Alternative Systems for Case Mix Classification in Health Care Financing

Shan Cretin, Linda G. Worthman
The research described in this report was supported by the Health Care Financing Administration, U.S. Department of Health and Human Services, Grant No. 99-C-98489/9-03.

Library of Congress Cataloging in Publication Data

Cretin, Shan.

Alternative systems for case mix classification in health care financing.

"R-3457-HCFA."

"Prepared for the Health Care Financing Administration, U.S. Department of Health and Human Services."

"October 1986."

Bibliography: p.


V. Title: Case mix classification in health care financing.

RA971.32.C74 1986 362.1'1'0681 86-26053

ISBN 0-8330-0763-7

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of The RAND Corporation do not necessarily reflect the opinions or policies of the sponsors of RAND research.
Alternative Systems for Case Mix Classification in Health Care Financing

Shan Cretin, Linda G. Worthman

September 1986

Prepared for the
Health Care Financing Administration,
U.S. Department of Health and Human Services

RAND
PREFACE

This report was prepared as background for the Health Care Financing Administration's report to Congress, *Refining Case Mix Adjustment in Medicare's Prospective Payment System: Severity Adjustments, Outlier Payments, and Other Options*. It compares the structural and performance characteristics of six case mix adjustment methods. The methods reviewed are Diagnosis-Related Groups (DRGs), APACHE II, Disease Staging, MEDISGRPS, Patient Management Categories (PMCs), and Severity of Illness Index (SOII).

This work was conducted at the RAND/UCLA Center for Policy Research in Health Care Financing, which is supported through Cooperative Research Agreement 18-C-98489/9-01. See also Linda G. Worthman and Shan Cretin, *Review of the Literature on Diagnosis-Related Groups*, N-2492-HCFA, October 1986.
SUMMARY

When Congress designed and implemented the Prospective Payment System (PPS), the best available method to adjust for differences in case mix was the Diagnosis-Related Groups (DRGs). The DRG system classified all patient cases, and had already been implemented in New Jersey's prospective payment system. Some of the problems in an earlier version of the DRGs have been resolved, but there remains concern that DRGs may need to be refined or replaced to better reflect variation in patient condition. As part of a continuing review of the Prospective Payment System, Congress requested the Health Care Financing Administration (HCFA) to report on Refining Case Mix Adjustment in Medicare's Prospective Payment System. The review in the present report, prepared as background to HCFA's report, compares the structural and performance characteristics of case mix adjustment measures being considered as replacements for or refinements to the Diagnosis-Related Groups.

We reviewed published papers, prepublication drafts, and technical reports on five case mix systems: APACHE II, Disease Staging, MEDISGRPS, Patient Management Categories (PMCs), and Severity of Illness Index (SOII). These systems are in different stages of development and testing, and comparable information is not available for each system. Comparing the systems is further complicated by the absence of a "gold standard" defining the factors for which case mix should and should not adjust. In the PPS, the method needs to adjust for the costs of medically necessary treatment, but at the present stage, neither "medically necessary" nor "costs" have been measured. In practice, the performance of these systems (including DRGs) is typically tested by predicting other measures of resource use such as length of stay or hospital charges.

The alternative systems differ in classification structure, data requirements, and stage of development. These characteristics (summarized in Table S.1), in turn, determine the feasibility of using each of these systems to replace or refine DRGs.

At present, no system appears to perform better than the DRGs. The two methods which on structural grounds could replace DRGs—Disease Staging and PMCs—do not predict resource use better than the DRGs in tests on the same data. Of the methods which could be used to refine DRGs—APACHE, Disease Staging, MEDISGRPS, and SOII—only the Staging modification of 17 DRGs (ADRGs) appears ready for more widespread testing. Neither of the clinical data systems
(APACHE and MEDISGRPS) has been tested, even on a small scale, as a refinement to DRGs. The Severity of Illness Index is not suitable in a reimbursement system because of its dependence on costly data collection and implicit rater judgments, and it is too early to evaluate its successor system, the Computerized Severity Index (CSI).

Table S.1

<table>
<thead>
<tr>
<th>System</th>
<th>Possible Uses in PPS</th>
<th>Classification Structure Based On:</th>
<th>Data Requirements</th>
<th>Subjective Data</th>
<th>Developed Using Probable Correlates of Resource Need</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRGs</td>
<td>—</td>
<td>Diagnoses and procedures</td>
<td>UHDDS data</td>
<td>Discharge diagnosis</td>
<td>Yes</td>
</tr>
<tr>
<td>APACHE II</td>
<td>Refine DRGs</td>
<td>Severity scale</td>
<td>Clinical data</td>
<td>None</td>
<td>No</td>
</tr>
<tr>
<td>Disease Staging</td>
<td>Replace or refine DRGs</td>
<td>Severity scale within disease conditions</td>
<td>UHDDS data</td>
<td>Discharge diagnosis</td>
<td>No</td>
</tr>
<tr>
<td>MEDISGRPS</td>
<td>Refine DRGs</td>
<td>Severity scale</td>
<td>Clinical data</td>
<td>Physical findings, X-Ray interpretations</td>
<td>No</td>
</tr>
<tr>
<td>PMCs</td>
<td>Replace DRGs</td>
<td>Diagnoses and procedures</td>
<td>UHDDS data</td>
<td>Discharge diagnosis</td>
<td>Yes</td>
</tr>
<tr>
<td>SOII</td>
<td>Refine DRGs</td>
<td>Severity scale</td>
<td>Medical record</td>
<td>Implicit coder judgments</td>
<td>No</td>
</tr>
</tbody>
</table>
CONTENTS

PREFACE .................................................. iii

SUMMARY .................................................. v

TABLES .................................................. ix

Section

I. A FRAMEWORK FOR COMPARISONS ...................... 1
   Introduction ......................................... 1
   The Gold Standard Problem: For What Should
   Case Mix Adjust? .................................. 1
   A Framework for Viewing Case Mix Systems .......... 3
   Criteria for Different System Uses .................. 11

II. COMPARISON OF THE STRUCTURAL
    CHARACTERISTICS OF ALTERNATIVE
    SYSTEMS ............................................. 13
    Diagnosis-Related Groups (DRGs) ..................... 13
    APACHE ........................................... 18
    Disease Staging .................................... 20
    MEDISGRPS ...................................... 23
    Patient Management Categories (PMCs) ............... 26
    Severity of Illness Index (SOII) .................... 28
    Summary of Comparative Structural Features
       of Competing Systems .......................... 33

III. COMPARING THE PERFORMANCE OF CASE MIX
    METHODS ............................................. 35
    Reliability ....................................... 35
    Predictive Performance ........................... 43
    Patient Management Categories (PMCs) ................ 52
    Comparative Performance .......................... 58

IV. CONCLUSIONS .......................................... 61
    Performance of the Alternatives as DRG
    Substitutes ..................................... 61
    Potential of the Alternative Systems for
    Refining DRGs .................................. 62
    Work in Progress and Future Research Needs ....... 64
    Summary .......................................... 66

REFERENCES ............................................. 67
TABLES

S.1. Structural Characteristics of Alternative Systems ............... vi
1. Reproducibility ........................................... 37
2. Performance of ICD-9-CM DRGs .................................. 44
3. Medical and Surgical DRGs .................................... 45
4. Performance of DRG Case Mix Index .............................. 46
5. Performance of APACHE II ...................................... 48
6. Performance of Disease Staging .................................. 50
7. Comparison of Variance Explained (R²) for DRGs and 10 SysteMetrics ADRGs ............................................. 51
8. Performance of MEDISGRPS in Predicting Mortality ......... 52
10. Performance of MEDISGRPS in Predicting Charges
    (Case Level, First Review) .................................... 54
11. Performance of MEDISGRPS in Predicting Charges
    (Case Level, Second Review) .................................... 54
12. Performance of Patient Management Categories .................. 55
13. Comparison of the Variation (R²) in Cost per Case at Case Level Explained by Three Case Mix Measures ............. 56
14. Performance of Severity of Illness Index ....................... 57
15. Comparison of Severity/Procedure-Adjusted DRGs
    With DRGs .................................................... 58
16. Medical and Surgical PMCs, Disease Staging, and DRGs .......... 60
I. A FRAMEWORK FOR COMPARISONS

INTRODUCTION

The purpose of case mix adjustment is to classify patients together whose inpatient care requires similar resources. The Prospective Payment System (PPS) or any other case-based payment system then determines the correct payment to be made for that care in the hospital in which it was rendered. Case mix adjustment began as part of hospital utilization review and quality assurance programs, and many case mix adjustment systems were designed to group together patients with similar expected length of stay or similar expected mortality. For PPS, however, the criterion variable is ability to class together patients with similar resource needs.

This report compares Diagnosis-Related Groups (DRGs) with five other case mix adjustment systems—APACHE, Disease Staging, MEDISGRPS, Patient Management Categories (PMCs), and Severity of Illness Index (SOII). There are three uses to which these alternative case mix systems could be applied within the context of PPS: assessing the adequacy of the DRG system, replacing the DRG system, and refining the DRG system. In comparing the five alternative systems, we will consider the suitability of each system for use in evaluating, replacing, or modifying DRGs.

The remainder of Sec. I lays out considerations for comparing alternative systems. Section II describes the structure of each of the alternative systems, Sec. III compares their performance, and Sec. IV assesses the feasibility of using these systems as refinements to DRGs and suggests a future strategy for research and development.

THE GOLD STANDARD PROBLEM: FOR WHAT SHOULD CASE MIX ADJUST?

The purpose of a case mix adjustment procedure is to group together patients whose care requires similar resources. There are problems, however, in defining what resources are required by any particular case. In the absence of a “gold standard,” the problem of assessing which of two case mix systems is “better” is difficult indeed. We can identify cases on which two different case mix systems give disparate estimates of needed resources. However, without an independent measure of truth, we cannot determine which system made the best call.
One approach to defining necessary services is to allow a panel of physicians to review the record of a hospitalization to identify the services and procedures that should be required in managing the case. These services and procedures could then, in theory, be translated into necessary costs. Implementation of this method, however, requires the resolution of both clinical and economic issues.

The first issue is the degree to which in-hospital events ought to be used in determining necessary services and procedures. Complications and idiosyncratic responses to therapy will occur in some percentage of cases, even under the best conditions. There are two ways in-hospital events might be included. First, the expected costs over all patients including those with and without in-hospital events can be used in determining the "average necessary resources." Second, separate categories can be established for patients with and without in-hospital events. In this case, the actual occurrence of an in-hospital event will raise the predicted "necessary resources." The first method assumes that in each hospital the burden imposed by in-hospital events will be approximately equal to the overall average burden imposed by such patients. The view is that excessive complication rates reflect poor quality of care and should not be covered by reimbursement.

The second method (including in-hospital events in the classification scheme) is consistent with the view that apparently excessive complication rates at some hospitals could come about by chance alone (particularly for hospitals with a small number of Medicare admissions) or because patients with a history of nonresponse and those most vulnerable to complications are systematically referred to certain hospitals which naturally experience high rates of complications during the hospital stay. Explicitly including in-hospital events in the case mix classification scheme protects small hospitals and referral hospitals, but runs the risk of giving higher reimbursement to those institutions at which poor medical care is the cause of higher complication rates. The inclusion of in-hospital events becomes even more vulnerable to criticism if the mere provision of services is taken as evidence that the services were dictated by a change in patient condition. In practice, the distinction between the occurrence of an in-hospital event and the provision of a service without independent evidence of its need can be hard to draw, especially if one is limited to abstracts of medical records, rather than the records themselves. Existing case mix systems differ on whether and how in-hospital events are to be included in the criteria for classification. The issue remains a controversial one.

The second problem is how to determine the cost of providing a given set of services. Many factors influence the costs of providing a service: the cost of capital and wages faced by the hospital, the
number of other Medicare and non-Medicare patients using the same service, and even characteristics of the patient (whether the patient speaks English, is conscious, requires special assistance to move from one place to another). The practical problem of incorporating these and other exogenous factors into a useful cost-determining algorithm has yet to be solved.

In the absence of a "gold standard," systems have been developed using either actual costs or, as a proxy for costs, actual charges or actual length of stay. All three of these approaches resolve the first issue raised above by implicitly including the hospital's response to in-hospital events. Using length of stay or charges ignores the problem of cost-finding. Methods using actual costs have relied on standard Medicare cost-accounting procedures to attribute "average costs."

The use of actual accounting costs as a proxy for necessary costs has very nearly emerged as a "silver standard" in the research community. How defensible is this position? There are wide variations in actual costs between regions of the country, between urban and rural areas, between large and small hospitals, and between teaching and nonteaching hospitals. Although much of this difference is attributable to differences in labor costs, diagnostic mix, and severity of illness, some is undoubtedly attributable to variations in the efficiency of different hospitals in producing services. There is substantial controversy as to whether a significant residue of the difference may be due to different practice styles for treating similar patients. Further, there are substantial differences between classes of hospitals in the frequency with which surgery is performed for patients with similar disorders, and it is unclear how much of this difference is attributable to referral of surgical patients to specialty centers and how much is attributable to differences in judgment as to how often surgery should be performed. Finally, there are differences in how aggressively various hospitals and physicians treat patients with similar conditions. For all these reasons, the use of actual costs in the development and evaluation of case mix methods may be misleading. We may reasonably expect that a case mix system which performs superbly in predicting necessary costs will fail to explain a high fraction of actual costs.

A FRAMEWORK FOR VIEWING CASE MIX SYSTEMS

Because case mix adjustment methods have been developed with diverse and distinct goals, comparing different systems first requires the establishment of a common framework. We propose to look at the case mix methods along two dimensions: structure and performance.
In comparing the structure of the case mix systems, we will look at six important characteristics:

- Classification structure
- Variant versions
- Data requirements
- Face validity
- Subjectivity or vulnerability to manipulation
- Availability of documentation

In comparing the performance of the case mix systems, we will look at two characteristics:

- Reliability
- Predictive validity

Having established this framework, we will then consider how the intended use of a case mix adjustment system affects the relative importance of these dimensions.

**Structural Characteristics of a System**

**Classification Structure.** Two approaches have been used in creating case mix systems. In the first approach, a set of distinct patient categories is created based on either clinical or statistical criteria or on a combination of the two; in a later step, a number is attached to each category representing the estimated resource needs for that category relative to the overall average. The DRG system, Disease Staging, and PMCs represent case mix systems of this type.

In the second approach, statistical (or other) methods are used to derive a scale or index number for each case. In this case, two patients with the same scale or index score may present entirely different clinical conditions. Usually the scale or index is envisioned as being used in conjunction with a diagnostic classification scheme. In that case, the meaning of a particular value of the scale or index in terms of predicting resource needs need not be unique, but may depend on other facts about the patient (such as diagnosis or procedure). APACHE, MEDISGRPS, and SOII represent systems of the second type.

Systems created using the first approach result in diagnosis-specific case mix categories. Diagnosis-specific case mix classification methods differ in how they handle patients with multiple diagnoses or procedures. Some systems uniquely classify every admission into exactly one category based on a series of decision rules (DRGs, for example). Others may place an admission in more than one category on the basis
of each diagnosis (Disease Staging or PMCs). A decision rule may subsequently be applied to assign the patient to one category (for example, the highest category across all diagnoses or the category associated with the principal diagnosis).

Systems differ in the number of categories they produce. Some systems producing a single severity statistic independent of diagnosis appear to produce very few categories; since the severity statistic must generally be combined with diagnostic categories to make the system usable for reimbursement, the number of categories is much larger in practice. The 467 categories of the DRG system are relatively parsimonious compared with some other systems. From the perspective of payment, the only limit on the number of categories are the data available to determine payment rates for the categories.

**Variant Versions.** All of the systems described in this report are still being developed, usually by the originator but sometimes by others as well. In consequence, a test of one version of the system may not apply, and a limitation in one version may not be present in another. In general, different versions constitute a developmental sequence, but in some cases there are versions which are being developed along different lines and are incompatible with one another. Variant versions complicate evaluation of the case mix systems.

**Data Requirements.** All hospitals must provide the Uniform Hospital Discharge Data Set (UHDDS) for each Medicare discharge and this information is available in computerized form. UHDDS includes the principal diagnosis and four secondary diagnoses, three procedures, age, sex, discharge status, race, ethnicity, and little else. Some systems can classify patients using UHDDS data alone; others require data readily abstracted from the chart; and others require data which can only be identified by trained abstractors. Since UHDDS data are already collected by all hospitals and converted into computer-readable form, UHDDS-based systems have the advantage of placing no additional data collection burden on the hospitals. This issue is discussed for each system in Sec. II.

**Face Validity.** The absence of a “gold standard” measure of needed resources makes it impossible to assess the criterion validity of any case mix method. We have, however, two other measures of validity. The first is the face or construct validity of the system. Is the classification system constructed in such a way that it ought to group together cases with similar need for medical resources? Is the categorization based on data elements which we would a priori assume to be related to resource need? Does the way in which the data elements are combined make clinical sense?
The second type of validity check available to us is predictive validity. Predictive validity in case mix systems is usually measured by determining how well the system groups together cases with similar actual costs. Predictive validity of the systems will be considered as a performance attribute in Sec. II. We are unaware of studies addressing what we consider true validity—whether the system groups together cases whose necessary costs are similar.

Subjectivity or Vulnerability to Manipulation. Some data in the medical record are relatively objective. Recording the value of the patient's white blood count at admission, for example, requires no interpretation by the person abstracting that datum. Coding a principal or secondary diagnosis, on the other hand, may call for judgment by the coder. Any abstraction which requires the coder to incorporate information from various places in the medical record and render an implicit judgment about the case is vulnerable to the introduction of bias in the coding process. This is not to say that a subjective measure cannot be reliable. Raters who share a frame of reference may produce highly reliable decisions which are very different from raters with different points of view. In payment policy the important question is how much difference in classification will occur between a rater who seeks to maximize Medicare payments (e.g., one working for the hospital) and one who seeks to minimize Medicare payments (e.g., one working for Medicare).

We will consider reliability of a system as a performance characteristic. Although we separate subjectivity and reliability, the two interact and are sometimes confused in the literature. We discuss the problems associated with the measurement of reliability in the section on performance characteristics.

Availability of Documentation. Systems differ in the degree to which they are in the public domain. A secret, copyright system which is being marketed by its developer for profit may be suspect if it is promoted for inclusion in a mandated public program such as Medicare. But a more complicated issue is the degree to which the systems are open for public inspection and research. A system can be unavailable for numerous reasons. An extreme is that the system may be copyright and the owners may refuse to allow dissemination of the key materials. A system may also be difficult to examine if it relies on lengthy, complex computer programs which have not been fully documented or if implementation of the system requires the use of specially trained coders for which no training manual is publicly available.
Performance Characteristics of a System

Reliability. The reliability of a case mix classification system depends on (1) the accuracy with which the necessary data can be abstracted from the medical record and (2) the reproducibility with which the classification can be generated from the requisite data.

For objective data, accurate coding depends on the ability of coders to consistently locate and record the necessary information. This task may be complicated by the presence of contradictory or ambiguous information in some records. When coding calls for subjective judgments based on the abstractor's implicit integration of information in the medical record, reliable coding will depend on the ability of the trainer to provide clear criteria for making the needed judgments. A trainer may, indeed, be able to train coders to carry out highly reliable abstractions. At the same time, another trainer with different motivations may be able to provide equally clear, equally defensible criteria which would consistently result in different coding of some cases. Although biased coders may produce biased coding of objective data, it is usually possible to produce conclusive evidence from the record as to what the correct code should have been. In the case of subjective measures, establishing "truth" may be more difficult. These considerations lead us to conclude that for all systems, but especially for systems relying on subjective measures, reliability must be judged not only over well-intentioned coders facing the same incentives, but over coders facing different incentives as well.

The reproducibility with which the input data is used to create the case mix classification is also a component of the reliability of the system. There is a general illusion that computerized systems are 100 percent reliable in this regard, whereas the accuracy of manual methods must be determined. In fact, computer systems, given the same set of input data, can safely be regarded as 100 percent reliable in generating the same classification. However, is the classification generated by the program the one intended? Since many of the computerized algorithms are quite complex, it is likely that unintended "bugs" may result in some misclassification by the programs. Such "bugs" may even be "machine dependent;" that is, the algorithm may work one way in some implementations of the program and in a slightly (or even grossly) different way in an implementation on a different computer system. Programs can and are tested for accuracy in classification, but it is quite possible (perhaps even likely) that in complex classification algorithms, errors in the program that only apply to rare cases will not be caught immediately.
Predictive Validity. We have reviewed the key structural features of case mix systems, including a discussion of face validity. In reviewing the performance characteristics of case mix systems, we will consider another way of assessing validity—predictive validity. Four questions arise in determining predictive validity:

- What is being predicted?
- What influence does the sample have on the results?
- Is the prediction being carried out at the case or the hospital level?
- What statistical methods are being used to evaluate predictive power?

We will consider each of these questions in turn.

What is being predicted? Although we have specified that performance should be measured against the costs of necessary care, we have also pointed out that data on necessary costs are not available. Even cost data for care actually provided is limited; only Medicare maintains a large-scale system for defining costs rather than charges, and in many data sets only length of stay rather than charges is available. Nevertheless, such cost data, adjusted for interhospital differences in wages and input prices, are the more appropriate criterion variable. Because such costs are not available, a number of evaluative studies have been conducted using charges or length of stay as the criterion.

We have already noted that events occurring during treatment are sometimes used as part of a classification system. When this is done, the resulting system will inherently predict more accurately than a system based on the same assessment approach but excluding in-hospital events. Understanding and controlling for differences in predictive power resulting from such structural differences is essential to a fair interpretation of differences in predictive performance.

Even more misleading is the predictive performance of a classification system that includes within its classification scheme direct measures of the variable being predicted. Thus a classification based in part on the use of services will always appear to be a better predictor of actual charges or actual costs than a classification scheme excluding the use of services. Such a system may not, however, be a better predictor of needed resources.

What influence does the sample have on the results? The case mix adjustment systems reviewed in this report are relatively new. As a result, research on them, particularly on their performance, has been conducted largely, and in some cases entirely, by their developers. Frequently the same data has been used for both development and testing.
The relative maturity of a system can be gauged by whether or not it has been applied to data sets other than those used in the initial development work and whether or not the system has been tested and applied by research teams other than the team which originally developed the system.

The sample of patients and hospitals used to test the performance of a system is important. First, the performance of a system in classifying patients from just a few hospitals may give a biased indication of the performance of the system on a national sample of hospitals. This is especially true if all the hospitals in a sample are from the same region of the country or are about the same size or have the same level of teaching involvement. In such cases, the data set may have little variation in a variable such as practice style, which influences costs but is not measured by the classification system. In the small test data set, such parameters are effectively controlled by the lack of variation, and the explanatory power of the classification system is artificially inflated compared to its true power on a more diverse data set.

Second, the performance of a system in a group of “volunteer” hospitals may not represent its potential in randomly selected or mandated hospitals. All of the systems except PMCs were originally developed and tested on "volunteer" hospitals or (in the case of APACHE) on selected cases within “volunteer” hospitals. Often these “volunteers” have actually subscribed to the system in question for internal hospital management functions. In addition to the fact that these volunteers are likely to be unrepresentative of a cross section of hospitals, the motivation of the facilities to make the system work effectively is quite different from the motivation of facilities under a government reimbursement program.

Finally, predictive tests carried out on the same data set used in developing a system (even when that data set is “representative”) tend to overstate the performance one can expect of the system when it is transported to a new set of cases. This type of bias can be reduced, but not entirely eliminated, by the use of appropriately adjusted statistics; however, these adjustments are not always made in practice.

The possible bias introduced by the use of idiosyncratic data sets to evaluate case mix systems is particularly troublesome when trying to compare the performance of two systems, each developed and tested on a different data set. For this reason, comparisons between systems applied to a single data set are more informative than simple measures of overall system performance derived on different data sets.

*Is the prediction being carried out at the case or the hospital level?* At the case level one measures the ability of the system to predict cost of individual cases; at the hospital level one measures the ability of a
system to predict the average cost for a hospital. A system which predicts well at the hospital level need not predict well at the case level because the patient characteristics which vary most between hospitals will not necessarily be those which vary most when all patients from all hospitals are considered as a single sample. Hospital-level performance is more important because it tells whether the system can pay hospitals appropriately. On the other hand, good patient-level prediction is important to reduce opportunities for hospitals to selectively skim or dump patients. In general, reports on patient-level performance are more plentiful than reports on hospital-level performance.

*What statistical methods are being used to evaluate predictive power?* Comparing the performance of classification systems is not a straightforward statistical problem. It is easy to agree that the purpose of the system is accurate prediction of some variable (for example, necessary costs). Statistical methods that try to summarize the accuracy of a predictive system must find some way to attach a penalty (or loss function) to any inaccurate predictions the system generates. The choice of the exact form of this penalty or loss function is somewhat arbitrary. The most popular methods of assessing predictive power use two statistics: $R^2$ and the coefficient of variation (C.V.).

- $R^2$ measures the percentage of the variation in admission cost which can be “explained” by the classification system
- C.V. measures the amount of variation in cost among cases in a DRG or other classification category as a percentage of the average cost for that category

Both these methods can be described as “squared error” methods, in that they penalize a prediction error in terms of the sum of squared error. The process of squaring the error term puts extra weight on large prediction errors. This means that a system which is off by $\$1$ on each of 100 cases will look better than one which is exactly right in 99 cases out of 100 and off by $\$11$ in the last case.

Both of these statistics are highly influenced by whether a data set is “trimmed” by the removal of unusual cases. Since hospital-level reimbursement will be off by the sum of the error, not the sum of the squared error, and since extremely large prediction error cases are often those which could be taken care of with some form of outlier policy, some statisticians have suggested that squared error statistics may not be the most appropriate to use in evaluating case mix systems for PPS.

Some more “robust” statistical measures have been suggested, based on the sum of absolute error rather than squared error. Although these
methods are less vulnerable to the influence of outliers, almost none of
the research has employed these statistics. Recent exceptions include
some work by Mitchell et al. (1984, 1985) on applying DRGs to physi-
cians and work by Horn et al. (1985b). In each case, statistics descri-
ing average profit or loss are much more robust than the “squared
error” statistics reported in most studies cited here.

A final warning note on the interpretation of statistical measures of
predictive performance is in order. The problem of using the $R^2$ sta-
tistic as the ultimate test of the performance of a system is especially
suspect when the dependent variable, the thing being “explained” by
the case mix system, is NOT our concept of “truth” (namely, necessary
resources) but some proxy measure which contains some elements of
“truth” but also contains some elements of resource use which we do
NOT wish to explain.

CRITERIA FOR DIFFERENT SYSTEM USES

We have suggested that a case mix adjustment method can have
three uses within PPS: research, refining DRGs, and replacing DRGs.
In addition, there is a fourth use for case mix methods—as a tool in
hospital management. These four purposes demand different struc-
tural characteristics for a case mix method and different performance
criteria.

Research

In many ways, a case mix system for research has to meet less
stringent requirements than a system for payment. A research system
can use more cumbersome data collection procedures than a system for
payment. Payment systems need to be based on objective measures,
resistant to manipulation. In contrast, classification methods based on
more subjective measures can be used in research, provided that care is
taken in training and double-checking the coders. “Gaming” for finan-
cial profit is not an important risk in research, but if coders use the
ambiguities of a system to classify a code according to actual resource
use rather than necessary resources, the system’s performance can be
unrealistically inflated. In some applications, however, a research sys-
tem must meet higher standards than a payment system. For example,
if another case mix system is to be used in evaluating the performance
of DRGs, that system must have higher reliability and higher validity
than DRGs.
Refinement

A system for DRG refinement need not be as reliable as one for research, but it must be resistant to manipulation and it must produce appreciable improvements in the extent to which similar cases are classified together. A refinement system need not be able to classify all cases since unclassifiable cases can be paid under existing DRGs after appropriate recalibration.

Replacement

A system to replace DRGs must be able to classify all cases in addition to performing better than DRGs.

Management

Hospital managers often find case mix systems difficult to use because there are not enough patients in any category to establish statistically reliable performance measures. By contrast, the Medicare payment system only needs accuracy at the hospital level, where the numbers are larger. In consequence, managers may need a system which is more precise than is necessary for Medicare. On the other hand, managers may be able to tolerate a system which might be manipulated since they usually have control over the abstractors who might do the gaming. Finally, case mix systems used in hospital management may be part of quality review, utilization review, or cost containment efforts. A case mix method used in reimbursement need only predict necessary costs effectively.

In this report, we will consider how well each of the five alternative systems might work in evaluation, refinement, and replacement of DRGs. We will not give much attention to other uses of case mix methods in hospital management.
II. COMPARISON OF THE STRUCTURAL CHARACTERISTICS OF ALTERNATIVE SYSTEMS

This section reviews the structure of six case mix methods: Diagnosis-Related Groups (DRGs), APACHE, Disease Staging, MEDISGRPS, Patient Management Categories (PMCs), and Severity of Illness Index (SOII). The main focus of our discussion is the five non-DRG systems. However, we include a brief review of the DRG system for comparison. In comparing the structures of the case mix systems, we will examine the six structural characteristics discussed in the last section:

- Classification structure
- Variants
- Data requirements
- Face validity
- Subjectivity or vulnerability to manipulation
- Availability of documentation

DIAGNOSIS-RELATED GROUPS (DRGs)

Classification Structure

The DRG patient classification system assigns each discharged inpatient to one of 467 Diagnosis-Related Groups based on five ICD-9-CM\(^1\) coded discharge diagnoses, three ICD-9 coded procedures, discharge status, age, and sex. Three additional DRG groups are used for cases in which there is a mismatch between the diagnoses and the procedure(s) or for which no other DRG can be assigned. The 467 DRGs are organized under 23 Major Diagnostic Categories (MDCs), which give clinical structure to the classification system. The number assigned to a DRG is simply a label, so that a higher number DRG is not necessarily more complex or more costly to treat than a lower number. Under the PPS system originally implemented in October 1983, each DRG was assigned a relative weight based on the average cost of treating that DRG derived from the 1981 MEDPAR file and 1981 Hospital Cost Report data (Federal Register, September 1, 1983).

\(^1\) *International Classification of Diseases, 9th Revision, Clinical Modification*, the coding scheme currently in use for diagnosis and procedures.
Variants

Partly because of its fairly long history and partly because of its actual use in both the New Jersey and Federal prospective payment systems, the DRG system exists in many versions. As of 1983, four complete versions existed (Averill, 1983). These variant versions are solely of historical interest, so we review them only briefly. The first version, based on ICD-8 diagnosis codes, had only 383 groups and used a clinical system different from the current Major Diagnostic Categories for organizing these groups into clusters. These original DRGs were envisioned as a tool for utilization review. Thus, length of stay in hospital was chosen as the dependent variable for creating the groups.

A modification of these original ICD-8 groups was designated for use in the New Jersey reimbursement system. Then, the DRGs were recreated, using similar statistical methods as the original system, but a different clinical organizing principle (the Major Diagnostic Categories). These new DRGs used ICD-9, not ICD-8, diagnosis codes, and both length of stay and costs were used as dependent variables. A modification of the new ICD-9 DRGs was used when PPS began in October 1984. A fifth version has now been completed for use in PPS beginning in FY 1986. The changes embodied in this new version are documented in the Federal Register (September 3, 1985) and new versions of the computer program and documentation are available (Health Systems International, Inc., 1985).

Data Requirements

The DRG patient classification system uses a computer algorithm to assign each patient to one of the 470 DRGs based on Uniform Hospital Discharge Data Set (UHDDS) elements. These include up to five ICD-9-CM coded diagnoses, up to three ICD-9-CM coded procedures, discharge status, age, and sex. These data elements are coded by each hospital’s medical record personnel from the medical record shortly after each patient’s discharge. By limiting the classification system to UHDDS data, the burden of data collection to the hospital is reduced to those items already collected on a routine basis. However, this limitation means that important indicators of patient condition or need for care are not available for use in assigning DRGs.

---

2The AUTOGRP statistical algorithm, a modification of the Automatic Interaction Detection (AID) program (Sonquist and Morgan, 1964), based on least-squares regression.
Face Validity

The current version of the DRGs is a product of both clinical and statistical input. First, 23 clinically based Major Diagnostic Categories (MDCs) were developed using physician input. Then, the AUTOGRP statistical algorithm was used to group diagnoses within MDCs into groups which would explain the most variance in length of stay (LOS). The same statistical methods were then used to further divide the groups based on the power to explain variation in costs. Although clinical judgment sometimes overrode statistically based decisions, the statistical information was always provided and used interactively in the decisionmaking process of creating these groups (Fetter et al., 1982; Averill, 1983).

The DRG system creates separate categories for patients with different clinical problems. The determination of the clinical problem is based on the ICD-9 coded discharge diagnoses. Within a cluster of diagnoses, other variables may be used to subdivide patients. For example, the presence of an operating room procedure was found to significantly alter LOS independently of diagnosis. Thus different DRGs are used to categorize patients in a given diagnostic cluster who did or did not undergo operating room procedures.

The use of procedures to define the DRG classification of some patients undoubtedly explains variation in both actual costs and length of stay. This increase in explanatory power is achieved by incorporating what is done for the patient, rather than using independent indicators of what is needed by the patient. Given the considerable regional variation in use of procedures (Wennberg and Gittelsohn, 1973; McPherson et al., 1982; Barnes et al., 1985; Chassin et al., 1986), it remains to be seen how well actual procedure use correlates with necessary procedure use. On the other hand, the problem of specifying independent indicators of patient need is not an easy one to solve. Ongoing research in this area is characterized by lengthy lists of possible indications for each procedure which can be difficult to obtain even from the complete medical record using specially trained abstractors (Chassin et al., 1986).

The DRG system also uses certain secondary diagnoses (complications or comorbid conditions) in classifying patients. Thus, different DRGs may be used to categorize patients with the same principal diagnoses with or without a comorbid condition. In the DRG system, the list of complications and comorbidities is not diagnosis-specific. In addition, since the DRG system is based on data abstracted after the patient's discharge, it is not always possible to distinguish conditions acquired during the hospitalization from those present at admission.
Therefore, the DRG system may not perform as well in predicting needed costs as a system in which diagnosis-specific complications and comorbidities are incorporated or which preexisting conditions are distinguished from those acquired during the course of hospitalization.

Other major variables used to differentiate patients with like diagnoses into DRGs are age and whether the patient was discharged alive. Age is used only as a categorical variable to indicate whether a patient is under 18 or over 70 years. In addition, some DRGs are defined on the basis of age over 70 or the existence of a complication or comorbidity. This limited use of age probably reduces the ability of the DRG system to explain the costs for very old or frail elderly patients (Berman and Pawson, 1984). The death of the patient is also restricted to use in DRGs where patients who die are less costly than patients who live. A more general use of patient death might improve the explanatory power of the DRGs but could also create a moral hazard in the form of higher payment for patients who are discharged dead.

The DRG classification system, by being constructed on a limited data set, has less intuitive clinical appeal than a system which also incorporates patient physiologic data or other clinical information. The use of clinical and statistical inputs in the creation of the original groups does not seem unreasonable. Although some groups are of questionable clinical homogeneity (for example, DRG 14 “Specific Cerebrovascular Disorder Except T.I.A.” [Smits, Fetter, and McMahon, 1984]), this is probably more a function of the decision to limit the DRGs to a number less than 500. Expanding the number of groups based on the same set of input variables would improve the clinical acceptability of the system, but would not be likely to affect the statistical performance to any large degree.

A more serious concern about the validity of the DRG system comes from its reliance on actual lengths of stay and costs in the construction of the system. Ideally, a measure of necessary costs would have been used. In fact, no case mix system has been able to achieve this ideal, and in relying on length of stay and costs (rather than just length of stay or charges), the DRG system has a better pedigree than many other systems.

**Subjectivity or Vulnerability to Manipulation**

The vulnerability of DRGs to manipulation lies principally in the coding of the discharge diagnoses and procedures. More aggressive coding of secondary diagnoses may lead to a higher fraction of cases in DRGs with complications and comorbidities. In addition, consistent “upcoding” of marginal cases can lead to higher case mix indices with
the same mix of cases (Simborg, 1981). DRGs have been tested on a number of large data sets representing discharges from whole states (New Jersey and North Carolina [Mitchell et al., 1984], Michigan and Washington [Mitchell et al., 1985], and Maryland [Coffey and Goldfarb, 1984]) and a sample of all Medicare cases nationally (Cotterill, Bobula, and Connerton, 1985). However, since these data sets were created from claims submitted for payment, they contain only one abstraction of the UHDDS data and so have not permitted an estimate of the degree to which DRG classification based on a second abstraction of the same case would agree with the classification based on the first.

One study using both Health Care Financing Administration (HCFA) and Commission on Professional and Hospital Activities (CPHA) data found evidence that the case mix index\(^3\) has increased since 1981 (Carter and Ginsburg, 1985). There appear to be two components to this increase: first, a consistent trend over time, and second, a one-time increase at the time hospitals first went on PPS. The long-term trend appears to be related to decreasing hospitalization rates and increased use of outpatient settings for simple cases, as well as a movement to higher technology treatment modalities for some diseases. The one-time jump seems to come from more “cost-conscious” coding. There may be room for more “enrichment” in the DRG case mix index through coding changes. Even after the beginning of PPS, abstracts coded by hospitals for CPHA yield a 3 percent higher case mix than abstracts coded by the same hospitals for HCFA. This somewhat counterintuitive finding may be because the coding for CPHA is done at a later time after discharge, when the medical record is more complete or because HCFA coding is being done by business office rather than medical records personnel at some hospitals.

**Availability of Documentation**

Extensive documentation for the DRG system is in the public domain. In addition to published articles (Fetter et al., 1980), and final reports of grant projects (Fetter et al., 1982), the PPS system itself is documented in the Federal Register (September 1, 1983; January 3, 1984; September 3, 1985). Detailed descriptions of the workings of the DRG assignment algorithm and a computer program to implement the algorithm are available from a private company (Health Systems International, Inc., 1985).

\(^3\)The case mix index for a hospital is the average DRG cost weight summed over all Medicare discharges from the hospital. The relative cost weights for each DRG have been standardized so that the average case mix index over all hospitals equals 1.0.
APACHE

The APACHE system was initially designed to assess the severity of illness (and probability of death) in intensive care unit (ICU) patients. The intended use of APACHE was to adjust for severity of illness when evaluating patient care (Knaus et al., 1981). The APACHE system has gone through several rounds of development and evaluation and has been proposed for possible use in reimbursement only recently (Wagner and Draper, 1984). In this review, we will focus on the current version of APACHE (called APACHE II) and its physiologic subscale (the APS12) (Knaus et al., 1985; Wagner and Draper, 1984).

Classification Structure

The APACHE system classifies patients, rather than classifying a disease or a diagnosis. The APACHE score is intended to assess the criticality of the patient's condition, independent of diagnosis.

The current version (APACHE II) builds a total score from three components (Knaus et al., 1985):

- A physiologic score (APS12) based on the values of 12 variables including vital signs, blood test, and blood gas results collected within 24 hours of admission to the Intensive Care Unit
- Age points for five categories of age
- Chronic health evaluation points awarded for patients with a history of organ insufficiency according to a list of conditions

Each physiologic variable contributes points to the overall score in a definitive fashion depending on how far the patient's value deviates from “normal.” The overall score is a cardinal scale with values ranging from 0 to a theoretical maximum of 71. So far, 55 is the highest observed value. The physiologic component contributes 60 of the 71 points (Knaus et al., 1985; Wagner and Draper, 1984).

Variants

The major variant of APACHE is historic. The original APACHE system with 34 variables in its physiologic component has been replaced by APACHE II, with its 12 variable APS physiologic score.

Data Requirements

APACHE requires physiologic and test data from the first 24 hours of admission to the ICU. These data elements are not now routinely abstracted in most hospitals and customized abstraction forms are used
to collect the data. The abstraction costs could be reduced if some of the APS12 elements could be obtained directly in machine-readable form, for example, directly from an automated laboratory reporting system.

In the long run, as hospitals put more and more patient information on a centrally accessible computer system, a system like APACHE (or MEDISGRPS) might be redesigned to eliminate the need for special abstraction forms. The clinical data could be extracted from patient data files with computer programs. Such technological changes in storage and retrieval of medical record information could dramatically reduce the marginal cost of implementing case mix adjustment methods using physiologic or other clinical data. At present, however, APACHE and other systems using physiologic data would place a burden of abstracting and reporting data not now routinely reported on each discharge.

Face Validity

APACHE has more distinctly clinical origins than the other systems reviewed here. Because APACHE was designed to assess ICU patients, the system tended to equate severity of illness with probability of death (Knaus et al., 1981; Knaus et al., 1985). This equation may be less useful when assessing patients suffering from conditions with extremely low mortality rates. The ICU focus of APACHE explains the timing requirements attached to the data (within 24 hours of admission to the unit) and the emphasis on physiologic parameters which ought to be related to life-threatening events.

The first round of the APACHE system was developed by clinicians who identified important variables and estimated relative weights of abnormal values. The initial physiologic score relied on 34 variables (Knaus et al., 1981). Subsequent clinical reviews bolstered by statistical analyses, using mortality as the yardstick, have helped winnow this down to 12 (Wagner and Draper, 1984; Knaus et al., 1985).

It has been argued that extension of APACHE to the prediction of inpatient costs is justified by the high proportion of hospital costs devoted to critical care. This argument is problematic for four reasons. First, despite the fact that a high proportion of costs are devoted to ICU care, only 11 percent of Medicare admissions include time in the ICU.4 Second, because the system was developed and tested on critical care cases only, the question of how well APACHE will work on non-ICU cases is still open. Presumably, non-ICU patients would have

---

4Based on a sample of FY 1984 Medicare admissions.
fairly low APACHE scores; however, this remains to be demonstrated. Third, the sickest patients may demand the most intense resources while they survive, but very sick patients who die soon after admission will not consume as many resources per admission as less critical patients who remain in the hospital for an extended period. Finally, the question of treatment goals will affect how much is done for “terminally ill” patients. Aggressive treatment will be more expensive than palliative care, even when the outcomes are no different (Garber, Fuchs, and Silverman, 1984).

In sum, the APACHE system, despite its methodical development, cannot lay strong claims to its face validity as a predictor of inpatient costs for all kinds of patients. Its applicability to critical care cases is far more obvious, but even here, the case is stronger for APACHE as a predictor of clinical status than as a resource need predictor.

Subjectivity or Vulnerability to Manipulation

The variables used in the physiologic and age components of APACHE are quite objective. Only the chronic health evaluation points leave scope for interpretation on the part of the abstractor. The relatively unambiguous nature of laboratory findings is a strong argument for their use in case mix classification.

Availability of Documentation

The APACHE system is well documented (Knaus et al., 1981; Knaus et al., 1985), and has been implemented in ICUs around the world (Wagner et al., 1984). In particular, the APS12 component of the score is simple to describe and implement.

DISEASE STAGING

Classification Structure

DRGs assign patients to case mix categories. In contrast, the concept of Staging applies to a disease, not to a patient. The Disease Staging classification system consists of two parts. First, the disease to be staged is assigned to one of 420 Staged Disease Categories (SDCs) according to its etiologic and pathologic attributes. Each of these 420 SDCs is then divided into four stages with the highest stage being death. When Staging is used to classify patients, rather than diseases, a third step is necessary to classify patients who have more than one SDC.
The Staging concept was developed in relation to clinical oncology, where stages represent progressive degrees of disease extension and generally correlate with morbidity and mortality. Disease Staging preserves this meaning (Gonnella, Hornbrook, and Louis, 1984). The categories in the Disease Staging system are:

- Stage 1 – conditions with no complications
- Stage 2 – problems limited to an organ or system
- Stage 3 – multiple site involvement
- Stage 4 – death

For each disease, criteria for each stage are specified.

The stages are ordinal with respect to severity of disease, but because Staging criteria are disease-specific, they are not directly comparable across diseases. Stage 2 hearing loss represents a different level of illness from Stage 2 breast cancer (Gonnella, Hornbrook, and Louis, 1984; Conklin et al., 1984b).

The Staging software assigns stages to all diagnoses for a given patient. Then, the patient’s assigned staged disease may be selected in two different ways. The most common method assigns the “underlying staged disease.” This is the highest staged disease related to the principal diagnosis (based either on the principal diagnosis itself or on any secondary diagnosis related to the principal diagnosis). Information concerning staged conditions unrelated to the underlying staged disease may or may not be retained for further study (Louis et al., 1983). A second method selects only the staged principal diagnosis (Coffey and Goldfarb, 1984).

Variants

The two major Staging variants are the manual and computerized versions. In addition, within the computerized version of Staging, the software has been modified to reflect alternative methods for handling principal diagnosis Staging, underlying disease Staging, and the Staging of unrelated secondary diseases (Louis et al., 1983; Conklin et al., 1984a,b).

In both computerized and manual Staging, substages are often specified as well, reflecting clinically detectable differences in severity. Substages are supposed to follow the rule of the major Staging categories with higher numbers representing a higher risk of morbidity and mortality (Gonnella, 1983). In practice, some substages do not seem to be necessarily ordinal; they simply specify the locus of the disease more precisely. Substages are rarely used in analyses of computerized Staging.
The original manual Staging could be used to assess stage at three points in a hospital stay: on admission, "peak" stage, and at discharge (Arthur Young and Company et al., 1983). The computerized version generates a single assessment of stage, one which is probably closer to the "peak" or "discharge" stage than to "admission" stage. However, Richards et al. have noted, in over 4000 admissions, that admission stage almost always equals peak stage.

Data Requirements

The manual and computerized versions of Staging have different data requirements. The manual format provides detailed criteria for assigning stage and substage based on a review of the entire medical record. The computerized version assigns stages and substages based on the UHDDS diagnoses, procedures, sex, and discharge status.

Face Validity

The Disease Staging system was developed with several potential applications in mind, including utilization review, patient care evaluation, hospital management, and prediction of morbidity and mortality. The genesis of the system was entirely clinical, with panels of physicians developing both the definitions of the 420 SDCs and the criteria for assigning stage and substage within each (Gonnella, Hornbrook, and Louis, 1984). The diffuse goals of Disease Staging and the dominance of clinical over statistical processes make it difficult to argue that the Staging system was designed to be (or ought to be, on the face of it) predictive of a patient's needed resources or even of actual resource consumption. The Staging system should predict resource use only to the extent that more severely ill patients will require more resources.

Subjectivity or Vulnerability to Manipulation

As with DRGs, computerized Staging is vulnerable to changes in criteria for coding diagnoses and procedures in the Uniform Hospital Discharge Data Set. The same kinds of pressures would exist under a prospective reimbursement system based on computerized Staging as under the DRG-based PPS. This includes incentives to find all possible secondary diagnoses and to resolve all marginal diagnosis codes by putting a case in the "most serious" category possible. The impact of such policies on "Staging creep" undoubtedly differ from their impact on "DRG creep." One might expect a greater amount of upward drift
in Staging than in DRGs if the “underlying staged disease” method of
Staging is used. So far, however, there has been no documentation of
the magnitude of “Staging creep.”

Potential problems of vulnerability to manipulation are even greater
in manual Staging than in computerized Staging. There is no informa-
tion available on the reliability of manual Staging, even under research
conditions, much less under conditions of use in payment. In fact, the
manual Staging system is not widely viewed as a viable system for use
in reimbursement for other reasons—principally, the special training
requirements for coders of manual Staging and the additional data col-
lection burden this system would impose on hospitals.

Availability of Documentation

Staging computer software and documentation are available for pur-
chase. A complete set of manual Staging criteria is also available.
Several researchers besides the original developers of Disease Staging
have used either the manual (Arthur Young and Company et al., 1983)
or computerized versions (Coffey and Goldfarb, 1984; Calore, 1985a).

MEDISGRPS

Like Disease Staging, MEDISGRPS was designed with diffuse goals
including utilization review, patient care evaluation, hospital manage-
ment, and prediction of morbidity and mortality (Brewster et al., 1985).
A relatively new case mix system, MEDISGRPS has yet to undergo the
development and testing of some of the older systems.

Classification Structure

The MEDISGRPS classification system assigns a patient to one of
five ordinal severity levels based on a computerized algorithm that
determines the highest severity level based on “key clinical findings.”
The levels are labeled 0 through 4, with higher numbers representing a
higher probability of failure of one or more organ systems. The
MEDISGRPS level, like that of SOII and APACHE, applies to the
patient rather than to the disease, and can be used to compare patients
within or across diagnostic categories (Brewster et al., 1985).

The MEDISGRPS system includes special abstraction forms and a
computer program designed to process the information from the forms
to generate a MEDISGRPS score. The abstraction form can accom-
modate information on up to 500 different findings, including physical
exam items, laboratory, x-ray, and pathology findings, intubation and
transfusions (Brewster et al., 1984a,b; 1985). Data are coded from the medical record for those findings which are present in a particular case and the system generates a score based on whatever variables happen to be available.

MEDISGRPS is designed to use two reviews. The first is review “on admission” (actually 48 hours after admission to allow relevant information to reach the patient’s chart). The second review is conducted at the tenth hospital day or the sixth post-operative day if surgery has been performed. At the time of the second review, data are also collected on morbidity during the course of treatment or discharge status (for those who have already been discharged) (Brewster et al., 1985). For patients with lengthy hospitalizations a third review may be carried out (MediQual Systems, 1984).

**Variants**

The MEDISGRPS system has not had time to spawn variant systems characteristic of the older case mix systems. The major variation is the possibility of implementing the system with only one review, at the time of discharge.

**Data Requirements**

The full version of MEDISGRPS is designed to carry out concurrent review of patients still in the hospital. In this version, two reviews of each patient’s medical record are required (Brewster, Bradbury, and Jacobs, 1985; Brewster et al., 1985). The estimated time for abstraction and data entry is about 12 minutes per case (Brewster, Bradbury, and Jacobs, 1985). It is possible that one review would suffice for a reimbursement application, but further work is needed to evaluate the feasibility of modifying the system in this way.

The costs of implementing this system include training abstractors to handle new elements and then abstracting those elements on all Medicare cases. In addition, the elements have to be converted into a score or grouping using a computer algorithm, thereby generating additional data processing costs. Given the large amount of input and the flexibility of the algorithm in its data requirements, the computational costs of this system could be quite large, compared with the computational costs of some of the UHDDS-based systems such as DRGs or Staging.

Long-range changes in technology may improve the accessibility of physiologic and other clinical data, reducing the costs of implementing MEDISGRPS (or APACHE). In the short run, however, case mix
adjustment systems such as MEDISGRPS place a significant data collection burden on the hospitals.

**Face Validity**

The MEDISGRPS system, like all the other case mix methods reviewed here, used expert judgment in conceptualizing the system and defining the criteria for each classification level. Associating the levels with increasing probability of organ failure is an appealing variation to simply using mortality as a measure of severity. On the other hand, the resources required to reverse or delay organ failure seem likely to depend on which organ is failing and on the mechanism of failure. Viewed in this way, the MEDISGRPS levels by themselves are likely to be a somewhat noisy predictor of needed resources. It may be possible to improve the specificity of this system by using it in conjunction with DRGs or some other diagnosis-based classification system.

**Subjectivity or Vulnerability to Manipulation**

Two features of the MEDISGRPS system seem to lend themselves to manipulation. The MEDISGRPS algorithm appears to be very flexible in its data requirements and unlike many systems, is prepared to handle missing data (Brewster et al., 1985). While this flexibility is a virtue in research or internal hospital management functions, it may be a liability in reimbursement. For example, one needs to verify that the system does not raise a patient's classification based solely on the number of tests done. Since the algorithm is designed to take the highest value when multiple values of a test are available (Brewster et al., 1985), such quantity-related creep seems possible.

The second feature is reliance on subjective data such as physical findings or radiology interpretations. Unlike the more objective laboratory results, these data could be manipulated inappropriately under reimbursement incentives.

**Availability of Documentation**

As a new, proprietary system, MEDISGRPS has yet to be completely documented. Full descriptions of the computer algorithm are not available, apparently not even to subscribers to the system. This lack of documentation is a serious problem for those who wish to evaluate the appropriateness of the system for possible use in a public program.
PATIENT MANAGEMENT CATEGORIES (PMCs)

Patient Management Categories are more comprehensive than any of the other case mix methods presented so far in attempting to capture needed resources. The system was developed under the auspices of Blue Cross of Western Pennsylvania, with the problem of hospital reimbursement specifically in mind (Young, Swinkola, and Zorn, 1982).

Classification Structure

The PMC system is a clinically based system that assigns patients to one of 800 Patient Management Categories based on their treatment and diagnostic needs (Young, 1985). PMCs were designed to create patient groups that are similar not only in expected resource use, but also similar in the components of care contributing to that total.

PMCs are grouped under 47 modules, each representing a disease or disorder group. The first two digits of the PMC number refer to the module, and the second two digits identify a particular PMC within the module. Although the number assigned to a PMC does not reflect any hierarchy of categories or modules, there is a hierarchy of categories within each module.

The PMC system also specifies the "components of care," a set of services needed to provide effective care for a typical patient in a PMC. These components of care were constructed by panels of physicians, and a normative "cost weight" is available for each PMC. In some PMCs, the cost weight is allowed to vary by patient discharge status (dead or alive). PMCs can also be used, like DRGs, with empirically-derived cost weights.

Variants

Two versions of PMCs have been developed (Young, 1985). An early version required data on the Reason for Admission (chief problem or specific elective procedure) in addition to UHDDS data. The two versions appear to function similarly. In comparing the two versions of PMCs, Young found the agreement in classification with and without reason for admission was just under 90 percent.

Data Requirements

As noted, the current version of PMCs operates on UHDDS data.
Face Validity

PMCs consist of two separate elements: (1) the PMC classification system and (2) the normative "cost weights" based on components of care. The validity of each of these two components needs to be considered separately. The cost weights by themselves can have no meaning unless the classification system is valid; however, the reverse is not necessarily true. The classification system could be valid, even if the cost weights bear no relation to true resource needs.

The PMC classification system was specifically developed to group patients according to their need for diagnostic and therapeutic services. This clearly distinguishes PMCs from other systems under review which were designed to explain clinical severity or length of stay. If the system is able to carry out this original intention, then the PMC classification system ought to be an ideal case mix adjustment method for use in reimbursement.

PMCs assign patients to categories using combinations of diagnoses and, when necessary, procedures. The order in which diagnosis codes are listed does not change the categorization process. Like DRGs, PMCs use procedure codes to complete the assignment of a patient to a category; approximately 36 percent of the categories use data on procedures. Age and sex are also used in a few instances.

There are two questions concerning the face validity of PMCs. The important question is the degree to which PMCs (like DRGs) are able to distinguish actual from needed resource use. A second question is whether PMCs are able to identify patient groups with similar diagnostic and therapeutic needs using limited input data. While there are good arguments for restricting the classification to operate on UHDDS data (or a slight expansion of UHDDS), it may be that any patient classification scheme developed on such a small number of elements will look and perform similarly in predicting costs, length of stay, mortality, or needed resources, regardless of the original conceptualization. We will return to this issue in Sec. IV when we discuss the predictive performance of the various systems.

The validity of the cost weights will depend on the degree to which expert panels were able to reliably identify needed components of care and the appropriateness of the methods used to convert a vector of components of care to costs. Since PMCs are the first comprehensive effort to assign normative costs to patient groups, we cannot at this point even estimate the reliability of the method (that is, the degree to which different expert panels might arrive at similar estimates), much less the validity.
Subjectivity or Vulnerability to Manipulation

As with other UHDDS-based systems, the reliability of PMCs depends on the accuracy of the discharge data abstract. As with DRGs and computerized Staging, PMCs will be vulnerable to changes in criteria for coding diagnoses and procedures in the Uniform Hospital Discharge Data Set. PMC-based prospective payment would exert the same type of pressure to "code up" as did the DRG-based PPS. This includes incentives to find all possible secondary diagnoses and to resolve all marginal diagnosis codes by putting a case in the "most serious" category possible. "PMC creep" would differ in its precise form from "DRG creep" or "Staging creep," but it would no doubt occur.

Availability of Documentation

Because the PMC system was developed under a research grant and because the system was intended for use in reimbursement, both the computerized algorithm and written documentation of the categories are available (Young, 1985). The system is a complex one, however, and the real test of the adequacy of this documentation will be the ability of other users to implement the system.

SEVERITY OF ILLNESS INDEX (SOII)

The Severity of Illness Index (SOII) was developed more recently than DRGs or Staging. The system has been actively developed and promoted as both a case mix management system for hospitals (especially teaching hospitals) and as a possible refinement for the DRG system for use in PPS.

Classification Structure

The Severity of Illness Index is applied to a patient, rather than to a disease. Each patient receives an overall score (an integer between 1 and 4) intended to rate the patient's "burden of illness" (Horn, Horn, and Sharkey, 1984).

Assignment of the overall severity level to a patient is based on a specially trained abstractor's review of the medical record. This overall level is derived by the rater's "implicit integration" (Horn, Sharkey, and Bertram, 1983; Horn, Horn, and Sharkey, 1984) of seven severity dimensions, each of which has four levels. (Later reports [Horn et al., 1985a,b; Horn and Horn, 1986] indicate the overall rating is usually "consistent with" the mode or the average of the seven dimensions.) The dimensions are:
• Stage of the principal diagnosis (NOT the same as the Disease Staging assignment)
• Complications of the principal diagnosis, or in-hospital events
• Concurrent interacting conditions affecting hospital course
• Dependency on hospital staff
• Extent of nonoperating room life-support procedures
• Rate of response to therapy or rate of recovery
• Residual impairment (Horn et al., 1985a)

The four levels of SOII are on an ordinal scale (Horn et al., 1985a). Level 4 is more severe than Level 3 but the distances between successive levels do not necessarily represent equal changes in severity. In some applications, each level of the SOII is further divided into three parts: without an operating room procedure, with a moderate operating room procedure, and with a major operating room procedure (Horn et al., 1985a). There is no simple ordinal rule for ranking the 12 categories in this “Procedure-adjusted SOII” (Horn et al., 1984; Horn et al., 1985b).

SOII can not be assigned independently of diagnostic information, since the assignment of the SOII severity level requires identifying the principal diagnosis, assessing its stage, and discriminating its complications from complications due to other factors (Horn and Horn, 1986). However, in contrast to the stages in Disease Staging, it is claimed that SOII is not disease-specific, enabling one to compare severity levels in Severity of Illness across different diagnoses (Horn et al., 1985b). In practice, the SOII is usually applied to patients after they have been grouped using a disease categorization scheme (Horn, 1985; Horn et al., 1985a,b; Horn and Sharkey, 1983).

Variants

Published studies on SOII all use one or more versions of a manual Severity of Illness scoring system.

Within manual SOII, the major variant is the “procedure-adjusted SOII.” In this variation, each level of the four-level overall severity score is divided into three categories: no operating room procedure, moderate operating room procedure, and major operating room procedure.

A computerized version (CSI, the Computerized Severity Index) is now being developed and tested. In this discussion, we will look at the earlier, manual systems, with some comment on how the new CSI differs from the old SOII.
The Computerized Severity Index has quite a different structure from the manual SOII (Horn, 1986). The Computerized Severity Index adds a sixth digit to the five digit ICD-9-CM code. This sixth digit severity code is a four-level code similar to that used in SOII. Unlike SOII, the criteria for coding CSI severity levels do depend on the patient's specific disease conditions. The criteria for coding severity differ for different five-digit diagnosis codes. In all, there are 700 different sets of severity criteria for use with the over 10,000 ICD-9-CM codes.

The sixth severity digit is added to every diagnosis coded in a particular case—the principal and all secondary diagnoses. Conceptually, the CSI is a weighted combination of the sixth digit severity codes and a "rate of response to therapy variable" (Horn and Horn, 1986). In principle, the weights can depend on what the principal and secondary diagnoses are. The problem is how to specify the appropriate weights.

Based on a sample of 2200 cases with the 60 most common diagnoses, a weighting system was developed to combine the sixth digit severity code for principal diagnosis and secondary diagnoses into an overall CSI score. A comprehensive software package containing all sixth digit criteria sets and weighting rules is under development (Horn, 1986).

The Computerized Severity Index is so different in concept and implementation from the manual SOII that it may be better to think of it as a separate system, rather than as a variation. An additional caveat: the additional data collection burden that would be imposed by this computerized case mix system greatly exceeds that required by other computerized systems such as Staging and DRGS.

Data Requirements

The Severity of Illness Index differs from the other five systems both in the type of data input it requires and in the use of subjective human "integration" to arrive at the overall Severity level. The SOII requires as input a four-level code for each of seven subscales coded by specially trained abstractors who review the patient's medical record. The coding of these scales, unlike the coding of the data elements of APACHE or MEDISGRPS, does not depend on locating a single clinical finding in the chart, but rather on the subjective integration by the abstractor of the entire content of the record. The developer reports that the abstraction takes a trained coder "a few" minutes per chart (Horn, 1986). Other sources estimate that an average hospital would have to hire two additional abstractors (Nathanson, 1986).
The Computerized Severity Index, unlike computerized Disease Staging or DRGs, is not simply a system for converting currently collected UHDDS elements into a case mix categorization. The input data for CSI, like that for SOII, requires the collection of new data, not now collected by most hospitals. CSI is based on different data elements from manual SOII (a new sixth digit on the diagnosis codes, rather than a multidimensional severity scale). From the sample criteria now available (Horn, 1986), the criteria for coding the sixth digit seem to be much more explicit than the old SOII criteria. Still, the problem of retraining abstractors to implement a new chart coding procedure invoking 700 separate sets of criteria is not trivial. Until more information about CSI is available, including a complete code book, it is difficult to estimate the additional data collection burden the system would impose, or the possible benefits that would accrue.

Face Validity

The Severity of Illness Index was developed at Johns Hopkins as an adjustment to DRGs for use in determining the hospital's reimbursement under the Maryland prospective payment plan introduced in the mid-1970s (Horn, 1985). Under that plan, each hospital was allowed to pick a case mix measure which would be used in rate-setting.

The origins of the system notwithstanding, SOII was not conceptualized as a direct predictor of resource consumption. Rather, SOII was designed to measure the patient's burden of illness by addressing the question: "How sick is the patient?" (Horn et al., 1984). If sicker patients need more resources, the SOII can be expected to predict resource need.

The seven dimensions of the original SOII have an intuitive appeal. Rate of response to therapy, interacting conditions, dependency, and complications are all factors which, on the face of them, could complicate a patient's recovery. Problems arise, however, in trying to measure these concepts. The concepts themselves are not precise. Explicit criteria general enough to apply to all patients are difficult to define. What was done for a patient may be hard to distinguish from what a patient needed to have done. The overall score is most often used in reporting data about the SOII. Yet there is little known about the process by which this overall score is derived, except that it is most frequently "consistent with" the mode or average of the seven dimensions of severity (Horn et al., 1985a,b; Horn and Horn, 1986).

There is a certain circular logic to the appeal of the SOII: a patient for whom life-support systems are used, who gets extra nursing care, and takes longer to recover (that is, has a prolonged length of stay)
gets a higher severity code. The higher severity code is justified because patients with higher codes use more resources (including life-support systems, nursing care, and longer stays) (Cretin and Buchanan, 1985).

Comparatively little has been written about the conceptual basis for the CSI or the details of the development of the criteria sets. Until this system is available for review, its content validity (or other characteristics) cannot be judged.

**Subjectivity or Vulnerability to Manipulation**

The SOII relies far more than other case mix methods on subjective and implicit judgments by the coder. When the subjectivity of the method has been criticized, the criticism has been deflected by pointing to the overall reliability of the system. Indeed, the overall rate of agreement between pairs of raters has been reported variously at between 90 and 95 percent, after two months of experience (Nathanson, 1985; Horn and Horn, 1986; Horn, Horn, and Sharkey, 1984).

Reliability reports of SOII when implemented in a research or hospital management application are beside the point, however. The as yet unanswered question is whether the system would be vulnerable to manipulation by abstractors with incentives to either code up (hospital-based coders) or code down (HCFA-based coders).

A secondary concern is that the relatively high overall reliability of the system may be misleading. The SOII system is only a four-level system, with less than 10 percent of cases in the third and fourth level, so high degrees of interrater reliability ought to occur by chance alone. We will return to a more detailed discussion of the reliability of SOII in the next section.

The development of the Computerized Severity Index is not yet far enough along to assess its vulnerability to manipulation.

**Availability of Documentation**

Much has been written about SOII and its predecessor system, the AS-SCORE. All of these reports have come from the original developer of the index (Horn, Rovetti, and Kreitzer, 1980; Rovetti, Horn, and Kreitzer, 1980; Horn, 1981; Horn and Schumacher, 1982; Horn, 1983; Horn, Sharkey, and Bertram, 1983; Horn, Chacich, and Clopton, 1983; Horn and Sharkey, 1983; Horn et al., 1984; Horn, Horn, and Sharkey, 1984; Horn et al., 1985a,b). Because the system depends on the ability of the abstractors to reliably code the SOII score, the key to implementing the system is the training of the abstractors. Little documentation on abstractor training has been publicly available.
The CSI is not yet documented. A complete code book for the sixth-digit system and a complete description of the weighting procedures have not yet been released.

SUMMARY OF COMPARATIVE STRUCTURAL FEATURES OF COMPETING SYSTEMS

The case mix adjustment methods reviewed here vary widely in their historical purposes, their developmental methods and their structural characteristics. At the same time some similarities emerge.

Three of the systems (DRGs, Staging, and MEDISGRPS) were developed from the start with diffuse goals. The developers of the three systems each sought a case mix adjustment that could be used in utilization review, hospital management, and patient care evaluation. Within these three systems, Staging (especially manual Staging) and MEDISGRPS have the strongest clinical flavor, whereas DRGs reflected a stronger “resource management” bias, emphasizing length of stay and charges, rather than clinical considerations in the early development. PMCs and SOII were developed in response to the perceived weaknesses of existing case mix methods used in a reimbursement context. APACHE is the most clinical of all the systems in its origins, emphasizing mortality and morbidity of critical care patients.

The development of the systems differed in the extent to which clinical versus statistical considerations were used to shape the patient groupings. The more focused systems which were able to identify a single criterion measure tended to make greater use of statistical techniques in designing and modifying the case mix system (DRGs using length of stay and APACHE using mortality).

DRGs, PMCs, and Staging approach case mix adjustment by categorizing patients into disease categories based on diagnoses, procedures, and age. In DRGs and PMCs, the case mix adjustment begins and ends with this assignment to a category. In Staging, the patients within a disease group are further divided on a four-part ordinal scale (the Stage). The Staging concept applies to the disease, not the patient, so that the interpretation of the Stage differs for different disease groups.

Severity of Illness Index and MEDISGRPS rely on an ordinal scale (a “severity level”) which is applied to patients, independent of their particular disease. In practice, both SOII and MEDISGRPS are used in conjunction with DRGs or some other diagnostic classification. The APACHE system also attaches a score to a patient (rather than to a disease). The APACHE II score is finer grained than the four- or
five-level ordinal scales used by SOII and MEDISGRPS, with theoretically attainable values from 0 to 71.

DRGs, Computerized Staging, and PMCs all have versions which limit the input data used in creating the case mix classification to UHDDS data. Manual Staging, MEDISGRPS, APACHE, and Computerized SOII developed definite criteria which could be obtained from the medical record by special abstraction. Of these systems, APACHE II uses the most limited data set, relying principally on the results from a dozen physiologic measures and age. The Computerized Severity score requires the implementation of a new system of diagnosis and procedure coding based on explicit criteria. Only manual SOII relies on rather general criteria and implicit integration of information by the coder.

All of the UHDDS-based systems (DRGs, computerized Staging and PMCs) have a computerized version of the case mix assignment algorithm. MEDISGRPS, APACHE, and Computerized Severity Index also have computerized algorithms. However, because these programs require the collection of input data not routinely available from all hospitals, implementation of the methods places an additional data collection burden on most hospitals.

Currently available case mix methods, then, differ substantially in their structure and approach. In the next section we will review the reliability and predictive validity of these systems.
III. COMPARING THE PERFORMANCE OF CASE MIX METHODS

In comparing the performance of the case mix systems, we will look at two characteristics—reliability and predictive validity. We first consider various indications of the reliability of the case mix methods and then look at their predictive performance.

RELIABILITY

We have already noted the problems that bedevil comparisons of the predictive performance of one case mix system with another: the lack of a “gold standard” against which to measure the systems, different samples, different levels of analysis (hospital versus patient), and different statistical criteria. Similar, although not identical, problems arise when trying to compare reliability or reproducibility of case mix methods.

Reliability is sometimes reported by simply comparing the fraction of cases in which two independent raters assigned the case to the same category (“the percent agreement”). This simple method turns out to be extremely misleading, however, especially when comparing a system having four possible categories with a system having over 400 possible categories. In assigning cases to one of a small number of categories with a high proportion of cases concentrated in one or two categories, multiple raters will frequently agree by chance alone. Since chance agreement is much less likely in a system with a large number of categories, some adjustment is needed to compare systems with different numbers of categories (for example, PMCs and SOII) (Horn and Williamson, 1977).

Meaningful assessment of reliability also requires taking account of the degree of disagreement. This in turn depends on the purposes of the assignment. For example, if DRGs are being used in reimbursement, then misassignment of a case to a wrong DRG with a cost weight similar to the right DRG is not as serious as misassignment to a wrong DRG with a dissimilar cost weight. On the other hand, if the DRG assignment is being used for clinical purposes, the distance from correct Major Diagnostic Category, rather than from the correct cost weight, might be the basis for measuring the egregiousness of a misassignment.
Because research on the different case mix adjustment systems has not been coordinated, reviewing evidence about the reliability of these systems will necessarily involve efforts to compare apples and oranges. The following review will suggest in broad terms what the reliability of the various case mix systems seems to be, but we will be limited in our cross system comparisons by what the researchers have chosen to report.

**DRGs**

The principal issue in the reliability of DRG assignment is the reliability of ICD-9-CM diagnoses and procedure coding. However, this is not a problem for DRGs alone, but for any system using ICD-9 codes—Disease Staging, PMCs, CSI, and (to the extent that they are applied within a diagnostic classification scheme) APACHE, MEDISGRPS, and SOII. The accuracy of diagnosis and procedure coding is discussed in detail in Worthman and Cretin (1986). We will briefly summarize the main findings here.

Agreement in coding principal diagnoses using the older ICD-8 classification scheme is generally reported to be between 60 and 70 percent, depending on whether agreement is at the three-digit or four-digit level (National Academy of Sciences, Institute of Medicine, 1977a,b, 1980; Demlo and Campbell, 1981). At the level of accuracy needed to assign a DRG, the agreement was just over 70 percent. The implications of such apparently high levels of disagreement on case mix index are of interest here. A recent study (Johnson and Appel, 1984) using ICD-9-based DRGs found disagreement in DRG assignment in 47 percent of the cases, but the average impact on the case mix index was 4 percent, with 77 percent of the hospitals showing no more than a 2 percent difference. This is similar to a finding that the average case mix index based on Commission on Professional and Hospital Activities (CPHA) versus HCFA data differs by about 3 percent (Carter and Ginsburg, 1985).

**APACHE**

The reliability of APACHE II has not been assessed. The early APACHE system depended on the abstraction of test results and other physiologic data, as well as assessment of a "Chronic Health Index." Repeat abstraction of the 34 physiologic measures resulted in agreement on 98 percent of the items (McClish et al., 1985). See Table 1. The agreement on the Chronic Health Index was reported to be "somewhat less" (Knaus et al., 1985). Agreement between APACHE scores based on two abstractions has also been reported: 100 percent agreement within two score points (Strauss et al., 1986).
Table 1

REPRODUCIBILITY

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>What Is Compared</th>
<th>Measure of Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>McClish et al., 1985</td>
<td>260 patient-days</td>
<td>1 hospital ICU 1983</td>
<td>Abstraction of physiologic items (APS)</td>
<td>Second abstraction of APS</td>
<td>% agreement item abstraction 98%</td>
</tr>
<tr>
<td>Strauss et al., 1986</td>
<td>30 cases</td>
<td>1 hospital ICU 1979–80</td>
<td>Abstraction of APS</td>
<td>Second abstraction of APS</td>
<td>Agreement on APS scores* 100% within 2 APS points</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease Staging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Louis et al., 1984</td>
<td>1380 ICD-9 cases</td>
<td>5 hospitals</td>
<td>Stage by manual abstraction</td>
<td>Stage by computer on UHDDS data</td>
<td>% agreement on stage (includes effect of cases with incorrect principal diagnosis) weighted 78%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

% agreement on stage (excludes effect of incorrect principal diagnosis) 80.5%
<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>What Is Compared</th>
<th>Measure of Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Richards et al., 1986a</td>
<td>4716</td>
<td>23 hospitals observations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stage at admission by manual abstraction</td>
<td>Adjusted $R^2$</td>
<td>.74</td>
</tr>
<tr>
<td>Stage at peak by manual abstraction</td>
<td>Stage at peak by manual abstraction</td>
<td>Adjusted $R^2$</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage at discharge by manual abstraction</td>
<td>Stage at discharge by manual abstraction</td>
<td>Adjusted $R^2$</td>
<td>.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDISGRPS</td>
<td>200 coders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brewster et al., 1985</td>
<td></td>
<td></td>
<td>Abstraction by MEDISGRPS expert Abstraction by hospital trainee</td>
<td>% trainees failing to meet 95% accuracy after 9 weeks experience</td>
<td>1%</td>
</tr>
<tr>
<td>Abstraction by hospital abstractor</td>
<td>Abstraction by second hospital abstractor</td>
<td>Contractual standard. Abstracts reviewed by MEDISGRPS expert</td>
<td>95%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1—continued

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>What Is Compared</th>
<th>Measure of Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young, 1985</td>
<td>28,790 cases</td>
<td>6 hospitals Western Pennsylvania FY 1981</td>
<td>PMCs based on UHDDS only</td>
<td>PMCs based on UHDDS plus reason for admission</td>
<td>% agreement case classification 89%</td>
</tr>
<tr>
<td>Young et al., 1982</td>
<td>4348 cases (5 modules)</td>
<td>6 hospitals Western Pennsylvania FY 1980</td>
<td>Criterion assignment by nurses</td>
<td>PMCs based on UHDDS only</td>
<td>% agreement case classification 90% after correcting ICD coding</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Source</th>
<th>What Is Compared</th>
<th>Measure of Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PMCs based on UHDDS plus reason for admission</td>
<td>% agreement case classification 95% after correcting ICD coding</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PMCs based on UHDDS only</td>
<td>% agreement case classification 72% before correcting ICD coding</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PMCs based on UHDDS plus reason for admission</td>
<td>% agreement case classification 75% before correcting ICD coding</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Sample</td>
<td>Source</td>
<td>What Is Compared</td>
<td>Measure of Comparison</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td>------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Horn and Horn, 1986</td>
<td>1775 cases</td>
<td>18 hospitals</td>
<td>SOII rating by expert rater</td>
<td>SOII rating by hospital coders (fourth check; &gt;2 months experience)</td>
</tr>
<tr>
<td></td>
<td>4489 cases</td>
<td>18 hospitals</td>
<td>SOII rating by expert rater</td>
<td>SOII rating by hospital coders (fourth and subsequent checks)</td>
</tr>
<tr>
<td>Richards et al., 1986b</td>
<td>3308 observations</td>
<td>5 hospitals</td>
<td>SOII rating by trained coder</td>
<td>SOII rating by trained coder</td>
</tr>
</tbody>
</table>

* Theoretical maximum is 92 points. Usual maximum is 50-55 points.
Disease Staging

Computerized Disease Staging uses the same UHDDS data as does the DRG system. Therefore the general conclusions about the inaccuracy of diagnosis and procedure coding raise the same concerns for the accuracy of Disease Staging as for DRGs. Unfortunately, we are lacking information about the overall effects of such diagnostic misassignment on a Staging case mix index.

Information on the reliability of computerized Staging has not been reported, per se. However, we do have data on the agreement between manual and computerized Staging in assigning stage (Table 1): between 78 and 80 percent, depending on how one handles cases with incorrect principal diagnosis (Louis et al., 1984). A study in progress (Richards et al., 1986a) indicates the reliability for admission, peak, or discharge stage using manual Staging is moderately good. In this study of 23 hospitals, less than 15 percent of the patients were at stage 3 or 4 at admission, and 70 percent of patients were at stage 0 by discharge. Raters agree on admission, peak, and discharge stage for 74, 74, and 77 percent of patients, respectively. After adjustment for differences between diagnosis and between hospitals, estimates of reliability based on the ratio of the patient sum of squares to the total sum of squares (adjusted for degrees of freedom) yield an adjusted $R^2$ of .74, .76, and .74, respectively.

MEDISGRPS

Information on the reliability of the MEDISGRPS abstraction process is sketchy. One report (Brewster et al., 1985) notes that only 1 percent of the trainees failed to meet 95 percent accuracy standards after nine weeks of experience (Table 1). Accuracy in this case was measured by comparing the items as abstracted by hospital trainees to those of an expert coder. There is no information about whether disagreement among raters leads to different MEDISGRPS severity assignments.

PMCs

The PMC system, like DRGs and Staging, uses UHDDS data in making its case mix assignment. An early version of PMCs used the Reason for Admission (chief problem or specific elective procedure) in addition to the UHDDS data set in classifying cases. These two versions of PMCs agree in 89 percent of the cases (Young, 1985) (Table 1). A smaller study tested manual assignment by nurses versus each
computerized version of PMCs to assess the validity of case classification in five PMC modules (Young et al., 1982). This study was not designed to examine reliability per se, but contains elements of reliability in that diagnostic coding in the UHDDS data influenced the level of agreement. In the first comparison, coding errors in the UHDDS data were not corrected. Manually assigned PMCs agreed with PMCs including Reason for Admission in 75 percent of cases, and with PMCs excluding Reason for Admission in 72 percent of cases. Following correction of obvious coding errors in the UHDDS data, classification improved to 95 percent with, and 90 percent without, Reason for Admission.

**Severity of Illness Index**

The manual Severity of Illness Index has reported more information relating to reliability than the other systems being reviewed. The reported levels of agreement between expert coders and hospital-based coders is high: between 88 and 93 percent (Table 1). These figures are impressive, but cannot be directly compared to the agreement in assigning three-digit ICD-9 codes, for example, without some adjustment for chance agreement. The SOII cases are strongly concentrated in severity levels 1 and 2, leaving only a small percentage of the cases at level 3 or 4. This distribution of severity level can lead to a high proportion of chance agreement (Richards et al., 1986b).

A detailed analysis of the reliability of SOII concluded that the overall Severity score is moderately reliable (Richards et al., 1986b). In a sample from five large teaching hospitals, raters agreed on the overall score for 68 percent of patients. After adjustment for differences between diagnoses and hospitals, the estimate of reliability based on the ratio of the patient sum of squares to the total sum of squares (adjusted for degrees of freedom) yields an adjusted $R^2$ of .64.

There is no information yet on the reliability of the Computerized Severity Index. Since it is based on adding a sixth digit to the ICD-9-CM codes, it presumably shares some of the same problems as DRGs, Staging, and other UHDDS systems.

**Summary**

The reliability of the various systems may be summarized as follows. We know almost nothing about the reliability of MEDISGRPS or the new Computerized Severity Index. APACHE seems to be quite reliable. SOII also seems to be moderately reliable, as is manual Staging (Richards et al., 1986a,b). Given their similar input data, it is likely
that DRGs, computerized Staging, and PMCs will have similar reliability, but this is only speculative, since no direct studies have been done to permit the conclusion.

PREDICTIVE PERFORMANCE

In reviewing the predictive performance of the case mix methods, it will be important to keep in mind that, as with reliability, studies on the different systems are usually on different samples, carried out at different levels and employing different statistics. Nonetheless, it is useful to obtain an overview of what has been done in testing each system.

DRGs

Although the original DRG system was developed on an idiosyncratic data set, before its use in PPS the system was used and tested on several statewide and national data sets, by many researchers. In fact, most reports on other case mix systems have used DRGs as a comparison system.

Performance at the Case Level. The principal validation of the DRG system has been its ability to predict length of stay, charges, and costs per case. When viewed at the case level, using reduction in variance as the statistical criterion, DRGs are clearly helpful in explaining variations in resource use. In a variety of studies (Table 2), DRGs explain from 17 to 30 percent of the variations in costs in untrimmed data and from 30 to 48 percent of the variations in cost on trimmed data (Mitchell et al., 1984, 1985). Performance in explaining charges is similar (West et al., 1985; Horn, 1985).

Surgical DRGs consistently perform better than medical DRGs (Table 3). Within surgical DRGs, at the case level, DRGs explain 48 to 57 percent of the variance in charges in trimmed data, whereas medical DRGs only explain 9 to 16 percent of the variance (Mitchell et al., 1984, 1985).

Performance at the Hospital Level. In reimbursement, the DRG-based case mix index is applied at the hospital level to explain differences in average cost per case. When used in a regression, together with variables similar to those used in the reimbursement formula (the wage index, the house staff ratio, bed size, and additional adjustments for the size of cities in urban areas), the total variance in average cost per case explained by the regression is over 72 percent of the variance in costs per case or charges per case (Table 4) (Pettengill and Vertrees, 1982; Cotterill, Bobula, and Connerton, 1985).
Table 2  
PERFORMANCE OF ICD-9-CM DRGs  
(Case level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>Trimmed/ Not</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>R^2</th>
<th>Average Coeff. V^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffey &amp; Goldfarb, 1984</td>
<td>394,000 cases^a</td>
<td>Maryland, 1979–81</td>
<td>Untrimmed</td>
<td>420 DRGs</td>
<td>LOS</td>
<td>.16</td>
<td>94%</td>
</tr>
<tr>
<td>Mitchell et al., 1984</td>
<td>272,087 cases^a</td>
<td>New Jersey, 1982</td>
<td>Untrimmed</td>
<td>DRGs</td>
<td>Costs</td>
<td>.18</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>237,639 cases^a</td>
<td>North Carolina, 1982</td>
<td>Untrimmed</td>
<td>DRGs</td>
<td>Costs</td>
<td>.16</td>
<td>.52</td>
</tr>
<tr>
<td>Mitchell et al., 1985</td>
<td>Not stated^a</td>
<td>Michigan, 1982</td>
<td>Untrimmed</td>
<td>DRGs</td>
<td>Costs</td>
<td>.17</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>Not stated^a</td>
<td>Washington, 1982</td>
<td>Untrimmed</td>
<td>DRGs</td>
<td>Costs</td>
<td>.30</td>
<td>.48</td>
</tr>
<tr>
<td>West et al., 1985</td>
<td>28,047 cases^a</td>
<td>South Carolina, 1981</td>
<td>Trimmed</td>
<td>DRGs</td>
<td>Charges</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>Horn, Horn, and Sharkey, 1984</td>
<td>19,122 cases^b</td>
<td>1 university teaching hospital</td>
<td>Not stated</td>
<td>436 DRGs</td>
<td>Charges</td>
<td>.33</td>
<td>85%</td>
</tr>
<tr>
<td>Horn, 1985</td>
<td>5580 cases^b</td>
<td>1 university teaching hospital</td>
<td>Not stated</td>
<td>401 DRGs</td>
<td>Charges</td>
<td>.29</td>
<td>76%</td>
</tr>
<tr>
<td>Horn et al., 1984</td>
<td>7500 cases^b</td>
<td>15 hospitals</td>
<td>Untrimmed</td>
<td>DRG cost weights</td>
<td>Cost/ Case</td>
<td>.28</td>
<td></td>
</tr>
</tbody>
</table>

^aMedicare cases.  
^bAll payors.
Table 3
MEDICAL AND SURGICAL DRGs

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample (Medicare)</th>
<th>Source</th>
<th>Trimmed/Not</th>
<th>Dependent Variable</th>
<th>Medical</th>
<th>Surgical</th>
<th>All</th>
<th>Measure of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitchell et al.,</td>
<td>272,087 cases</td>
<td>New Jersey, 1982</td>
<td>Trimmed</td>
<td>Charges</td>
<td>.11</td>
<td>.48</td>
<td>.32</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1984</td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td></td>
<td>.06</td>
<td>.31</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>North Carolina,</td>
<td>Trimmed</td>
<td>Charges</td>
<td>.09</td>
<td>.51</td>
<td>.32</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1982</td>
<td>Untrimmed</td>
<td></td>
<td>.04</td>
<td>.33</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>Mitchell et al.,</td>
<td>Not stated</td>
<td>Michigan, 1982</td>
<td>Trimmed</td>
<td>Charges</td>
<td>.10</td>
<td>.49</td>
<td>.30</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1986</td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td></td>
<td>.06</td>
<td>.34</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not stated</td>
<td>Washington, 1982</td>
<td>Trimmed</td>
<td>Charges</td>
<td>.16</td>
<td>.57</td>
<td>.48</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td></td>
<td>.09</td>
<td>.39</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>West et al.,</td>
<td>28,047 cases</td>
<td>South Carolina,</td>
<td>Trimmed</td>
<td>Charges</td>
<td>.07</td>
<td>.55</td>
<td>.26</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td>1979-1981</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coffey, 1985</td>
<td>268,358 cases</td>
<td>Maryland, 1979-1981</td>
<td>Trimmed</td>
<td>LOS</td>
<td>.12</td>
<td></td>
<td></td>
<td>Average coefficient of variation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td></td>
<td>.869</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>270,928 cases</td>
<td>Untrimmed</td>
<td>LOS</td>
<td></td>
<td>.09</td>
<td></td>
<td></td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.955</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank &amp; Leve, 1985</td>
<td>DRGs</td>
<td>HCFA</td>
<td>LOS</td>
<td></td>
<td>.964</td>
<td>.808</td>
<td></td>
<td>Average coefficient of variation SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.191)</td>
<td>(0.228)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HCFA</td>
<td>Costs</td>
<td></td>
<td>.963</td>
<td>.801</td>
<td></td>
<td>Average coefficient of variation SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.194)</td>
<td>(0.703)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maryland, 1979-1981</td>
<td>Untrimmed</td>
<td>Charges</td>
<td>.971</td>
<td>.857</td>
<td></td>
<td>Average coefficient of variation SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.666)</td>
<td>(0.292)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Sample</td>
<td>Source</td>
<td>Independent Variable</td>
<td>Coeff.</td>
<td>Standard Error</td>
<td>Dependent Variable</td>
<td>Result</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>--------</td>
<td>------------</td>
<td>--------------------------</td>
<td>--------</td>
<td>----------------</td>
<td>---------------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>Pettengill &amp; Vertrees, 1982</td>
<td>351 DRGs</td>
<td>MEDPAR, 1979</td>
<td>Case Mix Index (CMI)</td>
<td>1.081</td>
<td>.045</td>
<td>Medicare</td>
<td>Adjusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BLS wage index</td>
<td>1.000</td>
<td>.031</td>
<td>cost/case</td>
<td>R^2 = .72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident/bed ratio</td>
<td>.569</td>
<td>.042</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bed size</td>
<td>.107</td>
<td>.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.149</td>
<td>.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Medium city</td>
<td>.037</td>
<td>.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small city</td>
<td>.002</td>
<td>.012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotterill, Bobula, and Connerton, 1985</td>
<td>358 DRGs</td>
<td>MEDPAR, 1981</td>
<td>CMI (cost weights)</td>
<td>1.012</td>
<td>.043</td>
<td>Medicare</td>
<td>Adjusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1981 BLS wage index</td>
<td>1.023</td>
<td>.037</td>
<td>cost/case</td>
<td>R^2 = .72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident/bed ratio</td>
<td>.580</td>
<td>.045</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bed size</td>
<td>.119</td>
<td>.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.109</td>
<td>.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Medium city</td>
<td>.026</td>
<td>.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small City</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Constant</td>
<td>7.322</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CMI (charge weights)</td>
<td>.969</td>
<td>.041</td>
<td>Medicare</td>
<td>Adjusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1981 BLS wage index</td>
<td>1.022</td>
<td>.037</td>
<td>cost/case</td>
<td>R^2 = .72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident/bed ratio</td>
<td>.545</td>
<td>.045</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bed size</td>
<td>.117</td>
<td>.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.109</td>
<td>.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Medium city</td>
<td>.025</td>
<td>.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small city</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Constant</td>
<td>7.334</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horn et al. 1984</td>
<td>15 hospitals</td>
<td></td>
<td>Case Mix Index</td>
<td>—</td>
<td>—</td>
<td>Medicare</td>
<td>Adjusted</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cost/case</td>
<td>R^2 = .75</td>
<td></td>
</tr>
</tbody>
</table>
One small study of 15 hospitals reported that the case mix index alone explained 75 percent of the variance in cost per case (Horn et al., 1984). In view of the findings of the other larger studies, this must be viewed as an aberrant sample of hospitals.

Apache

APACHE II has been applied to intensive care cases at 13 hospitals, mostly university teaching hospitals (Knaus et al., 1985; Wagner and Draper, 1984). The original APACHE system has been applied more broadly (Knaus et al., 1981; Knaus et al., 1985), including a recently reported international study involving 14 hospitals from the United States, France, Spain, and Finland (Wagner et al., 1984). Analysis of all the APACHE data has stayed in the hands of the developers, although collection of the data has sometimes been carried out without their direct supervision (Knaus et al., 1985).

Performance at the Case Level. The principal predictive validity evidence on APACHE is related to the ability of the score to predict mortality (Table 5). As a predictor of death, the APACHE score has 97 percent sensitivity\(^1\) and a 49 percent specificity\(^2\) (Knaus et al., 1981). When components of APACHE II are used together with ICU admitting diagnosis and surgical status of the patient, about 32 percent of the variance in patient-level mortality status can be explained and about 85 percent of the cases correctly classified as to mortality outcome. This result is based on a study of 5030 ICU patients from 13 ICUs (Knaus et al., 1985).

APACHE II with surgical status is able to explain about 17 percent of the variance in resource use among 5790 ICU patients at 13 hospitals (Wagner and Draper, 1984). This study used the Therapeutic Intervention Scoring System (TISS) to generate a proxy for measuring resource use (Cullen et al., 1984). TISS is a measure for weighting ICU services according to nursing labor effort. An additional 5 percent of the variance in TISS score could be explained by adding variables identifying different hospitals. Only ICU cases were included and the patients were not limited to Medicare cases.

There are no studies examining the performance of APACHE in noncritical care cases or in a Medicare-only patient population.

Performance at the Hospital Level. The performance of APACHE in predicting hospital-level cost per case has not been tested.

\(^1\)Sensitivity is the proportion of correct predictions among all patients who lived.

\(^2\)Specificity is the proportion of correct predictions among all patients who did not live.
Table 5
PERFORMANCE OF APACHE II
(Case level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Samplea</th>
<th>Source</th>
<th>Trimmed/Not</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>R²</th>
<th>Percent Correct Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knaus et al., 1985</td>
<td>5030 cases</td>
<td>13 ICUs, 1979–1982</td>
<td></td>
<td>APS12</td>
<td>Mortality (n = 993)</td>
<td>.32</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td>(excludes coronary artery bypass)</td>
<td></td>
<td></td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chronic health</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Surgical status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ICU admit dx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wagner and Draper, 1984</td>
<td>5790 cases (includes coronary artery bypass)</td>
<td>13 ICUs, 1979-1982</td>
<td>“Trimmed&quot;b</td>
<td>APS12 “Cost” (TISS)b</td>
<td>“Cost” (TISS)b</td>
<td>.22</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chronic health</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Surgical status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ICU admit dx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hospital</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APS12</td>
<td></td>
<td></td>
<td></td>
<td>“Cost” (TISS)b</td>
<td></td>
<td>.17</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chronic health</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Surgical status</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a*All payors.

*b*TISS is the Therapeutic Intervention Scoring System, a measure of ICU services, weighted by nursing labor effort (Cullen et al., 1984). TISS data were trimmed by truncating TISS scores at 350 maximum, excluding 4 percent of patients and 16 percent of costs.
Until a data set including non-ICU cases is obtained, or until a proposal is made about how to limit the use of an APACHE-based modification of DRGs to just ICU patients, such a study is not likely to be done.

**Disease Staging**

Several large data sets have been used by the developers of Disease Staging, including the National Center for Health Services Research Health Care Utilization Project (HCUP) file for 1977 (with data from 373 short-term non-Federal hospitals) (Gonnella et al., 1984). In addition, other investigators have used Staging on large data sets: all Maryland Medicare discharges for 1979 through 1981 (Coffey and Goldfarb, 1984; Coffey, 1985), the HCUP file (Short and Coffey, 1984), and a sample of 500,000 discharges from 24 teaching hospitals (Butler and Bentley, 1982).

Medicare discharges from 54 Maryland hospitals for 1981 and from 80 New Jersey hospitals from 1979 have also been used in developing and testing a proposed modification of 17 DRGs using a Staging-based classification scheme (Conklin et al., 1984a,b; Conklin, 1985). This modification is discussed separately below.

**Disease Staging: Performance at the Case Level.** Some evidence of the predictive validity of Staging (within SDC) at the case level is available (Table 6). In both the Maryland Medicare data (Coffey and Goldfarb, 1984) and in the Michigan data (Calore, 1985b), Staging performed less well than DRGs in explaining costs per case or length of stay. Of course, these results do not mean that for some categories of disease, Staging might not outperform DRGs.

**Disease Staging: Performance at the Hospital Level.** All reported studies on Staging so far have reported results at the case level rather than the hospital level.

**DRGs Modified by Staging: Performance at the Case Level**

SysteMetrics has a proposal for modifying 17 existing DRGs using new patient groups based in part on computerized Staging (Conklin et al., 1984a,b; Conklin 1985). The 17 DRGs are organized into 10 groups of “adjacent DRGs,” DRGs which have the same defining variables except at the last “split,” such as with or without comorbid conditions. Special software has been developed to reclassify patients in the adjacent DRGs (ADRGs), using Disease Stage as one of the possible variables. At the case level, these new groups perform better as predictors
Table 6
PERFORMANCE OF DISEASE STAGING
(Case level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source*</th>
<th>Trimmed/Not</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>( R^2 )</th>
<th>Average Coeff. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffey &amp; Goldfarb, 1984</td>
<td>394,000 cases</td>
<td>Maryland, 1979-1981</td>
<td>Untrimmed</td>
<td>SDC-stages (principal diagnosis)</td>
<td>LOS</td>
<td>.12</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>805 groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SDC-stages</td>
<td>LOS</td>
<td>.10</td>
<td>93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>698 groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calore, 1985b</td>
<td>300,122 cases</td>
<td>Michigan, 1982</td>
<td>Untrimmed</td>
<td>SDC-stages</td>
<td>Costs</td>
<td>.10</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trimated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Medicare cases.

of length of stay and charges and costs than did the old DRGs (Table 7). The \( R^2 \) for the ADRGs is relatively low in all 10 diagnostic clusters. However, when compared with the variance explained by the DRGs in six of these clusters, the variance explained by the ADRG system is strikingly higher.

While these results suggest that the stage-based modifications to DRGs may be worthwhile, two steps are needed to verify that the stage-based groups result in more appropriate reimbursement. First, the hospital redistributive effects of the changes should be examined. Then, for a sample of cases where DRGs and the stage-based groups result in disparate estimates of reimbursement, a review of the actual cases is needed to determine which of the two estimates is the better reflection of needed care.

MEDIAGRPS

MEDIAGRPS was developed on a data set from one hospital (Brewster et al., 1985). Another report has included limited data from two additional hospitals (Brewster et al., 1984b).

Performance at the Case Level. The predictive validity of MEDIAGRPS has primarily been tested by the degree to which the system predicts mortality (Table 8), morbidity (Table 9) and charges in
Table 7
COMPARISON OF VARIANCE EXPLAINED (R²)
FOR DRGs AND 10 SYSTEMETRICS ADRGs

<table>
<thead>
<tr>
<th>Sample</th>
<th>Dependent Variable</th>
<th>Independent Variable</th>
<th>DRGs</th>
<th>Systemetrics</th>
<th>Systemetrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maryland</td>
<td>Medicare cost/case</td>
<td>Cirrhosis &amp; alcohol</td>
<td>—</td>
<td>.16 (1)</td>
<td>.18 (2)</td>
</tr>
<tr>
<td>1981</td>
<td></td>
<td>(DRG 202)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major rec. vasc.</td>
<td>.030</td>
<td>.115 (2)</td>
<td>.143 (3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 110 &amp; 111)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biliary tract</td>
<td>.07</td>
<td>.11 (3)</td>
<td>.19 (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 197 &amp; 198)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benign prostate Dia.</td>
<td>.02</td>
<td>.10 (4)</td>
<td>.14 (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 348 &amp; 349)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes</td>
<td>.0002</td>
<td>.08 (5)</td>
<td>.09 (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 294 &amp; 295)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kidney and UTI</td>
<td>.020</td>
<td>.073 (6)</td>
<td>.076 (8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 320,321,322)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pulmonary embolus</td>
<td>—</td>
<td>.061 (7)</td>
<td>.095 (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major chest/lung</td>
<td>—</td>
<td>.052 (8)</td>
<td>.075 (9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 75)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Major bowel proc.</td>
<td>.021</td>
<td>.050 (9)</td>
<td>.088 (7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 148 &amp; 149)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cereb. vasc. ex TIA</td>
<td>—</td>
<td>.014 (10)</td>
<td>.029 (10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(DRG 14)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SOURCES: Conklin et al., 1984a; Conklin, 1985.

*Severity Model includes condition, stage by condition, unrelated comorbidity and age.

selected conditions: abdominal pain, chest pain, shortness of breath, and extremity pain. (Table 10 shows the predictive power based on the first review and Table 11 shows the predictive power based on the second review.) Total charges for different severity groups are significantly different.

The concentration of the reported cases in one hospital and the selected nature of these conditions makes generalization difficult. In the hospitals and conditions studied, however, the five-level
MEDISGRPS score explained between 10 and 23 percent of the variance in charges per case based on the first review. Based on the second review (more similar to a review based on discharge data), the MEDISGRPS levels explained between 27 and 41 percent of the variance. Deaths and transfers are excluded in analysis of the charges. These numbers are difficult to interpret in the absence of a meaningful comparison group. A UHDDS system like DRGs, PMCs, or Staging may be able to explain a greater or a lesser fraction of the variance in charges in these cases. The performance of MEDISGRPS on an unselected sample of cases has yet to be reported.

**Performance at the Hospital Level.** Reports on hospital-level performance are not available.

**PATIENT MANAGEMENT CATEGORIES (PMCs)**

The developmental data set used for PMCs is far less "selected" than those used to develop the other systems. PMCs have been applied to all discharges from 90 Western Pennsylvania hospitals for 1981 through 1983 and all discharges from 55 Maryland hospitals in 1983. PMCs have now been used both on the developmental data set (Young, 1984, 1985) and on all Medicare cases from Michigan from

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>Dependent Variable</th>
<th>Mann-Whitney U Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brewster et al., 1985</td>
<td>1467 cases (Abdom. pain)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Mortality</td>
<td>n=43, Value = 13,327, p = &lt;0.01</td>
</tr>
<tr>
<td></td>
<td>1246 cases (Chest pain)</td>
<td></td>
<td></td>
<td>n=58, Value = 10,311, p = &lt;0.01</td>
</tr>
<tr>
<td>Brewster et al., 1984a</td>
<td>1313 cases (SOB)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Mortality</td>
<td>n=130, Value = 42,432, p = &lt;0.01</td>
</tr>
<tr>
<td>Brewster et al., n.d.</td>
<td>1007 cases (Extremity pain)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Mortality</td>
<td>n=24, p = &lt;0.01</td>
</tr>
</tbody>
</table>

*NOTE: Independent variable is MEDISGRPS category (5 groups).*  
*aAll payors, excluding transfers.*
Table 9

PERFORMANCE OF MEDISGRPS IN PREDICTING MORBIDITY
(Case level, first review)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>Dependent Variable</th>
<th>Mann-Whitney U Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brewer et al., 1985</td>
<td>1424 cases (Abdom. pain)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Morbidity, n=204</td>
<td>Value = 70,060, p = &lt;0.01</td>
</tr>
<tr>
<td></td>
<td>1188 cases (Chest pain)</td>
<td></td>
<td></td>
<td>Value = 40,294, p = &lt;0.01</td>
</tr>
<tr>
<td>Brewer et al., 1984a</td>
<td>1183 cases (SOB)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Morbidity, n=293</td>
<td>Value = 74,450, p = &lt;0.01</td>
</tr>
<tr>
<td>Brewer et al., n.d.</td>
<td>970 cases (Extremity pain)</td>
<td>Saint Vincent Hospital, FY1983</td>
<td>Morbidity, n=101</td>
<td>( \chi^2 ), p = &lt;0.01</td>
</tr>
</tbody>
</table>

NOTE: Independent variable is MEDISGRPS category (5 groups).

All payors, excluding deaths and transfers.

Performance at the Case Level. The predictive validity of PMCs is established both by using mortality within PMCs and by predicting length of stay (Table 12). The developers report statistics on the stability of death rates in selected PMCs over time (Young, 1985) as partial evidence of the validity of the system. The primary performance analysis, however, looked at the distribution around length of stay within PMC categories. The average coefficient of variation was only .59 in the trimmed data set and .79 in the untrimmed data. This is lower than the average coefficient of variation on LOS using DRGs in untrimmed data sets (.94 reported by Coffey and Goldfarb, 1984) or using Staging (.88 to .93 reported by Coffey and Goldfarb, 1984).

Despite this promising performance of PMCs in explaining length of stay, the performance of the system in explaining costs in the 1982 Michigan data was not better than that of DRGs and somewhat better than Staging in the same data set—explaining about 15 percent of the variance in cost per case in the untrimmed data set and 26 percent in the trimmed data set (Calore, 1985b) (Table 13).

Performance at the Hospital Level. There are no studies comparing performance of the system at the hospital level.
Table 10

PERFORMANCE OF MEDISGRPS IN PREDICTING CHARGES
(Case level, first review)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>R²</th>
<th>Coeff. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brewster et al., 1985</td>
<td>1424 cases (Abdom. pain)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.23 1.46 total</td>
<td>.11 weighted ave. for subgroups</td>
</tr>
<tr>
<td></td>
<td>1180 cases (Chest pain)</td>
<td></td>
<td>.14 1.11 total</td>
<td>.80 weighted ave. for subgroups</td>
</tr>
<tr>
<td>Brewster et al., 1984a</td>
<td>1183 cases (SOB)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.10 1.23 total</td>
<td>.13 weighted ave. for subgroups</td>
</tr>
<tr>
<td>Brewster et al., n.d.</td>
<td>983 cases (Extremity pain)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.18 1.17 total</td>
<td>.105 weighted ave. for subgroups</td>
</tr>
</tbody>
</table>

NOTE: Independent variable is MEDISGRPS category (5 groups); dependent variable is charges. All payors, excluding deaths and transfers.

Table 11

PERFORMANCE OF MEDISGRPS IN PREDICTING CHARGES
(Case level, second review)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>R²</th>
<th>Coeff. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brewster et al., 1985</td>
<td>1424 cases (Abdom. pain)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.41 1.46 total</td>
<td>.85 weighted ave. for subgroups</td>
</tr>
<tr>
<td></td>
<td>No R² morb. = 1220 R² morb. = 204</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brewster et al., 1984a</td>
<td>1188 cases (Chest pain)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.28 1.11 total</td>
<td>.67 weighted ave. for subgroups</td>
</tr>
<tr>
<td></td>
<td>No R² morb. = 1032 R² morb. = 156</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1183 cases (SOB)</td>
<td></td>
<td>.27 1.23 total</td>
<td>.81 weighted ave. for subgroups</td>
</tr>
<tr>
<td></td>
<td>No R² morb. = 890 R² morb. = 293</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brewster et al., n.d.</td>
<td>983 cases (Extremity pain)</td>
<td>Saint Vincent Hospital, Prize FY1983</td>
<td>.33 1.17 total</td>
<td>.97 weighted ave. for subgroups</td>
</tr>
<tr>
<td></td>
<td>No R² morb. = 881 R² morb. = 102</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Independent variable is MEDISGRPS category (5 groups);
Table 12
PERFORMANCE OF PATIENT MANAGEMENT CATEGORIES
(Case level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>Medicare/ All</th>
<th>Trimmed/ Not</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>R²</th>
<th>Average Coeff. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calore, 1985b</td>
<td>300,122 cases</td>
<td>Michigan, 1982</td>
<td>Medicare Untrimmed PMCs</td>
<td>Costs</td>
<td>.15</td>
<td>.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young, 1985</td>
<td>499,721 cases</td>
<td>Western Pennsylvania, 1983</td>
<td>All payors Untrimmed Single assignment LOS</td>
<td>.79 unweighted .76 weighted*</td>
<td>.59 unweighted .56 weighted*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aWeighted by number of patients.
Table 13

COMPARISON OF THE VARIATION ($R^2$) IN COST PER CASE AT CASE LEVEL EXPLAINED BY THREE CASE MIX MEASURES

<table>
<thead>
<tr>
<th></th>
<th>DRGs</th>
<th>PMCs</th>
<th>Staging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trimmed</td>
<td>0.30</td>
<td>0.25</td>
<td>0.17</td>
</tr>
<tr>
<td>Untrimmed</td>
<td>0.17</td>
<td>0.15</td>
<td>0.10</td>
</tr>
</tbody>
</table>

SOURCE: Calore, 1985b.
NOTE: Cost data are untrimmed and exclude DRGs 468 and 470.

Severity of Illness Index (SOII)

SOII research has been reported on a developmental data set consisting of 70 percent of the discharges from 15 “primarily large teaching hospitals” (Horn et al., 1984, 1985b). The training of abstractors in all these hospitals was done under the supervision of the developer. The actual data were collected by employees at the site (except for reliability checks).

Performance at the Case Level. Severity/Procedure adjusted DRGs appear to explain a significant part of the variance in cost per case (Table 14). In the same cases, DRGs alone were reported to explain only 28 percent of the variability in resource use. However, the reported statistics on SOII have been criticized as tending to overstate the performance of the system by failing to properly adjust for cell size and degrees of freedom. In addition, when coding some of the component dimensions of the SOII (non-operating room procedures or rate of response to therapy), coders may be heavily influenced by actual resource use, rather than the need for resources. This would produce a measure which explains a large amount of variance in length of stay and charges, although it would not be useful in reimbursement.

No data are available on the Computerized Severity Index and its performance, either at the case or hospital level.

Performance at the Hospital Level. In one study of SOII, charges were converted to operating cost per case and a hospital-level analysis comparing DRGs to Severity/procedure-adjusted DRGs was carried out. In the 15 hospitals in this sample, DRGs alone explained 75 percent of the hospital-level cost per case and Severity/procedure-adjusted DRGs explained 81 percent of the variance (Table 15). These numbers are both high compared with other hospital-level analyses of
**Table 14**

PERFORMANCE OF SEVERITY OF ILLNESS INDEX
(Case level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source*</th>
<th>Trimmed/ Not</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>Average R²</th>
<th>Coeff. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horn, Horn, and Sharkey, 1984</td>
<td>19,122 cases</td>
<td>1 university teaching hospital</td>
<td>Not stated</td>
<td>SOII and procedure type (12 groups)</td>
<td>Charges</td>
<td>.49</td>
<td>70%</td>
</tr>
<tr>
<td>Horn, 1985</td>
<td>5580 Cases</td>
<td>1 university teaching hospital</td>
<td>Not stated</td>
<td>SOII and procedure type</td>
<td>Charges</td>
<td>.46</td>
<td>64%</td>
</tr>
<tr>
<td>Horn et al., 1984</td>
<td>7500 cases</td>
<td>15 hospitals</td>
<td>Untrimmed</td>
<td>Severity/proc. adjusted DRG cost weights</td>
<td>Cost/ case</td>
<td>.61</td>
<td></td>
</tr>
</tbody>
</table>

*All payors.*
Table 15

COMPARISON OF SEVERITY/PROCEDURE-ADJUSTED DRGs WITH DRGs
(Hospital level)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Source</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horn et al., 15 hospitals 1984</td>
<td>Severity/procedure adjusted DRG Case Mix Index</td>
<td>Hospital cost/case</td>
<td>$R^2 = .81$</td>
<td></td>
</tr>
</tbody>
</table>

DRGs. Given the small sample size, the most likely conclusion is that the hospitals in the sample are not representative. Either the sample of hospitals is highly selected or the methods are flawed.

COMPARATIVE PERFORMANCE

As noted earlier, only two of the case mix adjustment system (PMCs and computerized Disease Staging) have matured sufficiently to be tested on comparative data by investigators other than the original system developers.

Two findings emerge from these studies. First, even though these two case mix methods are reasonably complete and documentation is available, implementation of the systems by a new user can be difficult: more difficult than implementing DRGs. This is not surprising, since the systems have not been around as long and have not had the opportunity for feedback from outside users that the DRG system has had. Partly because they are still somewhat experimental, the PMC and Staging software leave open questions about exactly how the algorithm will be applied in certain complex cases. In actual implementation by an outside user, both systems were unable to classify a significant number of cases. For example, Calore (1985a) found that 20 percent of Michigan cases could not be staged in the original application of the Staging algorithm. Subsequent classification of the same data by SystemsMetrics, using a new version of the algorithm, staged 89 percent of cases. An additional 10.5 percent were classified into "other"
“catchall” categories. Similarly, Calore (1985b) found that 10.5 percent of Michigan cases were not classified into PMCs, while 12.5 percent were assigned to two PMCs (compared to 5 percent unclassified and 21.7 percent two-PMC assignments in the developmental data [Young, 1985]).

The second finding is that viewed as self-contained systems neither Staging nor PMCs put on a convincingly better performance than DRGs (Table 16). This is particularly true for Staging, which actually explained about two-thirds of the variation in costs explained by DRGs. PMCs fared somewhat better. It is important to note, however, that both PMCs and Staging benefit from adding information about whether a case was medical or surgical. This information is already contained in DRGs, but not in Staging or PMCs. If one looks only within medical cases or only within surgical cases, then both PMCs and Staging perform as well as DRGs.

On performance alone, none of the alternative case mix adjustment systems stand out as clearly superior to DRGs. In the next section, we summarize our conclusions regarding the systems for use in improving the measurement of case mix.
Table 16

MEDICAL AND SURGICAL PMCs, DISEASE STAGING, AND DRGs

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Source</th>
<th>Trimmed/Not Variable</th>
<th>Medical</th>
<th>Surgical</th>
<th>All</th>
<th>Measure Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Patient Management Categories</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calore, 1985b</td>
<td>300,122 cases</td>
<td>Michigan, 1982</td>
<td>Trimmed Charges</td>
<td>.11</td>
<td>.51</td>
<td>.26</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td>.07</td>
<td>.35</td>
<td>.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Disease Staging</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calore, 1985b</td>
<td>300,122 cases</td>
<td>Michigan, 1982</td>
<td>Trimmed Charges</td>
<td>.13</td>
<td>.49</td>
<td>.17</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td>.09</td>
<td>.35</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Coffey, 1985</td>
<td>268,358 cases</td>
<td>Maryland, 1979–1981</td>
<td>Trimmed LOS</td>
<td>.09</td>
<td>–</td>
<td>–</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.847</td>
<td>–</td>
<td>–</td>
<td>Average coefficient of variation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>270,928 cases</td>
<td></td>
<td>Untrimmed</td>
<td>LOS</td>
<td>.07</td>
<td>–</td>
<td>–</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.916</td>
<td>–</td>
<td>–</td>
<td>Average coefficient of variation</td>
</tr>
<tr>
<td><strong>DRGs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calore, 1985b</td>
<td>300,122 cases</td>
<td>Michigan, 1982</td>
<td>Trimmed Charges</td>
<td>.10</td>
<td>.49</td>
<td>.30</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Untrimmed</td>
<td>.06</td>
<td>.35</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td>Coffey, 1985</td>
<td>268,358 cases</td>
<td>Maryland, 1979–1981</td>
<td>Trimmed LOS</td>
<td>.12</td>
<td>–</td>
<td>–</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.859</td>
<td>–</td>
<td>–</td>
<td>Average coefficient of variation</td>
</tr>
<tr>
<td>270,928 cases</td>
<td></td>
<td>Untrimmed</td>
<td>LOS</td>
<td>.09</td>
<td>–</td>
<td>–</td>
<td>$R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.955</td>
<td>–</td>
<td>–</td>
<td>Average coefficient of variation</td>
</tr>
</tbody>
</table>

*Medicare cases.*
IV. CONCLUSIONS

We present our conclusions in three parts: the ability of the systems to substitute for DRGs, their ability to refine DRGs, and further work which might result in improved case measurement for PPS.

PERFORMANCE OF THE ALTERNATIVES AS DRG SUBSTITUTES

Looking at the content of the DRG classification system has raised concerns about the use of procedures to define some DRGs. Likewise, there is concern about the concatenation of unrelated diagnostoses within certain DRGs. By and large, however, the criticisms levied against DRGs have not been that what they measure is not case mix, but that they fail to measure all of the case mix. The acceptance of DRGs as a valid “first cut” at case mix is underscored by the number of new systems proposing to build on the DRG system: SOII, MEDISGRPS, Staging, and perhaps even APACHE.

The sense that DRGs are a reasonable case mix adjustment method is strengthened when one considers substituting one of the alternative methods as the principal case mix adjuster in PPS. This is not to say that the DRG system cannot be improved. However, it is far from clear that any of the alternative systems offer a better base on which to build improvements.

Some of the systems can be ruled out on structural grounds as plausible substitutes for DRGs. SOII, MEDISGRPS, and APACHE all require a diagnostic framework within which the severity level or score can be applied. These systems, then, could not by themselves replace DRGs. Neither has the Computerized Severity Index been developed to the point where it can be considered a viable alternative to DRGs.

Computerized Staging and PMCs are plausible substitutes for DRGs. Both are reasonably complete, self-contained case mix adjustment systems and both systems have been developed and documented to the point that users other than the initial developers can implement the systems. However, there is little evidence that by themselves these systems outperform DRGs. Staging was compared to DRGs in the Coffey and Goldfarb (1984) study. Both Staging and PMCs have been compared to DRGs by Calore (1985b). In neither of these head-to-head comparisons did the alternative systems demonstrate their clear superiority to DRGs.
It is possible that even though a system is no better than DRGs in explaining overall costs, it might be preferred to DRGs on other grounds. One possibility is that the system might better reflect the case mix of tertiary care or teaching hospitals. A second possibility is that the system might be preferred on structural grounds, that is, that its classification structure is better than that of the DRG system. These possibilities have been suggested for Staging and PMCs.

Staging was seen as a system that would be better than DRGs for teaching and tertiary care hospitals (Butler and Bentley, 1982). However, when Coffey and Goldfarb (1984) and Short and Coffey (1984) further analyzed the performances of Staging and DRGs, the differences in performance were not in the hoped for directions. Staging did not appear to do a better job in differentiating cases seen at teaching hospitals, for example. Hospitals with a difficult surgical load also appear to lose under Staging, compared to DRGs. Pending further work that shows new advantages of Staging, there does not seem to be a compelling reason to replace DRGs with Staging on a wholesale basis.

PMCs are thought to have an advantage over DRGs for reimbursement, because the concept behind PMCs—grouping together patients on the basis of their needed components of care—is a better reflection of the goals of prospective payment than the procedure used in DRGs—grouping patients based on actual resources used. In practice (as opposed to concept), over a third of the PMC groups use procedures as part of the basis for grouping. So far, there has been no opportunity to simulate what redistributive and practical effects might follow from substituting PMCs for DRGs. Again, there is no compelling reason to replace DRGs with PMCs at this time.

POTENTIAL OF THE ALTERNATIVE SYSTEMS FOR REFINING DRGs

Four of the systems might be used as modifiers to DRGs: Disease Staging, SOII, APACHE, and MEDISGRPS.

Disease Staging

SysteMetrics has developed alternative groupings which incorporate information about Staging, along with other UHDDS variables used in DRGs to modify 17 existing DRGs (Conklin et al., 1984a,b; Conklin, 1985). Within these “adjacent DRGs,” the alternative groupings did outperform the DRG groups, sometimes markedly.
These results suggest the possible value of looking for better ways to classify cases in high-variance DRGs. Because the proposed alternative groups were based on more than simply stage of disease, and since stage was sometimes collapsed to two or three levels in these groupings, it is hard to separate the value of Staging, per se, in the improvement seen in the adjacent DRGs. It is possible that new groups based on UHDDS variables alone would capture much of the improvement demonstrated by the SysteMetrics groups.

Severity of Illness Index

The developer of the SOII has been a strong advocate of the use of SOII as modification to DRGs. At present, this would seem to be ill advised. Even the fairly objective DRG system suffered from an unexpected amount of “index creep.” The subjective SOII seems likely to encourage even more “upward coding.” The system has only been tested in 15 primarily large teaching hospitals (although more use the SOII system). From this biased sample it is impossible to conclude what the redistributive effects of the system would be to other hospitals. The importance of the sample has been demonstrated in tests of Disease Staging. When Staging was tested only in teaching hospitals, it appeared to capture the range and severity of cases (Butler and Bentley, 1982), yet when tested on statewide data, it did not appear to be an improvement over DRGs (Coffey and Goldfarb, 1984). Finally, the problems of implementing a manually coded, subjectively assessed measure nationally seem to overshadow the (remotely) possible gains.

Recognition of these problems has led to the development of a new version of the Severity of Illness Index called the Computerized Severity Index, a completely different system from the old SOII. This system has yet to be fully described and no data on the performance of the system have been reported. One thing is clear. The implementation of CSI first requires that the abstractors be trained in a new sixth-digit ICD-9 diagnosis and procedure coding system. It is premature to speculate on the suitability of this system as a replacement to DRGs.

APACHE and MEDISGRPS

These two systems offer promising future approaches to modifying DRGs, but in the short run both have the problem of acquiring several new data elements not now routinely coded on discharge. MEDISGRPS especially is in such an early stage of development that the consequences and possible advantages of using the system cannot be assessed. The APACHE system, which was originally intended for
intensive care patients, could be considered as an add-on for critically ill patients only, but more work is needed to demonstrate the feasibility of identifying such patients, of grafting APACHE on to DRGs, and of the redistributive effects such a grafted system would have at the hospital level.

Conclusions on Refinements

Two types of refinements to DRGs can be considered: those which, like DRGs, only require UHDDS elements to operate, and those which would require the expansion of UHDDS to include reason for admission, physiologic test results, or other elements. In the short run, UHDDS-based modifications are easiest to implement quickly.

The Staging-based modifications to 17 DRGs are the most highly developed of the proposed modifications. This system is UHDDS based, but it would require that all discharges in the 17 affected DRGs be run through Disease Staging, as well as DRG, software. Even for this system, further work has to be done to verify that the modifications would indeed solve perceived problems in the system and not merely compound them.

Of the remaining groups (APACHE, MEDISGRPS, SOII, and PMCs), APACHE seems closest to being implementable. MEDISGRPS (with its extensive data requirements) and the new CSI (with its requirement for implementing a new system of diagnosis and procedure coding) both seem to be in too early a stage of development for evaluation. The SOII system, with its labor intensive, subjective severity measures, is unsuitable for a reimbursement system.

WORK IN PROGRESS AND FUTURE RESEARCH NEEDS

As other reports suggest, the most important problem in DRG refinement is that we do not know where the current DRGs are weak. If problems are concentrated in outliers, then systems with fairly broad categories and emphasizing degree of disease progression are likely to be ineffective improvements. Results to date with Staging suggest that this is the case. If problems are focused on the purpose for which the patient is treated rather than the patient's condition, then new approaches to classification are needed and detailed attention to certain kinds of procedures are in order. If the difficulties result from differences in the degree of diagnostic and therapeutic aggressiveness between hospitals and between physicians, then Medicare will need either to pay for these differences without determining their efficacy or to find ways to pay for outcomes rather than treatments.
Work in Progress

A number of studies now in progress will yield significant clues as to the relative performance of a number of currently available measures. The Graduate Medical Education study will not only compare the performance of SOII and manual Staging (both now somewhat dated systems), but will look at the interplay between case mix, quality, cost, and appropriateness of care in a selected group of hospitals. Ashcroft and Thompson at the University of Michigan are pursuing case mix evaluation in a consortium of teaching hospitals. Alemi at Tulane is looking at the question of severity in myocardial infarction cases. HCFA-sponsored research at the Brandeis and RAND/UCLA Centers is continuing to review the adequacy of the current case mix adjustment, looking for evidence of skimming and dumping, regional practice variations, and appropriateness of care.

These studies should provide significant guidance as to how much information physiologic variables can add to case classification and supply guidance for possible large-scale studies involving primary data collection.

DRG Refinement Using Staging and PMCs

Conklin (1985) has already demonstrated that Staging and DRGs can be synthesized to produce a system which is somewhat more powerful than current DRGs but does not require additional data collection. Further work in this area is in progress at SysteMetrics and should be complete by the end of 1986. Young has viewed PMCs as a system so different from DRGs that synthesis is not possible, but PMCs may nevertheless provide useful tools in refining parts of the DRG system. Calore and colleagues at the Brandeis Center are exploring this approach. Although the improvements promised by these strategies appear to be modest, such improvements also appear to be appropriate. Such steps appear to be more useful than comparisons of the existing DRG, Staging, and PMC systems on larger data sets since PMCs and Staging do seem to provide enough advantages to consider using them as replacements for DRGs.

DRG Refinement Using Physiologic Measures

The studies in progress briefly reviewed above will provide considerable evidence on use of physiologic measures, but none will provide definitive information. This is due, in part, to the fact that the systems are evolving sufficiently rapidly so that new versions will almost certainly be in place by the time results from the current studies are
analyzed. A second problem is that each of these systems is developed and tested on a limited data set, so that cross-comparisons of systems on the same cases are difficult. The need for a large-scale comparison study is clear. The key questions to be raised before engaging in the design and implementation of such a study are:

- What do we gain by adding clinical data? For which kinds of patients?
- What clinical data are most useful for classification?
- Is it best to compare existing “off the shelf” systems or try to design a study which would measure the value of adding new data elements with demonstrated utility in one or more of these systems?

The answers may be more important than the particular construction of a score or of an assignment algorithm. Given the problems inherent in many of the existing systems and the proprietary nature of several of them, one might ask what the relative costs would be of developing “from scratch” a new case mix classification scheme based on both physiologic and UHDDS data elements.

Concurrent Issues Requiring Study

Work to measure quality of care or outcomes is an integral part of measuring case mix because it will not be possible to determine whether certain kinds of cases and certain kinds of treatment should be classified together and paid similarly until these issues can be settled.

SUMMARY

The most striking conclusion from this review is that no alternative stands out as clearly superior to DRGs at this time. The second is that modest improvements are now possible and that further improvements are very likely to be possible. But as important as any of these individual conclusions is the evidence that until the problems with the DRG system are more sharply defined it will be impossible to choose the most efficient strategy for DRG refinement.
REFERENCES


Demlo, L. K., and P. M. Campbell, "Improving Hospital Discharge Data: Lessons From the National Hospital Discharge Study," Medical Care, 19(1):1030–1040, October 1981.

Federal Register, "Medicare Program; Prospective Payments for Medicare Inpatient Hospital Services; Interim Final Rule with Comment Period," Federal Register, September 1, 1983.

Federal Register, "Medicare Program; Prospective Payments for Medicare Inpatient Hospital Services; Final Rule," Federal Register, January 3, 1984.

Federal Register, "Medicare Program; Changes to the Inpatient Hospital Prospective Payment System Fiscal Year 1986 Rates; Final Rule," Federal Register, September 3, 1985.


Richards, T., W. Rogers, D. Reboussin, G. Monsivail, N. Lurie, and R. Brook, "Reliability Analysis of the Severity of Illness Index: Preliminary Results Based on Data from Five Hospitals," The RAND Corporation, unpublished draft, January 1986b.


