Evaluating and improving the measurement of hospital case mix

by Stephen F. Jencks, Allen Dobson, Patricia Willis, and Patrice Hirsch Feinstein

The foundation of case-based prospective payment is the case classification system. The purpose of classification systems is to group together patients with similar treatment requirements. The systems described in this issue take a variety of theoretical and practical approaches to classification. The critical issue in comparing these systems is whether the variation in treatment requirements which is not explained by the classification system is associated with particular groups of patients, particular hospitals, or particular groups of hospitals in such a way as to result in unfair reimbursement. We suggest criteria for comparing classification systems and a research agenda for clarifying the fairness of different approaches.

This issue of *Health Care Financing Review* is devoted entirely to the accuracy of hospital case-mix measurement systems for reimbursing hospitals. We have brought together eight articles by active workers in this field, representing the most promising efforts to improve case-mix measurement. Although severity of illness is an organizing theme for these articles, their scope is much broader. The purpose of this paper is to give the reader four perspectives for reading these articles:

- An understanding of what case-mix measurement is and of the major approaches which are now being explored.
- An understanding of the importance of case-mix measurement to the Health Care Financing Administration (HCFA).
- An analytic framework which can be used to describe and compare different approaches to case-mix measurement.
- An outline of the research directions which appear likely, at this time, to be most useful to HCFA over the next five to ten years.

The nature of case-mix measurement

Case-mix measurement, as we discuss it in this issue of the *Review*, is the characterization of an inpatient stay in a way which permits us to predict the resources which are needed for care during the stay. This issue focuses on case-mix measurement as a tool for determining appropriate payments to hospitals, rather than for such uses as determining appropriateness of admission, quality assurance, hospital management, or epidemiology (Hornbrook, 1982).

For the purpose of paying hospitals, the accuracy of a classification system is measured by its ability to classify a case with other cases which have the similar costs for necessary care. When a case-mix system is used for payment, three steps are necessary: the hospital case must be placed in a class by the classification system; the class must be given a weight which indicates resource requirements relative to other classes of cases; and the system must assign multipliers which convert these weights to appropriate dollar prices. This paper and this issue of the *Review* are primarily concerned with the classification system rather than the weights and multipliers. However, to assess a classification system's ability to predict costs, it is necessary to calculate weights and compare them to observed costs, so the calculation of weights receives some consideration.

In reading about classification systems, it is useful to keep in mind that the data routinely collected about hospital inpatient care is the Uniform Hospital Discharge Data Set (UHDDS), a set of data elements found on the cover sheet of almost all medical records of a hospital stay. The UHDDS comprises age, sex, discharge diagnoses, procedures performed, type of place to which discharged, and whether the discharge was alive, dead, or against medical advice. Under UHDDS rules, the diagnosis which is listed first is the "principal" diagnosis—that diagnosis which, after study, is determined to be principally responsible for causing the admission. These data elements are routinely collected by Medicare and many other payors for all hospital admissions and are therefore routinely available in automated form for case-mix classification. Whether a classification system uses UHDDS data is obviously a matter of considerable practical importance, since other data is not routinely available either to test a system or to set weights.

This issue reports six major current classification strategies:

Diagnosis-related groups

The diagnosis-related group (DRG) system (Smits, 1984) has been developed at Yale University under John Thompson and Robert Fetter. DRG's are also being refined at Health Systems International, Inc. under direction of Richard Averill. DRG's are now used by Medicare and some other payors as a basis for payment and are the only classification system in

use for payment. DRG's are therefore the starting point for discussions of classification. The system operates on UHDDS data. The DRG system classifies a case into one of 23 major diagnostic categories (MDC) on the basis of the principal diagnosis. Once assigned to an MDC, the case is then further assigned to one of 467 DRG's on the basis of presence or absence of certain procedures, age of the patient, specific principal diagnosis, presence or absence of a significant comorbidity or complication, and, in a few cases, whether the patient died or was transferred to another hospital. For reimbursement there is an additional "wastebasket" for cases with surgery which is inconsistent with the principal diagnosis. The DRG's were developed using clinical judgment in concert with statistical analysis of a large number of hospital discharges.

Disease staging

Staging (Conklin, 1984) was developed at Jefferson Medical College under Joseph Gonnella's direction and is being further refined at SysteMetrics, Inc. by Daniel Louis and others. Staging has several forms, but in this discussion we consider only the computerized version which uses UHDDS data. Staging assigns each diagnosis to one of more than 400 conditions and then to one of 4 stages (5 in the case of cancer) of progression of disease within the condition category, depending on the exact diagnosis recorded. The stages are analogous to stages for cancer and reflect stage of progression from minimal disease to death. Unrelated secondary diagnoses are also staged. Under certain circumstances the computer program will reorder diagnoses if the principal diagnosis is vague or if another diagnosis appears likely on clinical grounds to be the principal diagnosis. The staging criteria were defined by specialists from their clinical experience rather than by consulting actual data on length of stay or resource consumption. Since a given stage of one disease does not predict the same resource consumption as the same stage of another disease, stage must be considered in conjunction with the diagnosis staged, giving a potential total of more than 1,700 classification groups. Staging makes use of secondary diagnoses in several ways, but other elements such as age, procedures, and transfer are considered in only a very small number of conditions (e.g., Caesarean section as a procedure is used to determine the type of delivery). A method for combining independent staged disorders in a single patient, which would be needed for reimbursement, is not yet automated.

Patient management categories

Patient management categories (PMC) (Young, 1984) has been developed at Blue Cross of Western Pennsylvania under direction of Wanda Young. The current, computerized version of PMC classifies patients into disease groups according to key diagnosis codes and then into a specific PMC. The categories were developed by panels of physicians, based on their clinical experience rather than on analysis of data. Each category is intended to represent a distinct type of illness requiring a distinct type of management. PMC places special emphasis on using multiple diagnoses to define the specific management category to which a patient is assigned; procedures are sometimes used to define categories when the diagnostic system does not have sufficient specificity. Within a disease group, a patient is assigned to the PMC which is expected to be most resource-intensive, but a patient could be assigned to PMC's in two or more disease groups. A method for combining multiple grouping for reimbursement is under development.

Severity of Illness Index

The Severity of Illness Index (Horn, 1984) has been developed at Johns Hopkins University under direction of Susan Horn. The severity score is assigned on a four-point scale using an "implicit" synthesis of seven sub-scale ratings, each of which is derived from chart reading. Rating of the subscales is done subjectively by trained raters who have had standardized training and use a system of benchmarks for the rating scale. Scoring has not yet been computerized although computerization is planned. In some studies, including one reported in this issue, the severity rating is further broken down according to a rating of the significance of surgery performed. Severity has been developed through an extensive process of testing against actual patient data in a variety of settings but has not been tested on national samples because of the cost of scoring the necessary number of cases. Because severity scoring is both the most ambitious and the most methodologically controversial technique, it receives special attention in this paper.

Acute physiology and chronic health evaluation

The acute physiology and chronic health evaluation (APACHE II) (Wagner, 1984) was developed at George Washington University under direction of William Knaus and Douglas Wagner. APACHE II uses admission values of 12 physiologic variables (acute physiology score or APS) such as blood pressure in addition to information about chronic health

conditions to produce a continuous variable; it is the only system which assigns cases a score on a continuous scale rather than assigning a case to a category. APACHE II has been extensively tested on intensive care patients but not on other patients, and it has not been tested as a general predictor of costs. This system is of special interest as a possible severity of illness measure to be applied in conjunction with other classification systems. APACHE II requires physiologic information which is not available from UHDDS or most other computerized abstracts, but direct acquisition of most of the necessary information from computerized laboratories might be possible. The necessary data is usually in the patient's record.

Medical illness severity grouping system

The medical illness severity grouping system (MEDISGRPS) (Brewster, 1984) was developed at St. Vincent's Hospital in Worcester under Alan Brewster's direction in conjunction with InterQual, Inc. MEDISGRPS uses admission values of a set of physiologic variables and clinical and X-ray findings to give the patient an admission severity score on a five-point scale. A second set of measurements is recorded a fixed time after admission, as further information for patient classification. MEDISGRPS was developed as a quality assurance tool and has been implemented at a number of hospitals for this purpose, but it is at an early stage of development as a reimbursement system. MEDISGRPS requires physiologic information which is not available from UHDDS. Medical record personnel need special training to abstract the data, which is usually in the patient's medical record.

HCFA's interest in case-mix analysis

As we pointed out earlier, this issue of the *Review* addresses case classification as a tool for hospital payment. A hospital's case mix is often considered its product in an economic sense, but a case is obviously what a hospital treats, while the care provided is the product. Nevertheless, if the variables used to classify cases in the case-mix system also determine the necessary diagnostic and treatment activities, case mix is an excellent proxy for care provided. We will discuss this important condition later in this article.

In the hospital prospective payment system (PPS) which Medicare has just put in place, case mix, as measured by the DRG system, is used as the measure of the hospital's product for purposes of payment. Under PPS a price schedule is published before care is rendered and the hospital is (in principal) paid an amount which is determined by the DRG into which the case is classified on discharge. PPS replaces a system in which reimbursement was based on retrospectively-determined costs. A classification system is essential to operating a prospective payment system based on per-case payment, since the classification of the case determines the payment.

In thinking about the use of classification in PPS it may be useful to think of PPS as actually comprising two separate reforms—paying for the case rather than for the individual services (this process of paying for larger aggregations of services is often called bundling) and determining the payment from a schedule published before care is rendered (prospective payment). The PPS reform has three distinguishable goals:

- To make Medicare's payments for hospital care predictable and controllable through prospectively determining payments for a unit of service (the admission) whose volume and composition has historically been relatively stable and predictable.
- To permit fair allocation of finite resources. Fairness means that different hospitals are paid comparable amounts for the care of like cases.
- To give hospitals incentives to provide care as efficiently as possible by eliminating the per-service retrospective payments which encouraged long stays and profligate use of services.

In the short term, the accuracy of case classification is far more important to the fairness of PPS than to its incentives for efficiency. Even if the DRG system is quite inaccurate, paying by the case creates strong incentives to shorten inpatient stays, to provide some services for the inpatient stay prior to admission or after discharge, and to provide fewer services. Even if the system does not classify together cases which require similar resources, it still encourages the hospital to treat each case as efficiently as possible.

On the other hand, if DRG's are not accurate, then reimbursement is not likely to be fair, since payments will not be well related to the resources necessary to treat each kind of patient. These inaccuracies could create perverse incentives to transfer or avoid some kinds of patients while selectively recruiting others. Such incentives could both disrupt the health care system and create access barriers for some Medicare beneficiaries. Furthermore, hospitals which continued to treat all patients in the face of these irrational incentives could face severe financial problems, not because of their own management shortcomings but because PPS did not reimburse them accurately. Such adverse effects could make the health care system less rather than more efficient.

Some investigators argue that there is so much flexibility and "fat" in the health care system that even fairly serious inequities will simply be absorbed. To put this argument differently, standard treatment will simply align with the resources available under the DRG system as health care providers respond to the PPS reimbursement levels. If a payment system resulted in every hospital being paid the same inadequate amount for comparable cases, this might be true, but such an outcome would reflect only errors in calculating weights and multipliers. If the classification system is inaccurate, hospitals will be paid the same amount for cases which differ clinically and

appear to clinicians to have different treatment requirements, and clinicians will probably be highly resistant to treating such patients in an identical fashion.

Because classification systems rely on relatively limited data to describe patients and generally classify patients into a limited number of categories, they are inevitably imprecise at the case level. Thus, it is not possible even in principle, to determine whether classification systems or case weights are accurate by looking at individual cases. The appropriate measure of accuracy is whether the patients in a DRG in one hospital have the same average treatment requirements as do patients in that DRG in other hospitals across the Nation. Since there is variation among cases, even the average treatment requirement for patients in a DRG in a hospital can only be expected to approximate national average treatment requirements for the DRG when the number of cases is large (the law of large numbers).

The concept of treatment requirements is important here. Treatment, obviously, is a shorthand term which includes diagnostic activity. More important, however, the concept of treatment requirements does not imply a treatment protocol; it simply implies the resources necessary to provide treatment. In the next section, when we consider economic neutrality, we touch on the problem of situations where there may be different costs for the same case.

The critical question in determining fairness of a classification system is whether hospitals can receive, select, or recruit patients who have either higher or lower treatment requirements than the national average for that DRG. This question can be put in a different way: Variation in treatment requirements does occur for patients within a DRG. Does that variation correlate with factors which can determine the hospital which a patient enters, or is variation "random" with respect to those factors? If the former is true, the system is either actually or potentially unfair. If the latter is true, the system is fair. Teaching hospitals, public hospitals, and specialty hospitals can easily demonstrate that their patients have distinctive characteristics (e.g., being referred, being poor, having special kinds of illness). The issue is whether these distinctive patient characteristics are correlated with the variation in treatment requirements within a DRG. If that correlation exists, then hospitals with disproportionate numbers of these patients will not be paid according to actual treatment requirements.

HCFA's most urgent research need in case-mix classification is to determine whether the DRG system is fair to classes of beneficiaries, to individual hospitals, and to classes of hospitals. Resolving this question is important because any problems which exist should be promptly identified and corrected. But demonstrating the fairness of the system convincingly is also important because erroneous beliefs that the system is unfair may create access problems for some Medicare beneficiaries. For example, if hospital managers believe that the poor, the frail elderly, or the disabled have greater treatment needs than other patients in the same DRG, then these groups of patients may have difficulty getting care. Such prejudices are especially likely to become a basis for action if a hospital finds its reimbursement under PPS falling from historical levels. It is therefore extremely important to the success of PPS both that the case classification system be fair and that it be perceived as fair for all classes of patients as well as for all kinds of hospitals.

When planning research in this area, it is useful to keep in mind that the accuracy of a classification system can be substantially compromised by a weighting procedure which does not accurately reflect the differences between cases or by multipliers which do not result in appropriate payments. Accordingly, it is important in assessing a system to assure that when tests which require weights, that the weights have been calculated accurately.

HCFA also has a broader interest in accurate case mix. Developing an accurate per capita price for a health maintenance organization or a competitive medical plan requires a sophisticated and precise measure to predict the needs of Medicare beneficiaries. Solving the problems discussed above supports development of capitated systems because a competitive capitation system is likely to need methods for setting rates which are more precise and sensitive than the methods of the current Medicare average annual per capita charge (AAPCC). The research agenda described later in this article should strengthen Medicare's capabilities in this area.

A framework for comparing case-mix systems

Comparing classification systems presents several difficulties, not least of which is the paucity of published information. Part of this scarcity is attributable to systems being very new (this issue contains the first journal publication on MEDISGRPS, and the computer program for PMC has just been completed). However, even for systems on which there are published articles in refereed journals, there are two significant problems. First, with the exception of DRG's, almost all published material is written by the developers of the system: a benchmark of scholarly evaluation is replication of results by investigators

other than the developers. Second, again with the exception of DRG's, there is no information on system performance using national data. Third, with the exception of Horn's work comparing DRG's and severity score, there are no published comparisons between systems.

Despite these extremely significant problems, we believe that it is useful to compare systems along three major axes: conceptual foundations, administrative implications, and empirical performance.

Conceptual foundations

Three conceptually important features of a classification system are the assumptions it makes about determinants of resource requirements, the economic neutrality of the system, and its clinical reasonableness. In addition, systems differ both conceptually and empirically in the degree to which they classify patients on the basis of care received as opposed to care actually required.

Determinants of resource requirements

Classification systems implicitly conceptualize determinants of resources required to diagnose and manage a patient. Among these determinants are the degree of diagnostic effort necessary, the nature and severity of the patient's condition on admission, the response of the patient to treatment, and the goals of treatment. Case-mix measurement systems differ in how many of these variables they recognize. For example, staging recognizes diagnoses but not variations in diagnostic needs, responsiveness to treatment, or treatment goals. DRG's include procedures, which can be a proxy for treatment goals, but DRG's do not consider diagnostic efforts or response to treatment. PMC addresses diagnostic efforts and treatment goals only indirectly, and does not consider treatment response. Severity of illness measures more variables (e.g., response to treatment) but does so less explicitly, while degree of diagnostic effort and goals of treatment do not seem to be considered. APACHE II and MEDISGRPS exclude diagnostic efforts or treatment goals, but are much more precise about severity of condition, and MEDISGRPS is more sensitive to response to treatment.

Economic neutrality

A system which neither rewards nor penalizes a hospital for performing an activity is economically neutral to that activity. For example, DRG's use major procedures to classify patients. If reimbursements are properly calculated for the DRG's with and without surgery, the difference in reimbursement is the cost of surgery, all other things being equal. Thus, the hospital is neither rewarded nor penalized for the performance of surgery and the provider is free to make this decision on clinical grounds. While large and small hospitals in Maryland have very similar case mix as described by staging, large hospitals have more complex case mix as described by DRG's. This presumably means that larger hospitals do more procedures for patients at a given stage of illness, since DRG's consider procedures while staging does not. One might argue that this represents over-utilization and should be discouraged or one might argue that it represents appropriate concentration of specialized procedures in specialty centers and should be treated with neutrality. If we adopted the former view, we would (if staging and DRG's are otherwise equal) adopt staging for case-mix classification since staging does not recognize procedures in classifying patients and thus creates an incentive for hospitals not to incur the costs of surgery. If we adopted the latter view, we would use DRG's for case-mix classification since DRG's are neutral to performance of procedures and therefore do not penalize a hospital for becoming a surgical referral center. The same argument can be applied, for example, to whether hospitals treat psychiatric patients in specialized units. If we regarded such treatment as potentially useful, we would classify it as a procedure and recognize it in the classification system; if we regarded it as wasteful, we would not recognize it in the classification system and there would be strong economic incentives not to use such treatment. These two examples suggest that the problem of neutrality is an intensely important policy issue. It is clear that neutrality with regard to equivalent treatments of different cost is a far-reaching issue since it raises the problem of the extent to which Medicare will pay for patient and physician preferences.

Clinical reasonableness

Clinical reasonableness means that a classification system classifies together cases which are clinically similar in their management. This is important for two reasons:

- A clinically reasonable system is likely to retain its power as relative costs of treatments change and even as clinical practices and technology evolve.
- A clinically reasonable system is more likely to be acceptable to administrators and physicians.

Clearly, the chart-reading methods of the severity score are the most closely connected to clinical data and therefore most clinically reasonable from that

perspective. The dependence on vital signs and critical findings in APACHE II and MEDISGRPS is highly reasonable for conditions where emergency admissions and critical illness are likely, but their applicability to elective surgery and noncritical illness is unclear. The detailed use of patterns of diagnoses (and sometimes procedures) in PMC, and to a lesser extent in staging, make them more clinically credible than coarser systems such as DRG's. However, many clinicians are suspicious of systems relying exclusively on diagnosis to determine what care is necessary.

Ability to distinguish necessary care from care rendered

A classification system which depends heavily on care actually rendered tends to return the payment system to cost-based reimbursement. Unfortunately, even diagnoses tend to depend on services rendered, since a large part of health care activity is diagnostic. The method most independent of care actually rendered is MEDISGRPS, which uses physiologic measures and presenting complaint and does not use specific diagnoses. APACHE II, which uses only chronic diagnoses, is also relatively independent of diagnostic effort. PMC's and staging depend on detail in diagnoses, and this may be correlated in some settings with either maintaining a teaching program or with extensive workups. DRG's depend heavily on whether an operating room procedure was performed although not on other elements of care, while PMC's and staging use procedures in a very limited way.

Administrative considerations

From an administrative perspective, systems differ in two important ways—data costs and risk of encouraging gaming or perverse incentives.

Data costs

All classification systems require data collection. However, DRG's, PMC's, and staging run on computers using UHDDS data which is already collected by Medicare. This is a major practical advantage. APACHE II and MEDISGRPS run on a defined clinical data set which could be collected as part of discharge abstracting by medical records technicians with minimal training. Severity, which is the most impressionistic or implicit method, requires chart reading by individuals with training well beyond that of most medical records technicians. All of the methods except severity can be performed by computer once the data is abstracted.

Incentives and gaming

A classification system, when implemented with accompanying weights, provides a set of incentives for hospitals. Some of these incentives are market incentives—it may be profitable to seek certain classes of patients while avoiding others. Other incentives are invitations to "gaming"—those maneuvers by which a hospital can enhance its revenue without improving its services and which run contrary to the intent of the system. Examples include some rearranging of the order of diagnoses in systems such as DRG's which are responsive to this order, dividing a single admission into two or more admissions, and reporting marginal complications and comorbidities to change the DRG classification (Gertman, 1984). If a relatively implicit system such as severity was used at hospitals by hospital personnel, the risk of gaming would probably be high enough to require an active audit system. Gaming shades into fraud, with such overt activities as reordering diagnoses in violation of accepted standards, deliberately miscoding diagnoses, or reporting procedures which were not performed. The importance of gaming in the real world of hospital management is hotly debated, but there is consensus that a classification system should place minimum emphasis on ambiguous information which might be gamed and that the incentives of the system must be carefully examined.

Empirical issues

Empirically, there are two critical tests of a system—reliability and ability to account for variation.

Reliability

A case-mix system should reliably classify the same case in the same way when that case is abstracted by different abstractors. Although all systems described in this issue except Severity of Illness Index are computerized and therefore appear reliable, significant variations actually exist.

Assignment of diagnosis by a physician and coding of diagnosis by a records technician involve judgment, particularly in determining the principal diagnosis. When the Institute of Medicine (1977) studied coding under the *International Classification of Diseases, adapted for use in the United States, Eighth Revision*, there were errors in 30 percent of cases, and in 10 percent of cases experts could not agree on the correct coding (Demlo and Campbell, 1981; Grimaldi *et al.*, 1983). The current diagnostic system (*International Classification of Diseases, Ninth Revision, Clinical Modification*) is more specific; coding conventions are more widely shared; and the incentives of PPS will almost certainly improve data quality. Indeed, the incentives of PPS might lead to consistent resolutions of ambiguous situations in a way which, by emphasizing reimbursement, will conceal the ambiguity. The Inspector General of the Department of Health and Human Services has, however, compiled a substantial list of situations in which judgment calls would result in assignment to different DRG's and consequently in different reimbursement. And there are common situations in which coding according to established conventions runs counter to clinical intuition and provides less information than "clinically appropriate coding." DRG, staging, and Severity of Illness Index

all depend on decision as to the principal diagnosis. Despite the importance of this issue, there is no data comparing the sensitivities of these different systems to empirically-observed coding error rates.

The Severity of Illness Index, an apparently more subjective method, has high reported interrater agreement for most raters, but the reliability figures are reported in a way which makes precise interpretation difficult. The way in which the subscales are scored is subjective, despite the guidelines provided, and there is no explicit way to combine the individual scales into the Severity of Illness Index. It is important to realize that, because Horn suggests using the Index as a refinement within DRG, errors in the Index add to the errors in diagnosis coding for DRG rather than replacing them; thus it is not appropriate simply to compare error rates for the two, even if data for such comparisons were available.

The Severity of Illness Index presents another potential problem which lies between reliability and validity: It is not clear whether raters may respond in scoring to the diagnostic efforts made and the care rendered rather than to the actual condition of the patient. Considerable clinical skill is required to discriminate between diagnostic efforts made and the clinical evidence which may have occasioned them, particularly when working from a medical record rather than from the clinical setting. Furthermore, extended and intensive care usually produces a more documented chart, and it is difficult to determine the degree to which such documentation influences the Severity of Illness Index, independent of the necessity for the care.

When examining reliability figures, it is extremely important to know whether errors are random (and can therefore be expected to cancel out for large numbers of cases) or whether they are or can be made systematic (so that they would bias overall payments even for a large hospital). If the latter is the case, rather modest error rates can jeopardize the integrity of an entire system and must be taken very seriously. For this reason, reliability should be reported in terms of its impact on overall payments to a hospital under a reimbursement model. Such data is not available for any classification system.

Explained variance

Investigators often measure the power of a classification system by the amount of variance in either resources used or length of stay which the system explains. Statistically, this is equivalent to the requirement that a good classification system should have substantial differences between groups and considerable coherence within groups. However, a classification system which explained 100 percent of the variance would not be good—it would implicitly assert that all care was necessary and would effectively return a reimbursement system to cost-based reimbursement. Unless it is desirable to pay for variations in physician and hospital efficiency, for example, it is not desirable to account for them in a classification system.

The data which is not considered by a system sets an upper limit on how much variance a system can reasonably account for. For example, variation in clinical practice between physicians is probably quite important and Horn has pointed out a useful direction by studying the contribution which considering inter-physician variation can make to assessing a case-mix system. But one might also wish to know the importance of such factors as whether the diagnosis was known before admission, the purpose of treatment, and disposition problems; one would need to be skeptical of a system which explained very large portions of the variance without considering such factors.

We have been unable to locate any published report of the power of any of the classification systems described in this report when applied to any national data base. The closest approximation is the data of Pettengill and Vertrees (1982) on the power of DRG's to account for variation in average hospital case cost using Medicare data and Horn's reports of Severity of Illness Index's explanatory power in a group of about 30 hospitals.

The problem of explained variance relates to the question of how many categories a system should have. Systems with a larger number of categories tend to explain more of the variance in resource use. A number of critics have suggested that the number of categories should be limited because the number of cases in individual patient categories in an individual hospital will otherwise be too small to support systems of payment based on averaging. Number of cases in a category is not a problem in itself, since the total number of cases in the hospital rather than the number in each category determines the degree to which the law of large numbers protects the hospital. It is true, however, that as the number of categories proliferates, larger data sets are required for calculating weights and that the overall system becomes somewhat less useful to the hospital as an internal management tool.

Finally, explanatory power may not be the most important determinant of the fairness of a reimbursement system. As noted in the last section, the critical question is whether the system is fair to groups of patients and hospitals. Although intuition suggests that the two questions are equivalent, they may not be in practice, and this question also invites empirical studies which have not been done.

A research strategy

The research problem described earlier is how to identify and correct variation in requirements which occurs between patient groups, hospitals, or types of hospitals. To this end we now describe a possible staged research agenda (Gertman, 1984).

For the short range (1985-86), this agenda might start with defining and making plans to acquire a data base for adequate comparisons of competing classification strategies. This agenda might also include activities which could yield immediate results such as new ways to compute weights for DRG's, ways to sharpen the DRG classification system and to merge DRG's with other UHDDS-based systems, and ways to use information about previous patient treatment.

For the intermediate range (1987-88), the agenda might include studying whether there are real unmeasured variations in patient treatment requirements between hospitals and studying ways to update both the DRG classification system and pricing to reflect changes in practice and technology.

For the longer range (perhaps 1990), the agenda might include developing and testing case-mix systems which represent a full generation of advance over the present DRG's and which consider variables such as the degree of previous diagnostic knowledge about the patient, the purpose of the admission, psychosocial characteristics of the patient, and whatever issues concerning severity of illness have been proven important in the intermediate range research.

Short-range research

It is clear that adequate comparisons for competing classification systems will require large and expensive data bases. Planning the form of such data bases and starting to acquire them has high priority in research to improve classification systems; although many results from such a data base may not be available in the short term, comparisons of staging, PMC, and DRG can be accomplished quite rapidly and it is possible that studies of other methods can also be completed quickly. These results will be of great importance in planning longer-range research.

In the short range we also need to study whether we can improve DRG system performance by using cleaner data or better procedures to weight the DRG's, by improving the DRG classification system, by incorporating successful aspects of other UHDDS-based systems such as staging and PMC's, or by using treatment history as data.

Weighting procedures

PPS might be significantly improved by using better data and procedures for developing weights, and improved performance might be apparent if the improved DRG's were tested on better data.

The discharge data used for creating weights should be as accurate as possible. The incentives of PPS should result in a steady improvement in data reported to Medicare, as should the application of automated editing procedures. In addition, physician bills could be used to validate hospital procedure codes and possibly other data. It would also be desirable to eliminate data pertaining to care which has been disallowed on review, which might imply a preference for data from regions where review has been especially vigorous or effective.

We can also test modifications to the weighting system which would counteract known or suspected biases. For example, we know that the use of average per diems understates the variation in nursing costs. Further study of nursing intensity (Thompson, 1984) would also shed light on this issue.

DRG classification system

We do not discuss here the efforts which HCFA is already making under Congressional mandate to include psychiatric and rehabilitation units of general hospitals, psychiatric and rehabilitation hospitals, childrens' hospitals, and long-term care facilities within PPS. There are three other directions for work on the DRG algorithm:

Refining the present algorithm: The existing DRG algorithm has a number of important limitations which could be explored and perhaps improved by very straightforward research in the following areas:

- The lists of comorbidities and complications are not specific to the MDC's in which they are applied and have not been validated. These lists should be refined and the methods of dealing with comorbidities and complications suggested by Young should also be explored.
- Certain cases which are highly similar in treatment strategy and resource use are assigned inappropriately to different DRG's because of an assumption that a patient should not reach the same DRG from two MDC's. Thus, for example, a patient admitted for end-stage diabetic renal disease goes into a different DRG from a patient with end-stage hypertensive renal disease.
- The possibility of classifying patients according to operative procedure rather than diagnosis needs to be explored in cases where a major operative procedure relates to a secondary diagnosis rather than a primary diagnosis.
- Methods are needed to classify multiple identical procedures such as bilateral hip replacements or bilateral cataract procedures.

Synthesizing DRG's with other UHDDS-based systems: Staging and patient management categories are computerized systems which operate on the current Medicare data set. Although merging the methods presents substantial technical problems, SysteMetrics has done work on a synthesis of staging and DRG's for a few DRG's which suggests that this strategy may be effective (SysteMetrics, 1984). A

broader effort involving both staging and patient management categories is conceptually straightforward and could proceed as soon as 1983 data is available. Other systems, which cannot operate from UHDDS data, cannot be merged with the DRG system at this time.

Using previous treatment as a patient variable: An important piece of information about patients—previous inpatient and perhaps outpatient treatment—is available from Medicare files but has not been used for case-mix analysis. Previous treatment may be a better indicator of comorbidity and the severity of disease than those data elements used by DRG's. The Medicare file of previous admissions may also be a useful tool for approaching the problem of "split admissions": cases in which medical judgment would permit managing a problem on either a single admission or two admissions.

Intermediate range research

In three to four years we should be able to complete studies to determine whether there are significant variations in treatment requirements among hospitals which are unmeasured by DRG's. There are two general approaches to this problem:

Overall characterization of hospitals using multiple methods

This is the research which should be done on the data bases planned under short-term research, perhaps including special on-site data collection to measure necessity for treatments rendered. This approach would probably characterize the case mix of hospitals using APACHE II, MEDISGRPS, Severity of Illness Index, and on-site data. These methods would be compared to the refined DRG's developed in the short range program. The different methods may have special power if they are used to cross-validate one another. The goal of this combined analysis would be to compare for different hospitals the degree to which differences in cost reflect differences in services rendered for comparable cases as opposed to differences in case mix. In particular, this strategy would allow us to determine whether public, teaching, and specialty hospitals truly have different case mixes from those measured by the DRG system.

Tracer conditions

A second approach to estimating the degree to which care rendered exceeds necessary care in different hospitals is to fully analyze certain tracer conditions to determine whether hospitals have different costs for conditions which are identical when fully corrected for intensity of illness and other variables. List's (1983) study of management of myocardial infarction in Maryland and Oregon provides an example of how such a study might be carried out. By examining common conditions for which very similar cases can be selected in different hospitals, such studies can shed great light on the relative efficiencies of different kinds of hospitals. The same methods can clarify the relative treatment requirements of different subgroups of patients in the same DRG.

Either of these strategies would permit us to accurately distinguish between hospitals which are winners and losers under PPS because of unmeasured case-mix variation and hospitals which are winners and losers because of efficiency or inefficiency. The former strategy would be somewhat more persuasive because it would consider all cases in a hospital, but it would also be more complex and difficult.

Other studies

Studies are needed of ways to update classification schemes to incorporate advances in practice and technology. This task would include studies of the differential impact of systems of classification and weighting on adoption of new practices and technology. Studies might also explore ways to appropriately recognize new practices and technology in the classification scheme as they are introduced. This problem is not simply a matter of maintaining the classification and weighting system—it also includes maintaining the diagnostic and procedure nomenclature. Over the next five years, a new diagnostic system, *International Classification of Diseases, Tenth Revision*, will be designed. Traditionally, diagnostic coding has been little influenced by considerations of reimbursement. A major research task will be determining the degree to which diagnostic nomenclature should be modified and the degree to which those modifications are possible within the international agreements which govern the system used in the United States. A similar major task will involve establishing a procedure classification system that is better suited to reimbursement.

Long-term research

Long-term strategies rest on more sophisticated concepts of case mix. These strategies assume that by the end of the intermediate phase we will have a classification system which better reflects patient condition. However, factors other than patient condition influence treatment needs and should be measured in a more sophisticated system.

The purpose of treatment

Physicians may have different goals in treating patients with similar conditions. Patients admitted for terminal care will receive very different care from a patient whose physician is determined to save him. Garber *et al.* (1984) found that a significant part of the greater costs of treatment at teaching hospitals may be related to the fact that in a nonteaching service more patients with a high risk of dying were admitted for essentially supportive care. It is interesting that, although the teaching patients appeared to have a lower in-hospital mortality, their mortality at 9

months followup was the same as the nonteaching patients in the same hospital, suggesting that the teaching service may simply spend more effort prolonging the inevitable. Treatment goals may vary widely for other reasons. For example, two patients in identical condition may be admitted to a rehabilitation hospital with quite different rehabilitation goals for the individual admission. A psychiatric hospital may set very different goals for two very similar schizophrenic patients depending on data not easily found in the record. Or a surgeon may select paliative treatment for one cancer patient and aggressive curative surgery for another, based on patient preference.

Knowledge about the patient at admission

When the diagnosis is known before admission, one can expect less need for diagnostic activity and diagnostic cost and, since the admission is more likely to be for treatment, we can expect more treatment activity and treatment cost. Several competing systems, notably PMC's, seek to address this problem, but the degree of success is uncertain.

Patterns of clinical practice

Equally-trained physicians treat similar patients in different ways. Some operate on asymptomatic gallstones, others do not; some hospitalize for unexplained chest pain, others do not. Wennberg (1984) has documented the wide variations in admitting and surgical patterns. Thus, there will be considerable variation in the diagnostic and treatment activities defined as necessary by different experts.

Referral practices

In communities there are often patterns of referral which result in differential assignment of the most difficult or treatment-resistant cases to certain hospitals. For example, in communities where general hospitals will not accept committed psychiatric patients, the most difficult patients are routinely triaged to hospitals which will accept commitments. Similar effects are said to occur with patients from nursing homes in some cities. Unless these referral criteria are included in the data set for testing the classification system, they may be very difficult to pick up indirectly. With other kinds of referrals, decisionmaking may be based on factors which are intuitive and difficult to quantify. Field studies could clarify the importance of these practices.

Psychosocial characteristics of patients

Large urban hospitals have argued that psychosocial factors (such as whether patients have a fixed address, have sufficient money to alter living arrangements, have someone at home to care for them) strongly influence length of stay and care costs for medical-surgical patients. These variables are especially important because, as suggested earlier, hospitals can easily create admission barriers for patients they do not want if they believe that the variables are important but not recognized in the classification system.

This research requires developing measures for new variables as well as collecting data to measure their impact and their importance for prospective payment; both the methodology and the analysis will be extremely challenging.

Conclusion

HCFA's interest in case-mix analysis is necessarily intense, and its stake in the accuracy of the system which it uses is high. With PPS in place, a systematic program of research to test the adequacy of the DRG system and simultaneously to improve the system is essential to making PPS fully effective. We believe that the papers in this issue of the *Review* provide the basis for long-range thinking and hope they will help to stimulate the needed research.

Acknowledgments

The authors wish to thank Rosanna Coffey, Ph.D., Judith Lave, Ph.D., and Julian Pettengill, Ph.D. for many helpful suggestions regarding this paper.

References

Brewster, A., Jacobs, C. M., and Bradbury, R. C.: Classifying severity of illness using clinical findings. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Conklin, J. E., Lieberman, J. V., Barnes, C. A., and Louis, D. Z.: Disease staging: Implications for hospital reimbursement and management. *Health Care Financing Review*. 1984 Annual Supplement HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Demlo, K., and Campbell, P. M.: Improving hospital discharge data: Lessons from the National Hospital Discharge Survey. *Med Care* 1981; *19*:1030-1040.

Gertman, P. M., and Lowenstein, S.: A research paradigm for severity of illness: Issues for the diagnosis-related group system. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Grimaldi, P. L., Michiletti, J. A., Zipkas, M. M.: The use of DRG's—selected implications for medical record departments. *J Am Med Rec Assoc* 1983; *54*:24-30.

Garber, A. M., Fuchs, V. R., and Silverman, J. F.: Case mix, costs, and outcomes: Differences between faculty and community services in a university hospital. *N Engl J Med* 1984; *310*:1231-1237.

Horn, S. D., Horn, R. A., and Sharkey, P. D.: The Severity of Illness Index as a severity adjustment to diagnosis-related groups. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing

Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Hornbrook, M. C.: Hospital case mix: Its definition, measurement, and use: Part I. The conceptual framework, *Med Care Rev*, 1982; *39*:1-43

Hornbrook, M. C.: Hospital case mix: Its definition, measurement, and use: Part II. review of alternative measures, *Med Care Rev*, 1982; *39*:75-123.

Institute of Medicine: Reliability of medicare discharge records. National Academy of Sciences. NTIS No. PB281680, Nov 1977.

List, N. D., Fronczak, N. E., Gottlieb, S. H., *et al.*: A cross-national study of differences in length of stay of patients with cardiac diagnoses. *Med Care* 1983; *21*:519-530.

Pettengill, J., and Vertrees, J.: Reliability and validity in hospital case-mix measurement. *Health Care Financing Review*. Vol. 4, No. 2. HCFA Pub. No. 03149. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Dec. 1982.

Smits, H. L., Fetter, R. B., McMahon, L. F.: Variation in resource use within diagnosis-related groups: The severity issue. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

SysteMetrics Inc., "Refinements to diagnosis related groups based on severity of illness and age": Draft Final Report, Contract HHS-100-82-0038, Office of the Assistant Secretary for Planning and Evaluation and Health Care Financing Administration, Department of Health and Human Services, July 9, 1984.

Thompson, J. D.: The measurement of nursing intensity. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Wagner, D.: Acute physiology and chronic health evaluation (APACHE II) and Medicare reimbursement. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.

Wennberg, J. E., McPherson, K., and Caper, P.: Will payment based on diagnosis-related groups control hospital costs?, *N Engl J Med*, 1984; *311*:295-300.

Young, W.: Incorporating severity of illness and co-morbidity in case-mix measurement. *Health Care Financing Review*. 1984 Annual Supplement. HCFA Pub. No. 03194. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Nov. 1984.