Evaluation is an increasingly important, time-consuming part of the medical care system. Use of explicit criteria in making such judgments has been recommended to promote consistency and fairness. This paper extends the use of criteria to screening so as to reduce the time physicians spend in evaluation. Can criteria be explicit and complete enough so that a suitably trained non-physician can use them with acceptable accuracy? If so, physicians could concentrate on cases selected by the non-physicians, rather than on cases arbitrarily or randomly selected.

Screening for Utilization Review: On the Use of Explicit Criteria and Non-Physicians in Case Selection

John O. McClain, Ph.D. and Donald C. Riedel, Ph.D.

Introduction

Although various programs for patient care appraisal have been used in hospitals and related health facilities on a voluntary basis in this country for many years, the conditions of participation under Medicare and Medicaid[1] now require formal utilization review procedures of a particular type, and the Joint Commission on Accreditation of Hospitals specifies the need for a utilization review "process" in each institution seeking accreditation. Debate continues as to whether these programs should be essentially educational, punitive, or deterrent in nature and obviously the characteristics of the "process" would vary according to the objectives of the review activity. While it is beyond the scope of this paper to argue the merits of these different objectives, it can be stated that a comprehensive system of patient care appraisal would employ three fundamental components: *collection* of basic demographic and clinical information on all patients, *statistical analysis and display,* and *case review* for selected patients, where indicated. The tremendous number of physician man-hours spent on routine tasks in the review process has underscored the need for further research on efficacious methods of case selection for detailed review and the use of non-physicians in some aspects of the review process.

Selection of Cases for Individual Review

There are a number of methods by which cases may be selected for review. One possibility is to review the care received by every patient. Such a task would be inordinately time consuming, and hence some means for sampling either patients or records is indicated. Random selection accomplishes a reduction in numbers, but may miss many of the cases that are worthy of review; on the other hand, the fraction of cases deemed inappropriate among those reviewed is an unbiased (statistically speaking) estimate of the fraction inappropriate in the entire population.

In order to increase the proportion of inappropriate cases among those that are reviewed, one may institute a purposive method of case selection, identifying categories that intuitively seem likely to contain disproportionate numbers of cases of inappropriate care or inappropriate utilization. Selection of all cases within certain diagnoses may accomplish this if there is reason to suspect that certain inappropriate practices occur more frequently in some medical situations than in others. Selection by other attributes such as length of stay may be instituted on similar grounds. Reviewing all cases of "extended duration," for example, is one of the requirements under Medicare and Medicaid. Other possibilities include selection of very short stays, or review of other selected points in the length of stay distribution.

All of these "purposive" selection devices have the disadvantage that they exclude review of large categories of patients (e.g., those with "normal" length of stay) and do not allow any meaningful statements to be made about the rate of inappropriate utilization for the hospital as a whole. Moreover, there is no guarantee that these methods will actually accomplish the goal of concentrating the inappropriate cases into the sample selected for review. These disadvantages may be overcome, however, by selecting a random sample from those records not initially selected for review by the purposive technique. A measurement of the rate of inappropriate utilization among these records will allow a test of the effectiveness of the purposive sampling. At the same time, the two rates may be combined in accordance with the sampling fraction to yield an estimate of the overall rate of inappropriate utilization.

The different methods mentioned above may be used in combination. For example, different length of stay levels may be specified for different diagnostic categories. Additional variables such as age, sex, and type of treatment may be used to specify differences among categories of patients. The number of different categories of patients proliferates quite rapidly as more descriptive variables are added, requiring some standard routine for setting the length of stay cutoff levels. Furthermore, there may be more catego-

ries specified by the various combinations of the variables than are meaningful, so that some method for determining which categories should be combined is also desirable.

Although the categorization of patients may make the selection process more sensitive to differences between patients, the use of a single criterion variable such as length of stay may be insufficient to separate appropriate and inappropriate utilization with any degree of success. This method could be extended by using other variables as criterion variables in concert with length of stay. An alternative is to specify medical criteria which the care of a patient should meet.[2] The problem with the latter is that the use of criteria might necessitate that a physician carry out the selection of cases, which defeats the purpose of the screening procedure. Alternatively, one might investigate the feasibility of programming a computer to apply the criteria, or training non-physicians in case selection according to the criteria. However, use of medical criteria is hampered by the form of the medical record employed in most hospitals; it is organized in a manner that is difficult for a physician to decode, let alone a non-physician.

Research is currently being conducted at Yale University to develop and evaluate alternative models of patient care appraisal. One of the projects has as its objective the construction of a utilization review case selection (i.e., screening) program which makes use of both statistical screening and screening based on preestablished medical criteria applied by non-physicians. This paper describes a study undertaken to make a preliminary test of the practicability of such a multidimensional screening program.

The Study—Non-Physician Case Screening

The goal of the study reported here was to elicit, from physicians, criteria by which a non-physician can make utilization review screening decisions. Gall bladder disease was selected as a test situation for two reasons: 1) it is a common disease, with well-established methods of treatment. Therefore, if the goal is not attained in this instance, it is not likely to be attained for any disease, using the methods of this study; 2) gall bladder disease is sometimes treated by internists rather than surgeons, which allows one to sample from a greater variety of points of view.

Physicians affiliated with the Yale-New Haven Hospital were approached with the permission of the appropriate chiefs of service. Participation was voluntary. The seven physicians who became subjects for this study represented both surgery and internal medicine, both private practice and full-time faculty at the Yale University School of Medicine. Six of the seven participants had no previous utilization review experience.

In order to attain some degree of uniformity in the experience of the participants vis-à-vis utilization review, the physicians were given a booklet, published by the American Medical Association,[3] containing selected articles about utilization review. A questionnaire was submitted to each physician, asking him to decide what areas of medical care should be included in utilization review, and how difficult it would be to collect the necessary information.

As a further preparation, they were asked to make utilization review judgments, retrospectively, by studying abstracts of 20 medical records. Each of the abstracts was submitted twice to each physician, for a total of 40 judgments per physician. These abstract forms were developed by another investigator;[4] their use allowed the physicians to carry out their judgment task more quickly, although there were some serious initial complaints about the validity of the abstracts. These complaints, together with the criteria gathered later in the study, were used to improve the abstract form for use later in the study.

Table 1—Categories of Utilization Review Criteria Gathered in this Study

1. Admission	7. Pathology report
2. History/physical	8. Operating room X ray
3. Biliary X ray	9. T-tube X ray
4. Other services	10. Post-operative complications
5. Pre-operative length of stay	11. Timing of discharge re antibiotics
6. Surgery	12. Post-operative length of stay

The Interviews

The criteria were collected primarily by interviewing the seven participating physicians. As a guide, criteria developed at the University of Michigan[5] were used in the first interview with the first physician. As new criteria were collected they were utilized in succeeding interviews to elicit comments from the physicians. There were two or three interviews with each doctor, resulting in a much more extensive list of criteria than the original set. This single list of criteria could be made specific to a physician by crossing out the criteria with which he disagreed. The criteria fell into twelve categories (Table 1) covering the various stages and aspects of the case and treatment of gall bladder disease.

At times, the considerations become too complicated for translation into criteria appropriate for non-physician screening. In those situations, the physician being interviewed was asked to invent a rule of thumb that would allow the non-physician to act correctly most of the time. Thus, the resulting criteria may be viewed as *approximate models* of the utilization review process of the participating physicians. A further report of this aspect of the study has been published elsewhere.[6]

After the interview sequence was concluded, each physician was given a questionnaire asking him to assign weights to reflect the relative importance of each category to his overall decision. Between zero and one hundred points were assigned to each category by making a mark on a linear scale. Before returning the questionnaire, each physician was asked to make a visual comparison of his twelve answers, modifying them as he saw fit.

Besides the interviews and questionnaires described, there was some written and telephone communication between the investigator and the physicians concerning revision and interpretation of the criteria. Beyond that, the remaining experimental effort consisted of: 1) writing a set of rules by which a non-physician could apply the criteria; 2) having a non-physician judge a set of abstracted medical records, according to the criteria and rules; and 3) com-

Table 2—Screening Accuracy on the Validation Sample (20 Records, Seven Physicians)

	Fraction rejected by the model		
	.15	.30	.45
Fraction positive of those rejected	.86	.71	.52
Fraction negative of those not rejected	.76	.83	.82
Hit rate = fraction correct	.77	.79	.67

Total number of judgments = 140
Total number of positive judgments = 47 (34%)

Table 3—Screening Accuracy on a Measure of Group Opinion (On a Sample of 20 Abstracts)

	Fraction rejected by the model (equal to fraction condemned by physician vote)		
	.15	.30	.45
Required positive votes*	5/7	4/7	3/7
Fraction condemned† of those rejected	1.00	1.00	.78
Fraction not condemned of those not rejected	1.00	1.00	.82
Hit rate = fraction correct	1.00	1.00	.80

* The number of records condemned (positive) by group opinion is determined by vote. The required number of positive votes to condemn a record is 5 out of seven in the first column, 4/7 in the second, and 3/7 in the third.
† Condemned by physician vote, that is.

paring the non-physician's judgments to those of the physicians he is supposed to be mimicking. The third step required that a new set of abstracts be judged by the physicians (a validation sample), since the criteria were developed and modified partly on the basis of the first set of judgments.

The detailed description of how the non-physician's comments on a record were converted into a screening decision will not be repeated here, for sake of brevity. The interested reader will find in reference[6] that some of the manipulations were based on a comparison with the validation sample. However, it was argued that the nature of the manipulations was such that the validity test retains its meaning, albeit with reduced confidence in the results.

Results

Because the effort reported here is dual in nature (a test of non-physician screening and a test of a modeling methodology as applied to the utilization review situation), the results are reported, with different emphasis, in more than one professional field (see reference 6). The accuracy with which the non-physician could mimic the utilization review physician was studied by means of a rather involved correlation analysis. The results[6] showed that there was no discernible difference among the seven sets of criteria, as applied to utilization review screening. That is, the differences of opinion (which were demonstrated to exist between physicians) were not captured in the criteria. For this reason, the seven "models" (model = criteria + rules for applying them + non-physician applying them) were combined into a single screening device by averaging their numerical outputs. In this way, the judgments of the non-physician, based on the criteria given by the physicians, were used to rank the 20 medical record abstracts.

The ranking of the records was used to test the validity of the use of a non-physician for screening cases for utilization review. This was accomplished by selecting a cutoff point in the ranking, such that all records below that point were rejected by the screen (they would be sent to a physician for further review), whereas records above the point would be considered acceptable, according to the criteria specified. This categorization was tested against the actual decisions of the physicians in two ways.

> Definition: A physician's judgment on an abstracted record is said to be *positive* if either (a) one or both of his test-retest judgments said "inappropriate" or (b) he indicated that the record needed further review.

According to this definition, it is the task of the non-physician screener to identify records where there is some question of appropriateness.

The first test was to compare the records rejected by the screen to those called positive by a physician. This was done separately for each physician, and the number of correct and incorrect screening decisions were added across physicians before being converted to percentages. The results are shown in Table 2 for three different cutoff levels in the screen's ranking. The second column of Table 2 shows that the screening accuracy was 79% when the bottom 30% of the cases ranked by the screening model were considered to have been rejected by the screen.

The second test was to compare the screen's ranking to a measure of group opinion. In this test, each record was given a "group score" equal to the number of positive judgments it received from the seven physicians. In order to decide which records should be rejected by the screen, someone must decide how many physicians must say "positive" to condemn a record.[7] This issue was finessed by trying several cutoff levels on the scale of physician group score. The number rejected by the non-physician screen was then set to match the number rejected by the group. The results are shown in Table 3. The second column of Table 3 shows that the screening model and the majority vote of the physicians were in complete accord as to which cases represent the bottom (worst) 30%.

When these results are contrasted to those of Table 2 (e.g., 79% accuracy on individual physician judgments contrasted to 100% on the vote) one concludes that the criteria were apparently capable of detecting the records where most physicians agreed that care was inadequate (the group opinion measure), but were less capable of making determinations when there was disagreement. This is supported further by noting, in Table 4, that the records on which the ranking of the screen was most out of line (cases number 3 and 4) also had 3 positive and 4 negative judgments—very high disagreement.

Table 4—Ranking of the Cases by the Screen vs. Number of Positive Judgments Given by the Physicians, for each of the 20 Cases

	Case number									
	1	2	3	4	5	6	7	8	9	10
Ranking by screen	1	1	1	1	5	6	7	8	8	10
No. of positive judgment	1	1	3	3	1	0	1	1	0	1
	Case Number									
	11	12	13	14	15	16	17	18	19	20
Ranking by screen	11	12	12	14	15	16	17	18	18	20
No. of positive judgments	2	0	0	3	4	4	4	6	7	5

The results reported so far have been based on a single non-physician (the first author) screening for the seven physicians. In order to test the generality of the results, another non-physician (the person who developed a medical record abstract) was asked to repeat the screening process on the same records. Based on the correlation analyses reported in reference,[6] there was no difference in the accuracy of the two screeners, and the correlation between the two screeners (.71) was higher than the average test-retest correlation of the physicians (.52). The hit rate analysis reported above was also repeated for the second screener, again with no important differences.

In order to decide how accurate a model or screen must be to be useful, it is important to compare the accuracy of the model to the reliability of the reviewing physicians. Table 4 shows that the degree of disagreement among physicians was by no means inconsequential. In six out of twenty records, the vote was 4 to 3 or 3 to 4 as to whether the case was positive. An additional measure of reliability is available through the test-retest methodology. In this study, each physician judged each abstracted case *twice*, with a separation of one or more weeks between test and retest. In making their judgments of appropriateness of utilization (exclusive of whether the record was judged to need additional review), the physicians reversed themselves an average of 6.4 times (16%) out of the 40 records they saw in this study. The range was from 3 to 11. Further analysis of the reliability issue may be found in reference.[8]

Discussion and Summary

The intended use of the models described in this article is to aid in the selection of cases for utilization review. Success is defined as being able to select a small sample that is concentrated in cases where there is reason to believe that utilization practices were inappropriate. To test this, the "judgments" of the models were compared to judgments of physicians. Unfortunately, physician judgment is not an ideal criterion, since there is inconsistency both between physicians and between repeated judgments of the same physician on the same record.

The results reported here show clear indications that the models can be used for screening in the diagnostic category studied. The major limitation seems to be that the differences of opinion between physicians are not well-represented by the criteria gathered. Several possible explanations exist for this limitation. If these differences were on complex medical issues, perhaps they got lost in the "approximation approach" used in collecting the criteria. It is also possible that there were certain items that the physicians did not discuss with the interviewer. Additional evidence will be available on this question, since the next application of this method is being carried out by a physician. The question of generality is also being addressed by extending the application to illnesses other than gall bladder disease.

The study of the use of explicit criteria and non-physicians in medical evaluation has potential impact on areas other than utilization review in hospitals. In fact, criteria are potentially useful in any situation where an assessment needs to be made, particularly when there is continuing or periodic assessment. Suppose, for example, that a tissue committee chose to specify a set of criteria. Using medical students to apply these criteria to cases would yield double benefit—the students would learn first hand about appropriate and inappropriate practices, and the tissue committee could work on cases that had already been screened and commented on according to their own criteria.*

Predetermined criteria can also be useful in assessment of the need for facilities throughout the spectrum of the medical care system, from home care to the acute hospital.† Publication of these criteria might also lead to more effective placement of patients, since the institutions

* Thanks to R. Touloukian, M.D., Yale University School of Medicine for this suggestion.

† The Genesse Region Health Planning Council used explicit criteria in this way in the Rochester, N.Y. area.

and professionals involved would then know exactly how the planners intended that the facilities be used.

Whatever the application, considerable work is involved in obtaining criteria explicit enough for a suitably trained non-physician to apply. However, the payoff in the long run is a considerable reduction in the time that physicians must spend in the review process.

References

1. Conditions of Participation of Hospitals, Federal Health Insurance for the Aged Regulations: Section 405.1035. Condition of Participation—Utilization Review Plan. Baltimore, Maryland: Social Security Administration, 1966.
2. Payne, B. C. Continued evolution of a system of medical care appraisal. JAMA 201:126 August 14, 1967.
3. Utilization Review, A Handbook for the Medical Staff. Chicago: American Medical Association, 1969.
4. Pallett, P. J. The Development of a Mechanism for Abstracting Medical Records for Utilization Review. M.P.H. thesis. Yale University, 1969.
5. Hospital Utilization Review Manual. Ann Arbor, Michigan: The University of Michigan Medical School, Department of Postgraduate Medicine, 1968.
6. McClain, J. O. Decision Modeling in Case Selection for Medical Utilization Review. Management Science. Vol. 18 No. 12, 1972.
7. McClain, J. O. On a rule for group decision-making. Medical Care 7:406 September-October 1969.
8. McClain, J. O. Physician Confidence and Reliability in Utilization Review. Medical Care. Vol. 10, No. 6, November-December 1972.

Dr. McClain is Assistant Professor of Quantitative Analysis, Sloan Institute of Hospital Administration, Graduate School of Business and Public Administration, Cornell University, Dr. Riedel is Professor of Public Health (Medical Care) Department of Epidemiology and Public Health, Yale School of Medicine, New Haven, Connecticut.

This study was supported in part by the following sources: Contract number HSM-110-89, Community Health Service, HSMA, Department of Health, Education, and Welfare; The Connecticut Regional Medical Program; Yale-New Haven Hospital (data). This paper was submitted for publication in June, 1971.

Nominations Invited for Browning Award

Nominations for the 1973 Edward W. Browning Achievement Award for outstanding contribution in the prevention of disease are being sought by the American Public Health Association. This prestigious award, established in 1971, is one of a five-part annual award for distinguished international achievement in five major areas—each of which is overseen by the following international organizations or groups:

Prevention of Disease, American Public Health Organization
Conserving the Environment, Smithsonian Institution
Improvement of Food Sources, American Society of Agronomy
Alleviation of Addiction, International Council on Alcohol and Addictions
Spreading of the Christian Gospel, a group of international religious leaders

The late Edward W. Browning conceived of the awards over 60 years ago when he was at the height of his career as a colorful and successful real estate entrepreneur. Mr. Browning, who had a profound interest in the "well being and happiness" of mankind, hoped that these awards would stimulate public concern in "religious, moral, social, economic, and intellectual" endeavors.

The Browning prize for preventive medicine was awarded for the first time at APHA's Annual Meeting in Minneapolis. The recipient was B. Russell Franklin, former chief of environmental health inservice training for the Philadelphia Department of Public Health. Last year's winner, honored during APHA's Centennial Meeting in Atlantic City, was E. Cuyler Hammond, Sc.D., vice president of the American Cancer Society. Selection of each year's nominee is made by APHA's five-member awards committee appointed by the Executive Board. Each award consists of an honorarium of $5,000 and a medal, bearing the likeness of Mr. Browning, the founder.

Nominations for the Browning Award may be made by any Association member. They must be accompanied by a biographical sketch of the candidate and reasons for the nomination, and should be sent to Browning Awards Committee, American Public Health Association, 1015 18th St., NW, Washington, D.C. 20036. Nominees must be living, and the recipient of no other Association awards in 1973. Nominations must be postmarked by May 15, 1973 to be considered for the 1973 Browning Award. This year's award winner will be honored at APHA's 101st Annual Meeting in San Francisco, Nov. 4-8. The Edward W. Browning Achievement Award is administered by the New York Community Trust, New York City.