did not in fact fall into the diagnostic category requested. However, it was found that care for any inhospital patient lent itself to evaluation, as an expected course of action should follow (a) discussion of the presenting problem, (b) diagnosis, and (c) therapy. The very fact that the patient was in the hospital, for whatever reason, took him out of the *minor illness category*, a category deliberately avoided in the HIP studies of ambulatory care.

Among the other methodological changes that were made was the use of two surveyors for each record in the fields of medicine, surgery, and pediatrics. The degree of difference of opinion on purely clinical matters was found to be of little importance. The final rate of disagreement between surveyors was 8 per cent. Even this did not entirely represent disagreement on purely clinical grounds, particularly as related to the necessity for hospitalization. The adequacy of records, the weight given to corollary aspects of patient care, and the problem of assessing each of a series of admissions for the same patient independently when a gross error of judgment had been made in an earlier ad-

mission, were also among the causes for disagreement. It was not felt that the level of disagreement would have influenced the study findings or recommendations.

The first audit had been prepared, as noted earlier, for a lay body—the Trustees of the Trust Fund—and great care had been taken to avoid the use of terminology that would not be meaningful to them. In the second audit the Appendix contained clinical summaries of all cases reviewed, in order to present to medically oriented readers the clinical bases for judgments made.

The impact of the Teamster studies was considerable, both for the group sponsoring the studies and for the community as a whole. For the Teamsters, a special center was established at Montefiore Hospital for providing consultation and treatment for certain specified conditions and diagnostic evaluation for all problem cases. An educational campaign was undertaken by means of newsletters and brochures for the purpose of informing the average Teamster family of the varying characteristics of physicians and hospitals and acquainting them with the components of good medical care. In their attempts at finding solutions to the problems facing them as the responsible agents for providing types of care to the Teamster members and their families, the emphasis of the Trustees broadened to include concern with quality of performance as well as with costs of care. Following the completion of the second audit, plans were initiated to offer as a choice to eligible members a type of medical practice that would guarantee the use of well-qualified specialists and a highly organized and respected teaching hospital. The Teamster Comprehensive Care Program (TCP) began operation at Montefiore on July 1, 1966, serving all eligible members in the area wishing to participate.

Using many of the findings of these studies, as well as others performed by the university, the commissioner of hospitals, at that time Dr. Ray E. Trussell, was able to establish and strengthen codes for the performance of surgery and other specialties in the proprietary hospitals under the department's jurisdiction. Subsequently, a State Hospital Code, with almost identical requirements, was adopted equally affecting all hospitals. Payment by the city for care in special disease categories in voluntary institutions was limited to those meeting special standards, particularly those related to hospital organization and physician qualifications—the two major factors associated with high quality performance in the Teamster studies.

The essential finding of these audits —namely, the relationship of the highest quality of medical care to the qualifications of attending physicians and to the type of hospital—has been used as a simpler method of reviewing care obtained by several other large consumer groups in the city. For such groups, a year's hospitalization experience was analyzed in regard to these two areas. The relationship between qualifications of physicians and the class of hospital has been markedly similar in all of these studies; the proportion of well-trained physicians providing care increases strikingly as the class rating goes from an unaccredited proprietary institution to that of a medical school-affiliated institution. One discouraging aspect of these reviews has been that as family income increased and as coverage for medical expenses increased, the proportion of care obtained in institutions where quality of care could be questioned also increased. A low-income group, with minimal hospitalization coverage ($15 per day) and only a limited indemnified surgical fee schedule, showed a considerably higher proportion of patients receiving care in the ward services of the city's teaching institutions—a plus factor in relation to quality. This group had only 12 per cent of hospitalizations in proprietary institutions, compared to the 30 per cent

experienced by the Teamsters with their far broader health coverage.

During the course of these studies, extensive efforts were made to develop indexes of patient care from the records that might lead to a simplified and more standardized way of obtaining evaluations of patient care. A registered nurse on the study staff, with extensive experience in working with and abstracting medical records, tabulated many items of patient care in equivalent cases. Many attempts were made to correlate such items with the surveyor's rating. The items ranged from patient identification data to salient factors in the history, laboratory and x-ray findings, to confirmation of pathology reports and other characteristics of the hospital stay. This was done for every group of diagnostic cases, from both studies combined, where more than ten cases of a similar diagnosis were found. The results were not encouraging, as in general the reasons for adverse judgment were unique for the particular patient and were not identified by the indexes no matter how detailed they were. This also had been the experience in the first audit, where the surveyor of cases of diabetes mellitus had, in addition to the form where he summarized his opinion and judgments, an extensive check-list with specific items relating to this disease up to and including strength and scheduling of medications. No profitable use, however, was ever made of these data despite repeated attempts to do so. The only suggestive relationship that occurred was in cases of peptic ulcer surgery, where those with less than favorable ratings all had ulcer history of less than one year, in contrast to those of longer duration whose ratings were more satisfactory. Even here, however, the short duration of the history was not always the major cause for the lower rating.

Needless to say, the value of the peer judgment approach to evaluation depends in no small part on the surveyor who has been chosen to review the records. In the period between the two Teamster studies, a considerable amount of time was expended in experimenting with different surveyors, with different types of evaluation forms, and with the usefulness of patient summaries to the surveyors.

It was thought that perhaps a case summary prepared by a nurse would facilitate a surveyor's orientation, particularly in excessively large records. The consensus of the surveyors was that such summaries, regardless of how well prepared, tended to interpose a completed picture between the clinician and the record and influence his train of thought as he followed the progress of any given patient. There was very strong feeling on the part of all the physicians involved in these projects that only through a detailed personal review of an entire medical record could a judgment on the quality of care be made with assurance; nurses' notes, in particular, were of value. Therefore, regardless of the cumbersome nature of the process of obtaining photostatic copies of records in their entirety, and the additional time and money required, all surveyors felt that there was no place for nonphysicians in the preparation of the material to be reviewed if the objective was to be the judgment of clinical performance.

Evaluation forms ranged from those where only over-all care was designated either *satisfactory* or *unsatisfactory* to those where many items of patient care were individually rated. There was experimentation with a supplementary evaluation sheet containing questions highlighting certain areas of patient care—adequacy of reporting, diagnostic formulation, and therapeutic management—of both major and minor aspects of the problem presented by the patient. These answers were examined in relation to the final judgments, but no consistent patterns emerged. In the end it was concluded that the simplified form

was preferable and that its most valuable component was the narrative justification given by the surveyor for his rating.

In any such method which does not attempt to deny the possible existence of a strong subjective element, it is obvious that the status of the reviewer as seen by his professional colleagues is very important. Younger physicians would have been more readily available, in relation to both time and money; but it was felt that regardless of how competent they might be, their findings would be subject to question where ratings of *unsatisfactory* were given. This is not to say that all senior physicians with outstanding reputations make good surveyors. One outstanding New York clinician and teacher, instead of basing his evaluations on how he would have handled a particular case himself, would invariably address himself first to "Was this patient helped by his hospital stay?" (which would bring up the problem of judging outcome again), and then to the even less desirable: "Was this patient harmed by his hospital stay?"

Problems were also encountered with the type of physician who found the tradition of not criticizing professional colleagues so strong that he was unable to give an unfavorable rating, even when his summaries contained references to inappropriate and ill-advised procedures identical to those made by surveyors who gave unsatisfactory ratings. Surveyors with highly specialized interests that interfered with an over-all view of the patient's problems were also occasionally encountered; they concentrated on the aspect of the disease that was of interest to them and tended to ignore important concomitant events.

The surveyors who were considered most suitable for this activity were generally senior clinicians, with a teaching appointment in one of the city's medical schools, and were also engaged in clinical practice. The majority had extensive experience with the review of clinical records either in a supervisory capacity in their hospital activities or as surveyors for the Joint Commission on Accreditation or both. They were able to summarize the essential aspects of patient care logically and concisely and to base their conclusions on the evidence presented and not on assumptions of intent. They concentrated on the interaction between the physician and the disease process and were not distracted by data or occurrences that were irrelevant to the assessment of diagnostic formulation and therapy.

Discussion and Summary

The "medical audit," as described in the foregoing pages, is by no means a perfected tool. Considerably more must be done to strengthen and to document the judgments which are the basis of this type of approach. This is particularly true before such studies can be more widely applied, as has been foreshadowed by the increasing interest of government in this area. However, within a defined setting, such as the organizations described, such studies can be used to focus attention on problem areas and aid in the redefinition of goals to be reached in obtaining high quality medical care. Their administrative usefulness by now seems beyond question.

The medical audits described concentrated on the professional performance of the attending physician as judged by a clinician surveyor in the review of a medical record. The same approach can be used to analyze other objectives if careful attention is paid to the study design. During the period of the HIP studies, one of the very strong teaching hospitals in the city requested that a review be made of the records in their medical outpatient department by the same technic. After a review of some 40 cases selected from the medical department's file, it was felt that the study was of little value in terms of ultimate recommendations; not one case received less than the allocated 100 points. Al-

APPENDIX A—QUALITY OF MEDICAL CARE STUDY

Example of the 4"x6" Card Used for Recording Data
by The Health Insurance Plan of Greater New York

Patient # _____ Medical Group _____

Age _____ Sex _____

Physician Studied _____

1 Diagnosis _____ Justification {Interviewer: G-F-P / Staff : 5-2-0

2 Months Under Care ____ No. of Visits _____

ITEM #		Rating by Interviewer G, F, or P	Weighted Numerical Equivalents given by staff to ratings G, F, and P					
			IF G	F	P			
3	History		8	4	0			
4	Physical		7	3	0			
5	Progress Notes		5	2	0			
6	Reports		5	2	0			
	DIAGNOSTIC PROCEDURES		IF consultations indicated			IF consultations not indicated		
7	Time Involved		2	1	0	4	2	0
8	Indicated Lab Work		6	3	0	8	4	0
9	Indicated X-rays		6	3	0	8	4	0
10	MANAGEMENT		20	10	0	20	10	0
	CONSULTATIONS							
11	Requested		6	3	0			
12	Replies							
	TREATMENT							
13	Acceptability		8	4	0			
14	Indicated Lab Work		7	3	0			
15	Follow-up Visits		10	5	0			
16	PATIENT CARE		5	2	0			
Signature of Interviewing Physician		Total Score	MAXIMUM: 100 points					

most as an afterthought, a second review was instituted: this time cases were selected at random from the admitting office of the outpatient department. This one change gave a considerably different picture of patient care. A major weakness was shown to exist by the lack of coordination between specialty departments. For example, a woman was followed for more than six years by the ophthalmology department where she went frequently for new glasses. Funduscopic reports over the years showed increasing evidence of degenerative changes and vascular disease. However, it was not until the sixth year, when

the woman suffered a stroke, that she came to the attention of the medical department.

By similar shifts in case selection this method of physician appraisal of clinical handling from medical records can be used to meet different objectives. The content of the reviews can be compiled to examine both administrative and professional policies. Weighting or scoring has advantages when large numbers of cases are involved and the study objectives are better served by presenting judgmental findings in a graduated scale; however, the arbitrary nature of such weights can be difficult to defend. In operating programs, particularly, where cost is always a consideration, the use of one surveyor would seem to suffice; the problem of clinical differences of opinion between highly trained physicians in the same field does not affect the over-all findings and recommendations.

This does not imply that operating agencies should use such methods exclusively for quality control. However, periodic use of the appraisal method can either examine in greater depth areas suggested by variation in reported indexes of operation or suggest other areas where monitoring by indexes is indicated.

REFERENCES

1. Esselstyn, Caldwell B. Principles of Physician Remuneration. Papers and Proceedings of the National Conference on Labor Health Services: Washington, D. C., June 16-17, 1958. Washington, D. C.: American Labor Health Association, 1958, p. 122.
2. Lembcke, P. A. A Scientific Method of Medical Auditing. J. Am. Hosp. A. 33:i–ix (June 16 and July 1), 1956.
3. Makover, Henry B. Quality of Medical Care: A Study of the Medical Care Provided in 1948 and Early 1949 by the Twenty-Six Groups Associated with the Health Insurance Plan of Greater New York. New York: Health Insurance Plan of Greater New York (July), 1950.
4. Daily, E. F., and Morehead, M. A. A Method of Evaluating and Improving the Quality of Medical Care. A.J.P.H. 46,7:848–854 (July), 1956.
5. Morehead, M. A.; Daily, E. F.; and Sharpiro, S. Quality of Medical Care Provided by Family Physicians as Related to Their Education, Training, and Methods of Practice. New York: Health Insurance Plan of Greater New York (May), 1958. (Unpublished, due to the not unjustifiable position of the medical groups that no other insurance plan was publishing data on the number of physicians performing at unsatisfactory levels.)
6. Peterson, Osler L., et al. An Analytical Study of North Carolina General Practice, 1953-54. J. M. Educ. 31:12 Part Two (Dec.), 1956.
7. Butler, J. J., and Quinlan, J. W. Internal Audit in the Department of Medicine of a Community Hospital. J.A.M.A. 167:567–572 (May), 1958.
8. Trussell, Ray E.; Morehead, M. A.; Ehrlich, J.; et al. The Quantity, Quality and Costs of Medical and Hospital Care Secured by a Sample of Teamster Families in the New York Area. New York: Columbia University School of Public Health and Administrative Medicine, 1962.
9. Morehead, M. A.; Donaldson, R. S.; et al. A Study of the Quality of Hospital Care Secured by a Sample of Teamster Family Members in New York City. New York: Columbia University School of Public Health and Administrative Medicine, 1964.
10. Baumgarten, H., and Trussell, Ray E. Teamsters Plan Next Step in Health Care. Mod. Hosp. (May), 1962.

Dr. Morehead is associate director of special research, Montefiore Hospital, Bronx, New York City, and adjunct assistant professor in administrative medicine, Columbia University School of Public Health and Administrative Medicine (630 West 168th St.), New York, N. Y. 10032.

This paper was submitted for publication in December, 1966. It was prepared at the request of the Program Area Committee on Medical Care Administration of the American Public Health Association for their Conference on Appraisal of Quality and of Utilization of Services, which will provide the basis for preparation of Volume III of "A Guide to Medical Care Administration."