REVIEW OF THE LITERATURE ON DIAGNOSIS RELATED GROUPS

Linda G. Worthman, Shan Cretin

October 1986
The research described in this report was supported by the Health Care Financing Administration, U.S. Department of Health and Human Services, under Cooperative Agreement No. 18-C-98489/9-01.

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.
A RAND NOTE

N-2492-HCFA

REVIEW OF THE LITERATURE ON DIAGNOSIS RELATED GROUPS

Linda G. Worthman, Shan Cretin

October 1986

Prepared for

The Health Care Financing Administration, U.S. Department of Health and Human Services
This literature review 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 reviews the literature on Diagnosis Related Groups, the case mix adjustment method used in the Prospective Payment System.

This work was conducted at the RAND/UCLA Center for Health Care Financing Policy Research, which is supported through Cooperative Research Agreement 18-C-98489/9-01. See also Shan Cretin and Linda G. Worthman, *Alternative Systems for Case Mix Classification in Health Care Financing*, R-3457-HCFA.
SUMMARY

In 1983, Medicare's Prospective Payment System (PPS) was enacted and implemented. In this single stroke, the system by which hospitals are reimbursed for Medicare inpatients was completely overhauled. PPS pays hospitals a fixed amount for each Medicare discharge. The payment is based primarily on the Diagnosis Related Group (DRG) into which the case falls, with adjustments for other factors affecting hospital costs (level of teaching activity, local wages, and location in an urban or rural environment).

In designing PPS, Congress intended to distribute hospital payments based on the resource needs of Medicare patients as defined by the mix of DRGs seen at each hospital. If factors not recognized in DRG classification are related to the expected costs of medically necessary treatment, hospitals may be over- or underpaid by PPS. Consistent over- or underpayment may jeopardize certain hospitals or patients. Concerned that the DRGs adjust appropriately for case mix, Congress requested the Health Care Financing Administration to prepare a report on Refining Case Mix Adjustment in Medicare's Prospective Payment System. This literature review provides background information on the DRGs in support of HCFA's report to Congress.

The review encompasses published and unpublished literature, primarily from the years 1982 through 1985. It focuses on the version of DRGs used in PPS and also reviews other PPS factors that influence DRG performance.

The review includes studies and commentary on the following areas:

- ability of DRGs to predict resource use
- three possible sources of variation in the DRGs
  - unmeasured "severity"
  - the quality of data DRGs use in classifying cases
  - variation due to regional and local variation in patterns of patient care
possible problems with the DRG relative weights
- critical commentary on the movement to national rates
- hospital and clinical strategies for adjusting to PPS
- effect of PPS on tertiary care hospitals
- problems in accommodating new technology.

Few studies of DRGs' predictive performance have appeared as journal articles, although several technical reports on this subject have appeared recently. This suggests that even basic information about the system is not readily available. We organized these predictive studies into three groups:

- performance of the case mix index at the hospital level
- performance at the case level
- performance of surgical and nonsurgical DRGs.

Studies show that the DRG case mix index—along with adjustments for wages, teaching activity, and urban/rural location—explain 72 to 75 percent of the variation in cost per case at the hospital level. At the case level, DRGs typically explain between 26 and 33 percent of the variation in costs or charges when outlier cases are removed but have explained as much as 48 percent of the cost variance in one statewide dataset. The power of the DRGs derives primarily from the surgical DRGs, which explain between 48 and 57 percent of the variation, rather than the 7 to 16 percent explained by the nonsurgical DRGs. High levels of unexplained variation among all cases or among nonsurgical cases is a problem for PPS if hospitals consistently treat patients who are more or less expensive than average, either by design or by chance.

The reasons for unexplained cost variation after adjusting for DRG case mix are not yet understood. The literature attributes unexplained variation in the DRGs to three general causes: unmeasured "severity," poor quality data, and variations in practice patterns.

Studies on "severity" vary greatly in quality and comprehensiveness. Several methods have been developed that claim to measure "severity" differences not captured by DRGs. However, direct
comparisons demonstrate that the case mix adjustment in PPS would not be improved by replacing DRGs with one of the existing alternatives. For example, comparisons on the same datasets show that Disease Staging and Patient Management Categories do not explain costs and length of stay better than the DRGs.

Many of the more narrowly focused studies of "severity" are difficult to interpret. Some studies that compare reimbursement with cost fail to adjust for important factors in the PPS formula, such as the indirect medical education adjustment. Others fail to consider that PPS payments are expected to "average out" to an equitable compensation and examine selected cases within a DRG or patients who use specific services such as intensive care. However, consistent underpayment for a DRG or for an identifiable subgroup of patients may lead to access problems for some patients.

Two concerns are currently raised about the quality of the data used in assigning the DRG. The first question is whether inherent limitations in the current diagnosis and procedure coding system account for much of the variance in costs not explained by the DRGs. The second question is the degree to which hospitals manipulate the data to maximize reimbursement. In the early days of PPS, the quality of the 1981 Medicare dataset used to set the DRG relative weights was also a concern. Reweighting based on more recent data has rendered this concern moot.

Although there is an extensive published literature on the effects of data quality on DRG classification, here as elsewhere there are no definitive answers. According to the literature, an earlier diagnosis coding scheme (ICD-8) was sufficiently unreliable to account for some of the unexplained variation in costs. The new ICD-9-CM diagnosis coding system may perpetuate unreliability or structural problems in the classification of diagnoses, but a study assessing the reliability of coding in ICD-9 has not been reported. Studies of diagnosis coding reliability indicate sufficient ambiguity, especially in complex cases, that hospitals could easily manipulate reimbursement by selecting codes judiciously. However, evidence suggests that this is not a major problem. Between 1981 and 1984, the Medicare case mix index increased 8.4 percent; however, investigators found that less than 3 percent of
the increase could be attributed to coding practice changes induced by PPS.

A third reason for unexplained variation in costs is variation in medical practice patterns. Inappropriate variation in practice patterns is not easily distinguished from all other sources of variation in costs within a DRG. Building on a history of well-documented variations both across regions and within small areas, studies have now been conducted using DRGs as the unit of analysis to show that variations exist in admitting patterns, length of stay, and use of other resources within DRGs. The reasons for these differences have not been determined, nor has the possible effect on medical outcomes been studied.

To gain insight into how the DRGs and PPS work or fail to work in practice, we reviewed the strategies hospitals are using to cope with the new system. The hospital management literature reveals that hospitals are using formal and informal "severity" adjustments to explain variation within DRGs. The adjustments help the hospitals to identify ways to contain costs. Hospitals are also attempting to influence physician practice patterns by developing standards of care, increasing concurrent and retrospective review activities, and instituting economic grand rounds. This literature is not rigorous and proves little. However, it does flag areas where there may be perceived problems with DRGs and where the responses to these problems may further generate undesirable effects.

In most of the published studies on "severity," one cannot easily distinguish possible problems in the DRG classification system from possible problems in the relative weights assigned to individual DRGs. The DRG weights have been criticized as being "compressed"—that is, high weights are not high enough and low weights are too high. If the weights are compressed, hospitals treating higher proportions of costly, complicated cases will be underpaid, while hospitals treating higher proportions of less costly, uncomplicated cases will be overpaid. Studies suggest that the original weights based on costs were compressed, and that basing the weights on charges reduces compression. HCFA recently recalibrated the weights using charges as a result of this evidence.
Criticism is also directed against another feature of PPS, the plan to reimburse hospitals based on a single national rate. National rates will not reduce Medicare costs, but some observers fear that they will create profits and losses simply based on the hospital's regional location. There have been no studies on this issue.

For teaching and public hospitals, discussion centers on two issues, neither of which are resolved in the current literature. The first issue is whether unmeasured variation in patients' treatment needs systematically jeopardizes this group of hospitals. The second issue concerns other factors (in addition to case mix) that may lead to higher costs in these institutions. Several studies using DRGs to control for case mix bear on these issues. First, within a single hospital, differences in case mix do not account for all of the difference in the intensity of treatment rendered to teaching versus nonteaching patients. In the same study, greater treatment intensity was associated with better short term, but not with better long term, outcomes for teaching patients. Second, patients with complex illness (as measured by complications and comorbidities, or by Disease Staging) occur in about the same proportion in teaching and nonteaching hospitals. Third, teaching hospitals perform more procedures than nonteaching hospitals, given the same DRG case mix of patients. These findings do not resolve the key question, however: Do teaching hospitals do more for the same patients, or are teaching hospitals seeing patients who differ in ways not captured by DRGs?

In calculating the true costs of teaching programs, it is necessary to adjust not only for case mix, but also for the costs of physician services, which are billed separately in private hospitals but included in the hospital bill in many teaching hospitals. Adjusting for both factors still produces a cost differential of about 10 percent, in a study of MediCal patients. Another analysis suggests that the indirect medical education adjustment adequately compensates teaching hospitals for the costs of teaching programs, but large public teaching hospitals (the "flagship municipal" hospitals) are financially at high risk from the combined effects of PPS, other cost containment strategies, and a high proportion of bad debt and charity care. The indirect medical
education adjustment appears arbitrarily high, but, in the opinion of one analyst, because it compensates for other factors omitted from PPS, it should not be reduced without redesigning the entire reimbursement formula.

The financial incentives in PPS may influence the development and adoption of new technology. Analysts cite infrequent weight recalibration, ICD-9-CM coding structure, DRG structure and incentives of per-case reimbursement as reasons why, in theory, technology may be stifled. The single study of this question, based on interviews, concludes that the intention to purchase magnetic resonance imagers has not been affected by PPS.

One important finding of our literature review is that the journal literature differs from the technical report literature in quality, comprehensiveness, and tone. The journal literature consists of early theoretical pieces and later studies that, for the most part, are critical of the system and its effect on hospitals. Unfortunately, these studies are often limited by the data and methods used. To the extent that these studies are accepted uncritically, they promulgate misunderstanding of the system. In contrast, the burgeoning technical report literature is based on larger, less selected datasets analyzed using methods that can accommodate the complex interrelationships among factors. This finding suggests that, at least at present, the Congress, HCFA, and some health policy analysts are working with an information set that differs markedly from that available to the majority of hospital managers, clinicians, and health researchers. The system as it is perceived and acted upon by each group is likely to be different.

The question of whether the DRGs adequately adjust for case mix in PPS is still open. Studies that compare the performance of DRGs with alternative methods for case mix adjustment suggest that the DRGs are still the best available method. Adverse effects of DRGs or PPS on hospitals or patients have yet to be documented in the literature. This may be due more to the lag in the availability of necessary data than to the lack of any effects. Studies now in progress or reported elsewhere in HCFA's report to Congress may lead us to reassess our conclusions in the future.
CONTENTS

PREFACE ................................................................. iii
SUMMARY ................................................................. v
TABLES ................................................................. xiii

Section

I. INTRODUCTION ......................................................... 1

II. PREDICTIVE PERFORMANCE OF DRGs ............................. 3
    Introduction ..................................................... 3
    Performance of the DRG Case Mix Index ..................... 6
    Performance of DRGs at the Patient Case Level ............ 10
    Surgical Compared with Nonsurgical DRGs ................. 11

III. UNEXPLAINED VARIATION IN DRGs: THE SEVERITY ISSUE ... 14
    Introduction ..................................................... 14
    The Severity Problem ......................................... 15
    Measuring Patients' Need for Treatment .................... 17
    Studies Comparing the Predictive Capabilities of Case Mix
    Adjustment Systems ............................................. 21
    Patient Condition and Use of Specific Resources ......... 25
    Summary .......................................................... 35

IV. UNEXPLAINED VARIATION IN DRGs: DATA QUALITY .......... 36
    Introduction ..................................................... 36
    Quality of UHDDS Clinical Data for DRG Classification ... 37
    Coding under Reimbursement Incentives .................... 46

V. VARIATION EXCLUDED FROM DRGs: PRACTICE PATTERNS ....... 50
    Introduction ..................................................... 50
    Admission Patterns ............................................. 52
    Charges .......................................................... 53
    Length of Stay .................................................. 53
    Use of Other Resources ....................................... 54
VI. CLINICAL AND HOSPITAL RESPONSE TO PPS .......................... 56
   Introduction ............................................. 56
   Response to Perceived Uncompensated Variation in Patient
   Condition .................................................. 56
   Change in Practice Patterns to Increase Efficiency ...... 58

VII. EFFECTS OF WEIGHTS AND PAYMENTS ................................. 62
    Introduction ............................................. 62
    Compression in the DRG Weights ........................... 62
    National Rates .......................................... 70

VIII. EFFECTS OF PPS: TWO EXEMPLARY POLICY ISSUES .................. 72
   Introduction ............................................. 72
   Teaching and Public Hospitals ............................ 72
   Responsiveness to Change in Technology ................. 80

IX. CONCLUSIONS .................................................................. 84

REFERENCES ....................................................................... 87
TABLES

1. Performance of DRG Case Mix Index .............................. 5
2. Performance of ICD-9-CM DRGs .................................... 7
3. Medical and Surgical DRGs ........................................... 8
4. Comparison of DRGs with Staging in Predicting Length of Stay ................................................................. 22
5. Percent Reduction in Sum of Squared Deviations of Nonsurgical DRGs and Staging in Predicting Length of Stay ................................................................. 24
6. Comparison of the Variation ($R^2$) in Cost/Case at Case Level Explained by Three Case Mix Measures ..................... 25
7. Correlation Coefficients: Total Adjusted Costs and Severity Measures ................................................................. 28
8. Comparison of Length of Stay, Mean Adjusted Costs, and Estimated Payment by Diagnosis-Related Group for Medical Intensive Care Unit and Routine Care Patients .......................... 30
9. Comparisons Between IOM and Original Data Source ............ 41
10. Decomposition of CMI Increase ........................................ 48
11. Percent Differences Above Nonteaching Hospitals in Direct Patient Care Costs, Adjusting for Case Mix and Physician Costs ................................................................. 78
I. INTRODUCTION

In 1983, the design and implementation of Medicare's Prospective Payment System (PPS) were swiftly accomplished. The old cost-based reimbursement for Medicare inpatients was replaced by a prospectively determined payment for each hospital discharge. The problem of adjusting for the differences in the types and complexity of cases seen at different hospitals was handled by using the best case mix adjustment method then available, Diagnosis Related Groups (DRGs).\(^1\)

Little was known about how the DRGs would perform within PPS. Developmental research on DRGs and experience with an early version of DRGs in New Jersey's prospective payment system had led to a revision of the first DRG system. The revised system classified patients into one of 467 categories based on diagnosis, procedures, age, and sex.\(^2\) Whether the shortcomings of the first system were adequately resolved in the subsequent revision of the DRGs was still a matter of debate. Congress therefore requested the Health Care Financing Administration (HCFA) to prepare a report on whether case mix adjustment in PPS could be improved by refining or replacing DRGs. The purpose of this literature review is to provide background information for this report to Congress on refining case mix adjustment in PPS.

---

\(^1\)This review assumes basic knowledge of DRGs and PPS. Overviews of history and issues concerning DRGs and PPS include HCFA's Report to Congress, Refining Case Mix Adjustment in Medicare's Prospective Payment System: Severity Adjustments, Outlier Payments and Other Options, Washington, D.C., 1986; U.S. Congress, Office of Technology Assessment, Medicare's Prospective Payment System: Strategies for Evaluating Cost, Quality and Medical Technology, Washington, D.C., 1985; HCFA, Health Care Financing Review, 1984 Annual Supplement; and Vladeck, 1984b.

\(^2\)In this report, we examine the DRGs as redesigned at Yale (Fetter et al., 1982) to conform to the revised coding scheme, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), and as implemented in PPS. We discuss the original DRGs (18 DRGs) developed at Yale in the mid 1970s (Fetter et al., 1980) only when the findings emphasize such issues as the differences between teaching and nonteaching hospitals.
We began the review with computerized literature searches such as MEDLINE, encompassing both academic and trade literature, from 1982 through 1985. The list of references was supplemented with published work cited in articles identified in the original search, prepublication drafts, and unpublished reports. The boundary years were selected to examine the DRGs in PPS planning and implementation. References before 1982 were limited to background material or descriptive work on case mix adjustment; references cited from 1986 were first reviewed in prepublication drafts. It was not possible to limit the review entirely to case mix adjustment because the DRGs are enmeshed with other adjustments in PPS.

The literature provides no definitive studies of whether the DRGs are performing adequately as the case mix adjustment in PPS. Although much of the literature identifies potential flaws in the DRGs and in PPS, the evidence is not convincing. Studies that critically examine the DRGs contain errors in analysis and interpretation, indicating that the complexity of PPS has proved difficult to understand. In particular, many investigators have apparently not grasped that payments are intended to be equitable on the average over a large number of a hospital's Medicare cases.

These findings suggest that further research and education are needed. Studies reported in other appendixes to the report, or still in progress, may yield more definitive conclusions than the literature supplies regarding case mix adjustment in PPS. Education is needed to achieve a balanced perception of the tensions in the system, which on the one hand encourages hospitals to look at DRGs as small units of service and on the other hand requires a broad, long-term view on financial performance.

We first collect the studies that document the ability of DRGs to predict patient care costs. We then review criticisms of DRG performance, specifically relating to unmeasured variation in patients' need for treatment, data quality and physician practice patterns. After that, we summarize literature that examines hospital response to DRGs and PPS. The next section examines the effects of the methods used in determining DRG weights. In the last section, we review two problem areas for PPS, the teaching hospital and technology.
II. PREDICTIVE PERFORMANCE OF DRGs

INTRODUCTION

Our primary interest in this review is discovering how well DRGs are able to predict necessary patient care costs at the hospital level. The success of PPS depends fundamentally on the ability of the DRG classification system to group cases with similar expected costs of treatment. Perfect prediction of costs at the case level is not necessary; however, if the DRG system does not accurately classify cases, reimbursement under PPS can jeopardize patients, hospitals, or both.

The sudden implementation of PPS has fueled a rapidly growing literature on the DRG case mix adjustment method. This literature, often critical, reflects perceptions that DRGs are not sufficiently accurate to justify their use for case mix adjustment in PPS. In fact, there is little published evidence on the performance of the ICD-9-CM DRGs in predicting costs, charges, or length of stay. Most recent journal literature referring to the low explanatory power of DRGs cites research on the performance of the I8 DRGs. Studies that examine the performance of I9 DRGs alone or in comparison with other case mix adjustment systems are only now beginning to appear as technical reports.

A major reason for the lack of literature on the performance of the I9 DRGs is that access to sufficiently large, representative datasets is limited. Of the investigators who have reported data on the performance of DRGs, three use a national database (Cotterill, Bobula, and Connerton, 1985; Frank and Lave, 1985; Pettengill and Vertrees, 1982). These studies used HCFA's Medicare files. Statewide datasets of Medicare claims are also being used to indicate DRG performance (Coffey, 1985; Coffey and Goldfarb, 1984; Frank and Lave, 1985, Mitchell et al., 1984, Mitchell et al., 1985; West et al., 1985).

The only other type of dataset used in these studies is from hospitals subscribing to the Severity of Illness Index (SOII) system (Horn, Horn, and Sharkey, 1984; Horn, Horn, and Moses, 1984). SOII data
include all patients (rather than Medicare patients) from a self-selected group predominantly composed of large teaching hospitals, but they do indicate the performance of the DRG system despite these limitations.

The studies reviewed in this section examine performance of the DRG system in predicting cost per case at the hospital level of analysis, and costs, charges, or length of stay at the patient case level of analysis. The hospital level of analysis examines the DRGs' ability to predict costs for one hospital compared with other hospitals. The case level analysis predicts costs from one case to another, either within or across hospitals.

For purposes of reimbursement, it is more important to explain costs at the hospital level of analysis. The effects of PPS on particular hospitals or groups of hospitals can be discerned at this level. At the case level, variation unexplained by DRGs is important because a hospital that believes it can predict which patients or patient groups are more expensive may discriminate against them.

The DRG classification system can be decomposed in several ways. Studies that examine the differences in performance of surgical and nonsurgical DRGs are discussed in this section. Studies of the performance of individual DRG categories are considered in Sec. III.

In general, the studies show:

- At the hospital level (Table 1), the DRG case mix index in combination with the wage index, teaching activity, and urban or rural status explains 72 to 75 percent of the variation in cost per case.
- At the case level (Table 2), the results of DRG performance studies are variable, depending upon the dataset and dependent variables used in the study:
  - 16 to 18 percent of the variation in length of stay or costs is explained in untrimmed data. An exception occurs in data from Washington state, where the power of DRGs in explaining cost variation is 30 percent.
Table 1

PERFORMANCE OF DRG CASE MIX INDEX (HOSPITAL LEVEL)

<table>
<thead>
<tr>
<th>Source</th>
<th>Data Source</th>
<th>Sample</th>
<th>Independent Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Dependent Variable</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pettengill and Vertrees, 1979</td>
<td>MEDPAR</td>
<td>351 DRGs</td>
<td>CMI</td>
<td>1.081</td>
<td>.045</td>
<td>Medicare</td>
<td>Adjusted $R^2 = .72$</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></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident-bed ratio</td>
<td>.569</td>
<td>.042</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.149</td>
<td>.011</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small city</td>
<td>.002</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Cost weights)</td>
<td>1.012</td>
<td>.043</td>
<td>Medicare</td>
<td>Adjusted $R^2 = .72$</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></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident-bed ratio</td>
<td>.580</td>
<td>.045</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.109</td>
<td>.014</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small City</td>
<td>.001</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>
</tr>
<tr>
<td>Cotterill, Bobula, and Connerton, 1981</td>
<td>MEDPAR</td>
<td>358 DRGs</td>
<td>CMI (Charge weights)</td>
<td>.969</td>
<td>.041</td>
<td>Medicare</td>
<td>Adjusted $R^2 = .72$</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></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resident-bed ratio</td>
<td>.545</td>
<td>.045</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large city</td>
<td>.109</td>
<td>.014</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small City</td>
<td>.000</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>
</tr>
<tr>
<td>Horn et al., 1984</td>
<td>15 hospitals</td>
<td></td>
<td>CMI</td>
<td></td>
<td></td>
<td>Hospital cost/case</td>
<td>Adjusted $R^2 = .75$</td>
</tr>
</tbody>
</table>
Between 26 and 33 percent of the variation in costs or charges is explained in trimmed data. The Washington data are again exceptional: 48 percent of the variation in costs is explained.

- Within the DRG classification system, the surgical DRGs provide most of the explanatory power (Table 3). In trimmed data, surgical DRGs explain from 48 to 57 percent of the variation in costs, while the nonsurgical DRGs explain from 7 to 16 percent.

PERFORMANCE OF THE DRG CASE MIX INDEX

One way to study overall DRG system performance is to examine the performance of case mix index (CMI) in predicting the hospital average cost per (Medicare) case. The CMI is the average cost weight of Medicare discharges from the hospital where the average case in the average hospital has a weight of 1.00. The CMI reflects the expected costliness of a hospital's Medicare cases compared with those of other hospitals.

Because the CMI is based both on DRG classification of cases and on the weights established for each DRG, this measure of DRG performance reflects the classification system and particular weights used. (Derivation of the weights, criticisms of the weighting structure, and alternative weighting methods are discussed in Sec. VII.)

Usual practice in studies of the CMI is to test the ability of the CMI, along with other independent variables, to predict the average (Medicare) cost per case among hospitals. The fraction of total variation in cost per case explained by the independent variables is expressed by the $R^2$. The partial effect of each variable is indicated by its coefficient.

Table 1 presents three studies that examine the hospital case mix index's ability to predict average cost per case: Pettengill and Vertrees' (1982) original testing of the case mix index; Cotterill, Bobula, and Connerton's (1985) comparison of case mix indexes based on cost versus charge weights; and Horn et al.'s (1984) testing of the case mix index in 15 teaching hospitals. This table shows the performance of the CMI in national datasets and in a very small dataset.
<table>
<thead>
<tr>
<th>Source</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 Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Coffey and Goldfarb, 1984</td>
<td>394,000 cases</td>
<td>Maryland 1979-81</td>
<td>Medicare Untrimmed</td>
<td></td>
<td>420 DRGs</td>
<td>LOS</td>
<td>.16</td>
<td>94%</td>
</tr>
<tr>
<td>2. Mitchell et al., 1984</td>
<td>272,087 cases</td>
<td>New Jersey 1982</td>
<td>Medicare Untrimmed Trimmed</td>
<td></td>
<td>DRGs</td>
<td>Costs</td>
<td>.18</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>237,639 cases</td>
<td>North Carolina 1982</td>
<td>Medicare Untrimmed Trimmed</td>
<td></td>
<td>DRGs</td>
<td>Costs</td>
<td>.16</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>Not stated</td>
<td>Washington 1982</td>
<td>Medicare Untrimmed Trimmed</td>
<td></td>
<td>DRGs</td>
<td>Costs</td>
<td>.30</td>
<td>.48</td>
</tr>
<tr>
<td>4. West et al., 1985</td>
<td>28,047 cases</td>
<td>South Carolina 1981</td>
<td>Medicare Trimmed</td>
<td></td>
<td>DRGs</td>
<td>Charges</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>5. Horn, Horn, and Sharkey, 1984</td>
<td>19,122 cases</td>
<td>1 University teaching hospital All Patients</td>
<td>Not stated</td>
<td></td>
<td>436 DRGs</td>
<td>Charges</td>
<td>.33</td>
<td>85%</td>
</tr>
<tr>
<td>6. Horn, 1985</td>
<td>5,580 cases</td>
<td>1 University teaching hospital All Patients</td>
<td>Not stated</td>
<td></td>
<td>401 DRGs</td>
<td>Charges</td>
<td>.29</td>
<td>76%</td>
</tr>
<tr>
<td>7. Horn et al., 1984</td>
<td>7,500 cases</td>
<td>15 hospitals All Patients Untrimmed</td>
<td>DRG Cost Weights</td>
<td></td>
<td>Cost/Case</td>
<td>.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3
MEDICAL AND SURGICAL DRGs

<table>
<thead>
<tr>
<th>Source</th>
<th>Sample Description</th>
<th>Source</th>
<th>Sample</th>
<th>Medicare /All</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>
</table>
| 1. Mitchell et al., 1984    | New Jersey         | 1982   | 272,087 cases | Medicare       | Trimmed      | Charges            | .11     | .48      | .32     | R²  
|                             |                    |        |        | Untrimmed     |              |                    | .06     | .31      | .18     | R²  
|                             | North Carolina     | 1982   | 237,639 cases | Medicare       | Trimmed      | Charges            | .09     | .51      | .32     | R²  
|                             |                    |        |        | Untrimmed     |              |                    | .04     | .33      | .16     | R²  
| 2. Mitchell et al., 1985    | Michigan           | 1982   | Not Stated | Medicare       | Trimmed      | Charges            | .10     | .49      | .30     | R²  
|                             |                    |        |        | Untrimmed     |              |                    | .06     | .34      | .17     | R²  
|                             | Washington         | 1982   | Not Stated | Medicare       | Trimmed      | Charges            | .16     | .57      | .48     | R²  
|                             |                    |        |        | Untrimmed     |              |                    | .09     | .39      | .30     | R²  
| 3. West et al., 1985        | South Carolina     | 1981   | 28,047 cases | Medicare       | Trimmed      | Charges            | .07     | .55      | .26     | R²  
| 4. Coffey, 1985             | Maryland           | 1979-81| 268,358 cases | Medicare       | Trimmed      | LOS                | .12     | --       | --       | R²  
|                             |                    |        |        |                |              |                    | .859    | --       | --       | Average coefficient of variation  
|                             |                    |        |        |                |              |                    |         | --       | --       | Average coefficient of variation  
|                             |                    |        |        |                |              |                    | 270,928 cases | Medicare Untrimmed | LOS   | .09 | -- | -- | R²  
|                             |                    |        |        |                |              |                    | .955    | --       | --       | Average coefficient of variation  
| 5. Frank and Lave, 1985     | HCFA DRGs          |        |        | Medicare       | LOS          |                    | .954    | .808     | --       | Average coefficient of variation  
|                             |                    |        |        |                |              |                    | (0.191) | (0.228)  | SD |    |    | Average coefficient of variation  
|                             |                    |        |        |                | Costs        |                    | .963    | .801     | --       | Average coefficient of variation  
|                             | HCFA               |        |        |                |              |                    | (0.194) | (0.703)  | SD |    |    | Average coefficient of variation  
|                             |                    |        |        |                | Medicare Untrimmed | Charges | .971    | .857     | --       | Average coefficient of variation  
|                             | Maryland           | 1979-81|        |                |              |                    | (0.666) | (0.292)  | SD |    |    | Average coefficient of variation  
|                             |                    |        |        |                |              |                    |         |          |          |        |
These studies rely almost entirely on the same set of independent variables used in the PPS formula. The set of variables is important because variables omitted from the equations may affect the apparent predictive power of the included variables. Thus, the coefficient on CMI (or any other included variable) may change depending on which other independent variables are included in the model.

In the Pettengill and Vertrees study and the Cotterill, Bobula, and Connerton study, the equations explain 72 percent of average hospital Medicare cost per case, whether cost weights or charge weights are used. Coefficients of the case mix index are all close to one. This means that each increase in a hospital's CMI is reflected in an approximately proportional increase in average costs; but the difference from true proportionality is an important issue called "compression," which is discussed in Sec. VII.

Cotterill, Bobula, and Connerton's study confirms the original findings of the Pettengill and Vertrees study and, additionally, shows that charges are an effective basis for weight determination. (The coefficients listed for cost data for Cotterill's study in Table 1 are essentially those used in the reimbursement formula until October 1985). Based on the findings of the Cotterill study, the fiscal year 1986 DRG weights were recalibrated using charge data.

In the Horn et al. (1984) study the DRG case mix index alone explains 75 percent of the variation in hospital average cost per case among 15 hospitals. As noted earlier, Horn et al.'s data are from a small sample of 15 teaching hospitals, and less variation in per case cost is expected in such a group than in the universe of hospitals. Although Horn's $R^2$ appears close to that of the other studies, it is actually much higher because other predictive variables--such as wage index, teaching status, and bedsize--are not included in the predictive equation. The apparent power of the DRG CMI in this study may be largely due to random variation, because the lower 95 percent confidence limit for a sample of 15 is .45.
PERFORMANCE OF DRGs AT THE PATIENT CASE LEVEL

The second group of studies examines the ability of the DRG classification system to predict charges, costs, or length of stay at the patient case level of analysis. Variation that is not explained at the case level is important because it can (1) give hospitals a reason to discriminate against predictably costly patients; (2) create financial hardship for hospitals whose patients have above average treatment needs not measured by the DRG system; and (3) leave a potentially large amount of random risk, which can hurt hospitals, especially small hospitals, by chance alone.

One important methodological consideration when looking at patient level datasets is whether extremely high cost patients ("outliers") have been excluded (or "trimmed") from the analysis. Because extreme outliers contribute so much to variance, DRGs usually explain a higher proportion of the variance in trimmed than in untrimmed datasets. In addition, because PPS has special provisions for outlier cases, performance with trimmed data is more relevant.

Table 2 shows that in five states, the DRGs explain 16 to 18 percent of the variation in resource use in an untrimmed dataset, and between 26 and 32 percent of the variation in a trimmed dataset. This is true when resources are measured by costs, as in the Mitchell et al. studies, and when resources are measured by length of stay, as in the Coffey and Goldfarb study.

Data from the state of Washington, however, indicate that DRGs can explain up to 18 percent more of the variation in costs. Mitchell et al. attribute the greater explanatory power in the state of Washington dataset to "greater homogeneity of physician treatment patterns"; average lengths of stay are shorter in Washington within four DRGs than they are in New Jersey, North Carolina, and Michigan.

Horn et al.'s studies (Table 2, nos. 5 through 7) represent DRG performance in small samples. Two of the studies (5 and 6) report data from a single university hospital, and the third included 15 teaching hospitals. Horn et al.'s data exclude cases that cannot be abstracted according to SOII as well as DRGs 468 through 470, a total of 30 percent of the admissions (Horn et al., 1984). In the 15 hospital study, DRG
cost weights were able to explain 28 percent of the within hospital variation in the observed cost per case.

These studies appear to indicate that DRGs explain less variation in resource use at the case level than at the hospital level. However, it is misleading to compare case level $R^2$'s with hospital level $R^2$'s because aggregating to the hospital level inherently reduces variation.

In sum, when considering hospital level and case level power of DRGs, it is important to realize that the hospital level statistics are of greatest immediate importance because they predict the overall ability of the system to pay hospitals appropriately. Case level analysis is important because of the influence residual variation can have on the system.

**SURGICAL COMPARED WITH NONSURGICAL DRGs**

The DRGs can be decomposed into two large subgroups, distinguishing between cases treated surgically and those managed without surgery. The studies in this section examine whether the surgical DRGs differ from the nonsurgical (often referred to as "medical") DRGs in explaining costs or length of stay.

In Table 3, we see that the power of DRGs in explaining costs of Medicare patients stems primarily from the surgical DRGs, which explain around one third of the variation in cost per case in an untrimmed dataset, to over 50 percent in a trimmed dataset. Medical DRGs explain one-tenth or less of the cost per case with untrimmed data, and in the trimmed dataset from Washington explain 15 percent.

Of the studies in Table 3, only one was undertaken to show the relative performance of DRG subgroups. Frank and Lave (1985) compared psychiatric, surgical, and medical DRGs to determine their relative homogeneity, in support of the mandate to evaluate options for appropriate reimbursement for psychiatric conditions. Using the 1981 HCFA dataset and the Maryland Medicare file for 1979 through 1981, Frank and Lave first compared average coefficients of variation, using the DRG as the unit of observation. The results of the DRG level analysis are those presented in Table 3. Analysis of variance found the surgical DRGs had significantly lower coefficients of variation in all three datasets. The medical DRGs had significantly smaller coefficients of
variation than the psychiatric DRGs in the HCFA length of stay data, but
the psychiatric and medical coefficients of variation did not differ in
the cost and charge data. This result also held in subsequent analyses
restricted to large volume DRGs.

At the case level of analysis (not reported in Table 3), Frank and
Lave found almost 77 percent of the surgical patients were in DRGs with
coefficients of variation between .25 and .79. In contrast, only 16
percent of the medical and none of the psychiatric patients occurred in
DRGs with this low range of coefficients.

Frank and Lave conclude that the surgical DRGs perform acceptably,
but the DRG classification system "is only slightly better for medical
cases than it is for psychiatric cases." Although they expect the
variation within both medical and surgical DRGs to decrease under PPS
(because of improved coding and reduced inappropriate utilization), they
observe there is no reason to believe that these factors should
selectively apply to the medical or surgical DRGs.

Surgical DRGs clearly explain resource use far better than do the
nonsurgical DRGs. There are several possible reasons why surgical cases
(and the DRGs) would have less variation in resource use than
nonsurgical cases:

- Patients cleared for surgery may on average be less variable
  than cases treated without surgery. The health status of
  elective surgery patients will be generally good. As Coffey
  (1985) notes, "The decision to perform surgery may mean, in
  many cases, that the prognosis is good and the complications
  are few." One alternative case mix adjustment method, APACHE,
  incorporates this concept by scoring patients admitted for
  elective surgery at lower risk than medical or emergency
  surgery admissions (Wagner and Draper, 1984).

- Medical DRGs are more likely to include cases admitted for a
diagnostic workup, as well as those with a confirmed diagnosis.
The decision to perform a particular surgical procedure may,
however, imply a high degree of diagnostic certainty. As
Coffey (1985) observes, "Surgical treatment may on average mean
that the diagnosis is certain and that the time spent
determining a diagnosis, which can vary tremendously depending on the problem, either is not necessary, has been done prior to hospitalization, or is of such an obvious or critical nature that the treatment is determined immediately or at an expected rate in most cases."

- Treatment patterns for the diagnostic workup or the medical management of a condition may vary more than treatment patterns associated with a particular surgical procedure (Coffey, 1985; Frank and Lave, 1985).
- Elective surgeries are planned in advance and may already exhibit efficiencies in outpatient diagnostic testing and inpatient scheduling.
- The reliability of coding medical diagnoses is lower than coding surgical procedures. This would increase variation due to error in the medical DRGs.

The 85 to 95 percent unexplained variation in nonsurgical DRGs poses a problem for PPS. As Frank and Lave note, hospitals with a high proportion of medical cases are at greater risk of financial loss due to random variation in patients than are hospitals with a high proportion of surgical cases. With such a high level of variation unexplained, it is difficult to tell whether the distribution of cases across hospitals is uniform or whether to adjust payments if it is not. Such unexplained variation subjects hospitals and patients to the risks of unexplained patient level variation discussed in the last section. Such unexplained variation does not reduce the overall explanatory power at the hospital level documented above. It is unlikely that the recent revisions of DRG categories and weights would dramatically improve case level performance of the nonsurgical DRGs, but this issue remains to be tested.

The magnitude of the policy problem posed by unexplained variation in the DRGs depends in part on the causes. Among the explanations appearing in the literature are unmeasured patient severity, data that vary in quality, and practice pattern variation. These issues are discussed in the next three sections.
III. UNEXPLAINED VARIATION IN DRGs: THE SEVERITY ISSUE

INTRODUCTION
Variation in patient treatment needs not captured by DRGs poses a problem for PPS if it is unevenly distributed across patients, hospitals, or groups of hospitals. If high cost patients within a DRG can be systematically identified at admission, such patients may be denied access to some hospitals. This section examines possible sources of variation in treatment needs that are not measured by DRGs.

The literature conveys widespread perceptions that DRGs do not adjust sufficiently for differences in the "severity" of a patient's condition or differences in the stage or complexity of disease. Several alternative classification systems have been proposed to replace DRGs or refine them by further dividing each DRG category. Unfortunately, the issue is clouded by misunderstanding of case mix adjustment and PPS.

Errors in analysis or interpretation mar many of the studies published in clinical, hospital, and health services research journals. Some of the studies discuss patients who may be argued to be severely ill (emergency admissions or intensive care patients, for example). Other studies stratify cases within DRGs according to various alternative measures of clinical severity. All of these studies attempt to reconcile historical charges or "costs" to PPS reimbursement. However, many studies fail to make important PPS adjustments such as the indirect medical education adjustment, or to consider that PPS payments are expected to balance on average. Although these studies reflect only early experience because of publication lag time, they may promulgate early misunderstanding of PPS because they appear in a widely read literature.

The technical report literature is more current and often more sophisticated in approach. These studies underscore the complexity of the case mix adjustment problem. Simply adopting another way to operationalize severity of illness is unlikely to improve PPS. Researchers at the National Center for Health Services Research have shown that the Disease Staging method of case mix adjustment predicts
length of stay no better than DRGs, and that Medicare payments under
that method would not account for the greater intensity of service
rendered in teaching hospitals. Research conducted at the Center for
Health Economics Research demonstrates that neither Disease Staging nor
Patient Management Categories performs better than the DRGs at
predicting costs in statewide data. Both groups of researchers use
large datasets and accepted methods to test the case mix measures.
Unfortunately, these and other studies still in progress are not
available in the published literature to inform clinicians, hospital
managers, and other researchers.

Practicing physicians and hospital administrators who rely
primarily on the journal and professional literature may be unaware of
important findings in the technical report literature. It is not
surprising then, when practitioners, whose information is based on an
incomplete or even misleading picture of PPS, come to different
conclusions about the system than policymakers who have access to a
wider range of information. Although this situation may improve as the
published literature begins to reflect more experience with PPS, more
effort could be devoted to disseminating current research findings.

This section is divided into five parts. We first review some of
the misconceptions underlying the "severity" issue. Then we briefly
compare DRGs and other case mix adjustment systems and review some
proposals for improving the measurement of patient condition in DRGs.
In the third part we review studies that compare distribution of
reimbursement under different case mix adjustment systems. We then
examine studies of individual or selected groups of DRGs and summarize
the literature.

THE SEVERITY PROBLEM

The literature reflects an unfortunate tendency to label unmeasured
variation in the DRGs as "severity." Three issues are hidden in this
misconception.
Case Mix Measurement for Reimbursement Versus Clinical Purposes

The problems of measuring case mix for reimbursement purposes and for clinical purposes are not the same (Hornbrook, 1982). For reimbursement, we need a measure that predicts the costs of medically necessary treatment. It might include measurement of severity, treatment complexity, and treatment intensity. For clinical purposes, we need to predict mortality, difficulty of decisionmaking, degree of impairment, progression of disease, or some other clinical measure.

Differences in Clinical Perspective

Severity is defined and used differently by different people. For example, Smits, Fetter, and McMahon (1984) note that physicians use the term to refer to "the impact of the particular disease process on the patient's physiologic integrity," the probability of death or disability. The nursing definition adds psychological and dependency needs to the medical definition.

Differences in Suggested Sources of Variation

No studies have successfully distinguished differences in patients' treatment needs from other known sources of variation, such as variation due to data quality and practice patterns. Case mix studies are designed to predict dependent variables (costs, length of stay) that reflect historical patterns of these known sources of variation.

In a commentary on studies testing case mix adjustment measures, Eisenberg (1984) notes that they "emphasize the point that illness severity and appropriate patient treatment vary greatly across hospitals and patient categories." Varying levels of inappropriate care and physician practice patterns are cited by others (Frank and Lave, 1985; Gertman and Lowenstein, 1984; Lave, 1985b; Ommen and Conrad, 1984; Smits, Fetter and McMahon, 1984), which also contribute to the problem. Stern and Epstein (1985) observe that hospitals vary greatly in the quality and composition of services they render, both locally and across regions.
Eisenberg (1984) reiterates the point that all the variation within DRGs should not be attributed to variations in patient attributes. He warns that we do not want to "equate differences in hospitals' historical resource usage with differences in severity of illness." Young (1984) adds, "it should not be assumed that severity will necessarily be related consistently to costs, charges, or to any other measure of resources used in patient management."

MEASURING PATIENTS' NEED FOR TREATMENT
How DRGs Measure Case Mix Complexity

In a description of the design and development of the 19 DRGs, Averill (1983) distinguishes among five different dimensions of case mix complexity:

Severity of Illness refers to the relative levels of loss of function and mortality that may be experienced by patients with a particular disease.

Prognosis refers to the probable outcome of an illness including the likelihood of improvement or deterioration in the severity of the illness, the likelihood for recurrence and the probable life span.

Treatment Difficulty refers to the patient management problems which a particular illness presents to the health care provider. Such management problems are associated with illnesses without a clear pattern of symptoms, illnesses requiring sophisticated and technically difficult procedures and illnesses requiring close monitoring and supervision.

Need for Intervention to the consequences in terms of severity of illness that lack of immediate or continuing care would produce.

Resource Intensity refers to the relative volume and types of diagnostic, therapeutic and bed services used in the management of a particular illness.¹

¹Reprinted with permission from the publisher.
Averill further distinguishes between the clinical and administrative points of view regarding case mix complexity. Clinicians tend to interpret case mix as "the patients treated have a greater severity of illness, present greater treatment difficulty, have poorer prognoses and have a greater need for intervention." Administrators regard case mix complexity as "the resource intensity demands that patients place on an institution."

According to Averill, DRGs are most closely related to hospital administrators' view of case mix complexity:

The purpose of the DRGs is to relate a hospital's case mix to the resource demands and associated costs experienced by the hospital. Therefore, a hospital having a more complex case mix from a DRG perspective means that the hospital treats patients who require more hospital resources but not necessarily that the hospital treats patients having a greater severity of illness, a greater treatment difficulty, poorer prognosis or a greater need for intervention.

DRGs do take account of severity. As Eisenberg (1984) observes, "The question is not whether DRGs adjust for severity of illness; they clearly attempt to by assigning DRGs on the basis of surgical procedures, comorbidities, complications, and, in some cases, age and sex. The crucial question is whether DRGs adjust for severity of illness consistently enough."

The DRG system recognizes that patients' need for services is linked to other concurrent diseases of the patient (comorbidity). Comorbidities are handled by DRGs through the presence or absence of secondary diagnoses: if a patient has any secondary diagnosis found on a list of secondary diagnoses thought to result in greater resource use, the DRG classification may change.

Young (1984) believes that the problem of comorbidity deserves much greater attention. In the Patient Management Categories (PMCs) case mix adjustment system she developed, comorbidity is defined in great detail. Patients who have many related diagnoses may have only a single disease process (no comorbidity) or several diseases (true comorbidity). Different levels of comorbidity are also recognized; a distinction is
made between comorbidity that implies a more severely ill patient and one that does not. Whether comorbidity is actively managed during the hospital stay is also examined. Finally, comorbidity that requires separate patient management is distinguished from comorbidity where a single management strategy subsumes treatment for comorbid conditions.

DRGs, in addition to allowing for the presence of comorbidities, uses the concept of outliers. Mullin (1985) believes that these two features make DRGs "as sensitive to severity as any other reproducible system yet developed." DRG development work at Yale included specifying outlier trim points that were clinically and statistically specific for each DRG (Mullin, 1985). High trim points identified more severely ill patients, and low trim points were expected to identify less severely ill patients and possible inappropriate hospital utilization. The Yale trim points were not employed in PPS, and PPS trim points identify a smaller proportion of cases as outliers. Mullin argues that critics of severity adjustments within DRGs have not compared their case mix adjustment systems with the Yale DRGs, but rather with DRGs as implemented under PPS.

Proposals to improve DRGs include incorporating a measure of nursing intensity (Thompson, 1984; Smits, Fetter, and McMahon, 1984). The "value and amount of nursing resources used by individual patients during a hospital stay" has never been incorporated in traditional hospital accounting, and such data were therefore unavailable when DRGs were designed.

Refining DRGs with other case mix adjustment systems has also been proposed. For example, where information on medical severity is lacking because of problems in the coding scheme or medical nomenclature, Smits, Fetter, and McMahon (1984) suggest research into modifications with case mix measures such as APACHE. DRGs with high variation have been studied to show how classification improves with additional information supplied by other case mix adjustment measures such as SOII (Horn, Horn, and Sharkey, 1984; Horn et al., 1984) and Disease Staging (Conklin et al., 1984a; Conklin et al., 1984b; Conklin, 1985).
How Other Case Mix Adjustment Systems Measure Patients' Needs

Five case mix adjustment methods are frequently mentioned in the literature as potential refinements or replacements for DRGs: APACHE, Disease Staging, MEDISGRPS, Patient Management Categories (PMCs), and Severity of Illness Index (SOII). For a detailed comparison of the structural and performance characteristics of these systems, see Cretin and Worthman (1986).

Most of these systems were developed for diverse purposes, such as utilization review, patient care evaluation, or assessing the clinical severity of illness. Only PMCs were developed specifically for reimbursement (Young, Swinkola, and Zorn, 1982).

The clinical origins of APACHE, MEDISGRPS, and Disease Staging are reflected in clinically meaningful distinctions among patient scores or categories. Both APACHE (Knaus et al., 1981, 1985) and MEDISGRPS (Brewster et al., 1985) score patients' physiologic data and define severity in terms of the probability of death. They are applied as generic measures across diseases but are normally used in conjunction with a disease categorization system. Disease Staging focuses on the progression of the disease (Gonnella, Hornbrook, and Louis, 1984); higher stages mean a greater degree of body system involvement and severity. In contrast to the generic measures, Disease Staging is specific for the disease.

SOII is more similar to DRGs than the other systems in that it measures severity (by stage of primary diagnosis, severity of complications), complexity (by responsiveness to treatment, for example), and resource use (by the use of life support measures). The entire measure is said to measure the patient's "burden of illness" (Horn, Horn, and Sharkey, 1984b).

Of the case mix adjustment systems, only PMCs attempt to define patients' treatment needs explicitly. Panels of physicians defined 800 PMC categories based on the expected management of the disease ("components of care") (Young, 1985).
STUDIES COMPARING THE PREDICTIVE CAPABILITIES OF CASE MIX ADJUSTMENT SYSTEMS

Only two of the case mix adjustment systems (PMCs and Disease Staging) have been tested on comparative data by investigators other than the original developers of the system. This sub-section reviews three studies comparing Disease Staging with DRGs. We then present comparative data on predictive performance of Disease Staging, PMCs and DRGs.

Comparisons of Disease Staging with DRGs

On the assumption that clinical severity as defined by disease progression might define patients' treatment needs more appropriately than DRGs, investigators at the National Center for Health Services Research undertook a comparison of Disease Staging with DRGs (Coffey, 1985; Coffey and Goldfarb, 1984; Short and Coffey, 1984). These studies are the first comparison of case mix adjustment systems undertaken in large databases by investigators other than those who developed the system. They show that defining the costs of medically necessary treatment is considerably more complex than severity under the Disease Staging model of disease progression.

The major finding of Coffey and Goldfarb's comparison of DRGs and Disease Staging was that the two case mix measures produced large differences in case mix index and projected reimbursement among different types of hospitals. Indexes derived from Disease Staging varied little across types of hospitals. Although the indexes were higher for teaching, urban, public, and large hospitals than for nonteaching, rural, voluntary, and small hospitals, the differences were much smaller than under the DRG CMI. This finding was confirmed in the national database from the Hospital Cost and Utilization Project by Short and Coffey (1984). In a later comparison limited to nonsurgical cases, Coffey (1985) found no significant difference between the two systems in gains or losses by type of hospital.

Coffey and Goldfarb (1984) used state-wide data from Maryland from 394,000 patients (1979-1981), while the Short and Coffey study (1984) used 770,809 records from 1977. Coffey (1985) used 270,928 records, the
nonsurgical subset of the Maryland data. All three studies were limited to Medicare patients and compared Disease Staging with DRGs to explain case level variation in length of stay.

Table 4 presents selected results from the Coffey and Goldfarb study. Staging is shown two different ways. In the first, the Staging algorithm classifies the stage of the principal diagnosis only; this strategy requires the Staging system to classify the same data that DRGs use. The preferred use of Staging is to let the computer software search among the principal diagnosis and related secondary diagnoses to identify the underlying staged disease. In this study, many other combinations of DRGs with Staging were produced as the data dictated, with the number of groups ranging from almost 1500 to almost 4000. Clinically meaningful combinations of DRGs with Stages are being developed and tested by SysteMetrics (Conklin et al., 1984a; Conklin et al., 1984b; Conklin, 1985).

Overall, DRGs and Staging explain similar amounts of the variation in length of stay in untrimmed data, but the systems perform very differently in distributing revenue. Reimbursement effects were compared holding maximum reimbursement equal for DRGs and Staging, and examining differences in reimbursement by hospital type. Coffey and

Table 4

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Number of Groups</th>
<th>Percent Reduction in Sum of Squared Deviations</th>
<th>Average Coefficient of Variation Within Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRG</td>
<td>420</td>
<td>15.7</td>
<td>94.4</td>
</tr>
<tr>
<td>Staging (principal diagnosis only)</td>
<td>805</td>
<td>12.0</td>
<td>87.8</td>
</tr>
<tr>
<td>Staging</td>
<td>698</td>
<td>10.2</td>
<td>93.3</td>
</tr>
</tbody>
</table>

Goldfarb examined differences by teaching status, location, type of hospital control, and bed size. Holding other characteristics constant, Coffey and Goldfarb found that significant differences existed only for bed size: Larger hospitals would lose 3 percent on the average under Disease Staging compared with DRGs. Coffey and Goldfarb speculate that the large hospitals may treat a large volume of less severely ill patients as measured by the Disease Staging system. They also compared case mix indexes calculated for each case mix system. Both DRGs and Staging have higher case mix indexes for teaching, urban, public, and large hospitals. But the range of values found for nonteaching, rural, investor-owned, and small hospitals was much greater under DRGs than under Staging.

The greater spread under DRGs leads Coffey and Goldfarb to observe, "One wonders if DRG-based case-mix measures 'overstate' the true differences in severity among hospitals by confounding true severity with use of procedures." The DRG case mix index may reflect "the medical technology of each hospital" and establish "payment based on existing allocation of medical resources."

In exploring this issue further, Coffey and Goldfarb looked more closely at the composition of case mix in Maryland teaching hospitals. The distribution of nonsurgical DRGs was equal in teaching and nonteaching hospitals: In both teaching and nonteaching hospitals, 41 percent of patients in medical DRGs had complications and comorbidities (ccDRGs). Surgical DRGs, however, showed a different pattern: 34 percent of patients in teaching hospitals and 37 percent of patients in nonteaching hospitals fell into ccDRGs. The distribution of ccDRGs also showed different patterns by hospital type. Among all ccDRGs, 31 percent of patients were surgical in teaching hospitals, compared with 25 percent in nonteaching hospitals. DRGs without complications were distributed with 38 percent surgical in the teaching hospitals versus 29 percent in nonteaching hospitals. Finally, when Coffey and Goldfarb held case mix constant under Staging, they found that for every 100 patients, teaching hospitals perform 137 procedures and nonteaching hospitals perform 83.
The distribution of ccDRGs in this study may be confounded by data quality. Studies of coding consistently show unreliable coding in complex cases. One study (Johnson and Appel, 1984) indicates that tertiary hospitals are more likely to underreport complex cases than other hospitals.

In a further analysis, Coffey (1985) concentrated on the nonsurgical DRGs using the methods of the original study. Again, the Disease Staging and DRG case mix adjustment systems performed similarly in predicting length of stay for nonsurgical patients as noted in Table 5. Both trimmed and untrimmed data were examined.

Coffey observes that the failure of Disease Staging (which specifically focuses on the classification of medical conditions) to improve on nonsurgical DRGs may occur because nonsurgical DRGs and Disease Staging rely on the same medical diagnostic information for classifying patients, and neither attempts to classify according to variation in medical treatment differences. The study also suggests that differences between the two systems in the treatment of unrelated comorbidities (absent in Disease Staging and present to a degree in DRGs) and disease-specific severity distinctions (fundamental to Disease Staging and handled primarily through the comorbidity and complication list in DRGs) are insufficient to cause reimbursement differences between the two systems.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>Trimmed</th>
<th>Untrimmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsurgical DRGs</td>
<td>12.1</td>
<td>9.2</td>
</tr>
<tr>
<td>Disease Staging</td>
<td>9.2</td>
<td>7.2</td>
</tr>
</tbody>
</table>

Comparison of Disease Staging, PMCs, and DRGs

In a recent comparison of Disease Staging, PMCs, and DRGs, Calore (1985) found that neither PMCs nor Staging explain costs in the 1982 Michigan Medicare data better than DRGs. Table 6 shows that the PMC system explains about 15 percent of the variance in cost per case in the untrimmed data set and 26 percent in the trimmed data set, while Staging explained 10 percent of cost variation in untrimmed data and 17 percent in trimmed data.

Both PMCs and Staging benefit from adding information about whether a case was a medical case or a surgical case. 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.

PATIENT CONDITION AND USE OF SPECIFIC RESOURCES

Several studies of the clinical severity issue have addressed specific patient characteristics, diseases or treatments. Some of these studies suggest specific solutions, while others suggest that the authors do not fully understand how PPS pays hospitals.

Table 6

<table>
<thead>
<tr>
<th>Cases</th>
<th>Patient Management Categories</th>
<th>Disease Staging</th>
<th>DRGs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trimmed</td>
<td>Untrimmed</td>
<td>Trimmed</td>
</tr>
<tr>
<td>Medical</td>
<td>.11</td>
<td>.07</td>
<td>.13</td>
</tr>
<tr>
<td>Surgical</td>
<td>.51</td>
<td>.35</td>
<td>.49</td>
</tr>
<tr>
<td>All</td>
<td>.26</td>
<td>.15</td>
<td>.17</td>
</tr>
</tbody>
</table>


aData are from 300,122 Medicare admissions, Michigan, 1982.
Identifying Expensive Patients: Emergency Admissions

Some authors have examined admissions generated by the hospital emergency department as a proxy for patient severity. Munoz et al. (1985) assess the financial effect of admissions to the hospital through its emergency department. Over 8000 admissions from 1983 and 1984 were examined, of which 40 percent were Medicare patients. Rate calculations for 1983 used 75 percent hospital-specific and 25 percent national rates. For 1984, rates were 50 percent hospital-specific and 50 percent national. Results showed consistent losses for all admissions when charges were compared with projected DRG reimbursement, although losses under Medicare were greater. When costs were compared with expected DRG revenue, results varied by year and Medicare status. Both groups showed profit in 1983, with Medicare profit around 25 percent of that for non-Medicare patients. Decreased revenue projected for 1984 resulted in loss for the Medicare patients.

This study, however, needs attention to the appropriate comparison groups and appropriate denominator, as is true for many of the studies spawned by PPS. First, it is important to understand how the experience of emergency admissions compares with that of nonemergency admissions, and whether charge differences reflect actual costs. Second, the potential financial risk for admissions generated by emergency departments must be set off against the potential profits both from operating the emergency department and from having the department part of the hospital's array of services. For example, a report by Powills and Matson (1985) indicates that emergency departments are establishing themselves as profitable by successfully competing with freestanding minor emergency clinics for patients who are less seriously ill.

Identifying Expensive Treatments—Intensive Care as an Example

Concern has been expressed that the DRG relative weights may influence the choice of patient management strategy. As Lave (1984) has observed, "if the payment to marginal cost relationship varies across the alternative treatment modalities the treatment selected may be influenced by payment levels." These effects may be observed across DRGs, in the choice of surgical versus medical management of the same
condition, the use of medical intensive care, and specialized treatments in the care of burn, trauma and oncology patients.

Butler, Bone, and Field (1985) and Coulton et al. (1985) examine the effects of treatment in a medical intensive care unit (MICU) on the difference between reimbursement and costs over selected DRGs. The discrepancy in costs can be readily identified for hospital action to improve the profit margin.

Coulton et al. (1985) studied the costliness of intensive care versus routine care for patients across several DRGs. The sample consisted of 1,485 patients from 1983, of whom 305 spent a portion of the hospital stay in the MICU. Both Medicare and non-Medicare patients were included, but the proportions are not reported. Charges were reduced to costs using cost-to-charge ratios; capital and direct medical education costs were excluded, but indirect teaching costs were included. Reimbursement rates were calculated to approximate an all-payer DRG-based payment system. New Jersey 1983 DRG weights were used because these were developed for a combined Medicare and non-Medicare population. The hospital's 1984 DRG payment rate was adjusted downward twice: by 8 percent to approximate costs in 1983, and by an additional 20 percent to adjust for the difference in costs for Medicare patients and all patients at this hospital in 1983. Indirect medical education was excluded from payment calculations, and rates were 75 percent hospital, 25 percent regional.

The authors studied 13 DRGs in which the proportion of MICU patients ranged from 7 to 48 percent. Three of the DRGs were examined more intensively, by assigning severity scores (Acute Physiology Score--APS--of APACHE) to the patients. APS was used in three different ways: on admission, total over the entire length of stay, and an average to eliminate the effect of length of stay.

In 10 of the 13 DRGs, MICU costs were significantly greater than those for routine care, and the costs of MICU patients exceeded estimated payment rates for all 13 DRGs. Average loss per patient using the MICU was $1,795; average gain for routine patients was $337 per patient. Overall loss was $101 per patient. Medicare patients examined separately had losses of $153 per patient.
Severity results, as measured by APS, were highly variable for the three DRGs studied. Correlation coefficients for hospital admission APS are presented in Table 7. Although MICU patients with chronic obstructive pulmonary disease (COPD) have strong positive correlations of APS with cost, MICU patients with bronchitis show a strong negative correlation coefficient. The elimination of death cases did not change the severity results. Differences in severity among the small number of patients in this study did not appear to be useful in distinguishing the MICU patients from the routine patients, and there was a wide range of severity of illness among MICU patients.

Coulton et al. conclude that the costs of MICU care relative to PPS reimbursement could lead to hospital decisions to limit admission of patients requiring intensive care or to reduce the supply of intensive care beds.

Table 7
CORRELATION COEFFICIENTS: TOTAL ADJUSTED COSTS AND SEVERITY MEASURES

<table>
<thead>
<tr>
<th>Measure of Severity</th>
<th>MICU</th>
<th>Routine</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRG 88: COPD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital admission APS</td>
<td>0.70&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.10</td>
<td>0.57&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total APS</td>
<td>0.96&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.49&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.92&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Average APS</td>
<td>0.65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.21</td>
<td>0.57&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DRG 89: Simple Pneumonia and Pleurisy, Age &gt;69 and/or CC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital admission APS</td>
<td>0.20&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.38&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.45&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total APS</td>
<td>0.65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.81&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.78&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Average APS</td>
<td>-0.33</td>
<td>0.33&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.26&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DRG 96: Bronchitis and Asthma, Age &gt;69 and/or CC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital admission APS</td>
<td>-0.54&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.21</td>
<td>0.06&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total APS</td>
<td>0.55&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.82&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.76&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Average APS</td>
<td>-0.29</td>
<td>0.16</td>
<td>0.10</td>
</tr>
</tbody>
</table>

SOURCE: Coulton et al., 1985. Reprinted with permission from publisher and first author.

<sup>a</sup>P < 0.05.
These data could lead to other plausible conclusions. For example, the lack of association between level of care and the case mix adjustment measure could also indicate that the MICU is being used inappropriately. McClish et al. (1985), reporting on the same MICU sample, note, "Between 35% and 38% of the MICU patients were not critically ill at the time of admission (APS no greater than 10); only 10% of these patients ever received a score greater than 10 during their ICU stay. Of the ward patients, 27% had an APS over 10 sometime during their ward stay." Table 8 also shows the following:

1. Seven of the DRGs were overall "winners" for the hospital regardless of the patients' level of care. Two of them are among the three DRGs with the highest proportions of patients receiving MICU care.

2. Of the six "loser" DRGs, three have routine care costs (as well as MICU costs) in excess of the estimated payment. These DRGs would lose money for the hospital regardless of the level of care.

3. The costs for all patients in one DRG (#316: Renal failure without dialysis) are so high relative to reimbursement that eliminating this DRG from the study would give the hospital a net gain of $26.00 per patient, given the mix of MICU and routine patients.

Finally, the methods used to calculate payments virtually assure that payment will not cover costs. First, indirect medical education costs are included in cost calculations, but excluded from payment calculations. Second, reducing payment by 20 percent in addition to using the New Jersey weights (which adjusted adequately for payment differences under an all-payer system) arbitrarily created additional payment deficit. The flaws in this study are not atypical and underscore the difficulty of designing studies that adequately address the complexities of PPS.
<table>
<thead>
<tr>
<th>DRG</th>
<th>Description</th>
<th>Patient</th>
<th>N</th>
<th>Mean Length of stay</th>
<th>Mean Adjusted Costs</th>
<th>Estimated Payment</th>
<th>Overall Winners/ Losers</th>
<th>Loss on Routine</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.</td>
<td>Specific cerebrovascular disorders except transient ischemic attack</td>
<td>MICU</td>
<td>15</td>
<td>11.9</td>
<td>5,402</td>
<td>4,028</td>
<td>4,275</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>182</td>
<td>12.2</td>
<td>5,210</td>
<td>4,028</td>
<td>4,275</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>197</td>
<td>12.2</td>
<td>5,210</td>
<td>4,028</td>
<td>4,275</td>
<td>L</td>
</tr>
<tr>
<td>24.</td>
<td>Seizure and headache, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>11</td>
<td>8.6</td>
<td>6,933</td>
<td>2,254</td>
<td>3,386</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>70</td>
<td>6.6</td>
<td>6,933</td>
<td>2,254</td>
<td>3,386</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>81</td>
<td>6.9</td>
<td>6,933</td>
<td>2,254</td>
<td>3,386</td>
<td>W</td>
</tr>
<tr>
<td>88.</td>
<td>Chronic obstructive pulmonary disease</td>
<td>MICU</td>
<td>15</td>
<td>9.8</td>
<td>8,364</td>
<td>3,008</td>
<td>3,892</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>51</td>
<td>5.5</td>
<td>8,364</td>
<td>3,008</td>
<td>3,892</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>66</td>
<td>6.5</td>
<td>8,364</td>
<td>3,008</td>
<td>3,892</td>
<td>W</td>
</tr>
<tr>
<td>89.</td>
<td>Simple pneumonia and pleurisy, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>14</td>
<td>13.3</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>93</td>
<td>7.5</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>107</td>
<td>8.3</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td>96.</td>
<td>Bronchitis and asthma, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>15</td>
<td>8.9</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>68</td>
<td>6.4</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>83</td>
<td>6.8</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td>97.</td>
<td>Bronchitis and asthma, age 18-69 without CC</td>
<td>MICU</td>
<td>11</td>
<td>5.5</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>101</td>
<td>5.9</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>112</td>
<td>5.8</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td>127.</td>
<td>Heart failure and shock</td>
<td>MICU</td>
<td>108</td>
<td>7.8</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>163</td>
<td>7.3</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>271</td>
<td>7.5</td>
<td>3,008</td>
<td>3,008</td>
<td>3,008</td>
<td>W</td>
</tr>
<tr>
<td>132.</td>
<td>Atherosclerosis, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>32</td>
<td>6.3</td>
<td>3,277</td>
<td>3,277</td>
<td>3,277</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>35</td>
<td>3.4</td>
<td>3,277</td>
<td>3,277</td>
<td>3,277</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>67</td>
<td>4.8</td>
<td>3,277</td>
<td>3,277</td>
<td>3,277</td>
<td>W</td>
</tr>
<tr>
<td>174.</td>
<td>Gastrointestinal hemorrhage, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>12</td>
<td>6.6</td>
<td>3,686</td>
<td>3,686</td>
<td>3,686</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>89</td>
<td>6.0</td>
<td>3,686</td>
<td>3,686</td>
<td>3,686</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>101</td>
<td>6.0</td>
<td>3,686</td>
<td>3,686</td>
<td>3,686</td>
<td>W</td>
</tr>
<tr>
<td>182.</td>
<td>Esophagitis, gastroenteritis and miscellaneous digestive diseases, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>18</td>
<td>8.7</td>
<td>5,189</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>177</td>
<td>5.5</td>
<td>5,189</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>195</td>
<td>5.5</td>
<td>5,189</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td>316.</td>
<td>Renal failure without dialysis</td>
<td>MICU</td>
<td>18</td>
<td>10.0</td>
<td>3,270</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>37</td>
<td>5.4</td>
<td>3,270</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>55</td>
<td>5.9</td>
<td>3,270</td>
<td>3,270</td>
<td>3,270</td>
<td>W</td>
</tr>
<tr>
<td>403.</td>
<td>Lymphoma or leukemia, age &gt;69 and/or CC</td>
<td>MICU</td>
<td>9</td>
<td>12.6</td>
<td>10,377</td>
<td>3,747</td>
<td>3,747</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>85</td>
<td>7.6</td>
<td>10,377</td>
<td>3,747</td>
<td>3,747</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>94</td>
<td>8.0</td>
<td>10,377</td>
<td>3,747</td>
<td>3,747</td>
<td>W</td>
</tr>
<tr>
<td>450.</td>
<td>Toxic effects of drugs, age 18-69 without CC</td>
<td>MICU</td>
<td>27</td>
<td>2.7</td>
<td>1,558</td>
<td>1,343</td>
<td>1,225</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine</td>
<td>29</td>
<td>4.0</td>
<td>1,558</td>
<td>1,343</td>
<td>1,225</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>56</td>
<td>3.4</td>
<td>1,558</td>
<td>1,343</td>
<td>1,225</td>
<td>L</td>
</tr>
</tbody>
</table>

**SOURCE:** Coulton et al., 1985. Reprinted with permission from the publisher and first author. We are responsible for the two right columns.

*p < 0.05.*
Butler, Bone, and Field (1985) restricted the sample for their study of MICU treatment costs to Medicare patients. Costs were derived using 138 detailed cost-to-charge ratios, in contrast to the 20 used within the Medicare cost report. Reimbursement calculations accounted for capital and direct medical education passthroughs as well as the indirect medical education adjustment and outlier payments. The authors found that the average loss per patient in this group was $10,567; and for the 28 percent of these patients who died, a $21,651 loss per discharge. They conclude that hospitals in financial difficulty may find it necessary to "decrease or discontinue provision of medical intensive care units and other types of high technology care to severely ill patients."

In an accompanying editorial, Weinberg (1985) notes that hospitals faced with financial losses actually have two additional choices: pressing for a change in reimbursement or changing the way intensive care medicine is practiced. He argues that changing our priorities for medical intensive care would not only reduce costs, but also "eliminate unnecessary suffering for patients in their final days of life."

Butler's own data suggest, however, that the hospital as a whole is doing well under PPS. It is never precisely stated whether the comparison group of non-MICU patients in this study is restricted to Medicare patients, but this is likely to be the case. Assuming a Medicare comparison group, and using the stated proportion of MICU patients ($n = 446, 4.6\%$ percent of the total population), we can calculate that their analysis presents data on 9,695 Medicare admissions, of which 9,249 did not spend any time in the MICU. Average profit for these patients was $578 per discharge, for an estimated gross profit of $5,345,922. The gross loss on MICU patients was $4,712,882, for a net profit of $633,040.

In general, the studies of MICU treatment and its effects on hospital financial status are most difficult to evaluate because of problems with cost data, and their implications are difficult to assess without data on clinical outcome.
The actual costs of MICU and of routine care may be quite different from those estimated from charges. As Coulton et al. note, some routine patients may actually require a higher nursing intensity than some MICU patients, and efficiencies achieved through technology of an MICU may actually render ancillary costs lower than the costs of the same service provided under routine conditions.

Outcome data would be most useful in evaluating suggestions such as Weinberg makes regarding changes in medical practice. Garber, Fuchs, and Silverman (1984) provide a good model in that outcomes were determined not only at discharge, but also within the subsequent year.

Coulton et al. found no consistent relationship between severity as measured by APS and whether patients were cared for in routine or special care units. Although we can say little about true severity or the severity score itself from this study, the usual notions of the relation between intensity of care and clinical severity of illness are not supported by Coulton's findings.

Other Specialized Treatment Modalities

The reports in this subsection involve the use of specialized treatment units and technologies, but the patients are restricted to a single patient problem (with or without other complications), spread through several DRGs. Financial analysis is limited to comparing charges with reimbursement.

Jacobs (1985) used two trauma indexes to assess severity among 1,018 patients in DRGs 444, 445 and 446--Multiple trauma in patients aged >69 (444), 18-69 (445), and 0-17 (446). The trauma indexes--Champion's Trauma Score (1981) and Baker et al.'s Injury Severity Score (1974)--are physiologic scales that predict the probability of death. Jacobs does not describe the methods for determining "costs" or reimbursement. He found that patients who are the most severely injured die early and are therefore not so costly as those with less severe multisystem injuries.
In an address to the American Burn Association, Curreri (1985) indicates that burn centers experience greatest financial loss for the least severely burned patients. These patients had been referred to the burn center because of complications rather than the need for specialized burn care. Curreri's presentation was not intended to be a formal study, because his data include only eight Medicare patients overall, of whom only two were least severely burned.

In the field of oncology, there is concern with the costs of patients who are participating in clinical trials for cancer research. Katterhagen and Mortenson (1984) present data from 50 patients, six of whom required inpatient care for a total of 16 admissions during the first half of 1984. Each of these admissions generated losses, from $650 through $42,000. (The hospital treating these patients normally has lengths of stay below the Washington state average, which is already low.) Patients on the trials are compared with nonprotocol patients, who on the average produce a profit for the hospital. The authors conclude that the difference between cost and reimbursement risks the future of the clinical research. They propose a new DRG for these patients, to be reimbursed at cost--an "incentive neutral" solution. This suggestion opens the sensitive issue of how far the Medicare program should support clinical research by paying for experimental treatments.

Studies of Selected DRGs

Profitability of DRGs. A recent analysis of DRG "profitability" (Mendenhall, 1985) sought to determine which DRGs were likely to be consistent winners or losers. Mendenhall examined more than 580,000 Medicare cases submitted to the Commission on Professional and Hospital Activities (CPHA) by over 350 hospitals in 1983 and 1984. Medicare payment was estimated based on urban/rural and wage adjustments on national rates. Capital and medical education adjustments were not made, nor was the hospital-specific portion calculated. The difference between hospital-reported charges and estimated Medicare payment provided the basis for determining what was mislabeled in the article as "profit" or "loss." This analysis produced lists of the ten most and
least "profitable" DRGs, and comparisons of DRG pairs with and without complications and comorbidities. Without the medical education adjustment, however, cases treated in teaching hospitals automatically become least "profitable." In addition, comparison of charges with payments without adjusting for the difference between charges and costs can be very misleading.

**Clinically Oriented Analyses.** Beginning in 1985, a small number of DRG studies began to appear in the clinical journals (Ephgrave and Hunt, 1985; Munoz, Margolis, and Wise, 1985; Weinberger, Potts, and Brandt, 1985). The sample sizes are small and not restricted to Medicare patients.

Ephgrave and Hunt report on pancreatic pseudocyst, a single diagnosis within DRG 191 (Major Pancreas and Liver Procedure with Shunt), which only had 162 patients in the 1981 MEDPAR file. The small numbers problem is amply demonstrated by the fact that two hospitals and eight years of data were necessary to derive the sample of 115 patients; 23 patients from one hospital (and four years) were used for the cost comparisons in this study. Yet the authors conclude that DRG development should have separated the pancreatic from the liver procedures to create two DRGs, thereby further decreasing the number of patients in each. It seems unlikely that further reducing the applicable patients would improve estimates of appropriate costs, however. The relative weight for this DRG was increased in the recent revision of weights.

Munoz, Margolis, and Wise (1985) studied 46 patients with uncomplicated gastrointestinal hemorrhage, further subdividing the sample into groups with and without transfusion to assist in the analysis. The use of the transfusion "identifier" created two groups of patients differing in resource use. Not surprisingly, the patients with transfusions had greater costs, almost entirely associated with hematology and blood products. Despite the obvious nature of this analysis, it did help to identify possible efficiencies in each group.

Weinberger, Potts, and Brandt (1985) reviewed 96 patients with Rheumatoid Arthritis (RA) or Systemic Lupus Erythematosus (SLE) who would be classified into Connective Tissue Disorders, with (DRG 240) and without (DRG 241) age > 69 or comorbidity. DRG 240 showed consistent
use of resources among the patients, regardless of diagnosis. However, DRG 241 showed marked differences in charges and use of services, depending on diagnosis. SLE patients accounted for 36 percent more resource use than RA patients. Because only 15 percent of the sample were Medicare patients, who would be more likely to be classified in DRG 240, the authors conclude that PPS is unlikely to pose an immediate problem for practicing rheumatologists. They express concern, however, that if prospective payment is extended to non-Medicare patients, and if hospitals transfer their most severe patients to university hospitals, the heterogeneity in DRG 241 will be problematic.

Severity and Outcomes

The severity discussion needs to be enlarged by knowledge of medical care outcomes. Garber, Fuchs, and Silverman (1984) have shown that the intensity of service rendered to patients does not necessarily reflect differences in severity, nor does it necessarily result in improved long-term outcome. Until more studies on the effectiveness of medical care treatment patterns are available, the policy decisions will be made in the absence of clear information on this issue.

SUMMARY

The literature typically labels unmeasured variation in patient treatment needs "severity," but such variation must include not only the clinical severity of the patient's condition, but other factors that may influence clinical judgment as to necessary treatment. The literature tends not to address this problem clearly or explicitly and provides incomplete information regarding the success of DRGs (or other case mix adjustment systems) in accomplishing that purpose.

Despite lack of good evidence, there is a perception that DRGs do not adjust sufficiently for differences in patient condition that determine needed treatment, and some authors propose using alternative methods to modify high variation DRGs. At present, none of the case mix adjustment methods have demonstrated a general and reliable capacity to measure the differences in patient condition that determine the costs of medically necessary treatment. Testing all of these methods risks confusing variation in "severity" with variation caused by data quality problems or differences in practice patterns.
IV. UNEXPLAINED VARIATION IN DRGs: DATA QUALITY

INTRODUCTION

The second reason for unexplained variation in the DRGs may be the quality of data DRGs use to classify cases. The issue of data quality actually poses three problems for PPS. The first problem is the quality of the MEDPAR file used for calibrating DRG weights. This issue is now historical but it stimulated criticism and several small studies. The second problem is the contribution of imperfect data to unexplained variation in DRGs. The third problem is whether hospitals can manipulate data to maximize reimbursement.

Despite the number of studies cited in this section, the last two questions are still open because no studies have examined these issues directly. Answering these questions requires studies that can factor out several simultaneously occurring changes. Studies should exclude the effect of changes in hospital and physician practice patterns, whether inspired by PPS or not. Three additional factors need to be separated, preferably by means of direct examination of the medical record. First, hospitals are reporting more complete data, whether they are manipulating it or not. Second, the data also will inevitably contain honest errors and ambiguous cases. Third, a few hospitals may even attempt to falsify cases. All of these changes affect the data that DRGs use to classify cases, creating differences between the data used to design the DRGs and data currently available. On current data, the DRGs could improve, deteriorate, or remain approximately the same in explaining Medicare cost per case.

Below we review the literature on data quality, including the clinical basis for diagnosis, the structure of the ICD-9-CM coding scheme for diagnoses and procedures, and the reliability of coding. This literature examines the Uniform Hospital Discharge Data Set (UHDDS), the standard data that Medicare requires for payment and the basic input data for DRGs, PMCs, and Disease Staging. We then discuss specific concerns raised within the context of PPS, including the effects of reliability on Medicare reimbursement and the quality of data.
in the MEDPAR file. We then review evidence on coding changes under the incentives of PPS.

QUALITY OF UHDDS CLINICAL DATA FOR DRG CLASSIFICATION

Unexplained variation in the DRGs could stem from three possible sources related to the UHDDS dataset: standards for assigning diagnoses, the structure of the coding scheme, and coding reliability. These problems will surface for any case mix adjustment system dependent on UHDDS diagnostic and procedure data. They exist independently of incentives to manipulate the data but could contribute marked variation in DRG classification.

Diagnostic Standards

Underlying the issue of clinical data quality is the problem of the validity of the diagnosis. In a recent discussion of the severity problem for DRGs, Gertman and Lowenstein (1984) questioned the standards on which certain diagnoses are made. The evidence for common conditions such as diabetes, coma, angina, gastrointestinal hemorrhage, and post-operative wound infections vary from the merely suggestive through the most dramatic clinical findings. Without explicit standards for these diagnoses, they argue, the diagnosis itself describes patients whose condition and need for medical treatment vary.

The concern with diagnostic standards is not new. In a summary of three Institute of Medicine (IOM) studies concerned with the reliability of coding, Demlo and Campbell (1981) recommend new developing standards for diagnoses because "part of the cause of unreliable diagnostic data is ambiguity in the criteria for designating diagnoses."

Some diagnostic uncertainty is to be expected, however. Simborg (1981) cites "medical vagaries and uncertainties in many diagnostic situations" as a reason for legitimate disagreement. For example, distinguishing "abdominal pain with a duodenal scar" from a "probable duodenal ulcer" may depend on the physician's practice style.

Iezzoni and Moskowitz (1984) provide an example of how the diagnostic "level" selected by the physician is eventually reflected in different DRG assignment. Depending upon the level of clinical investigation and diagnostic style of the physician, a given case may be
labelled "chest pain" (the symptom), "angina pectoris" (the clinical diagnosis, or "atherosclerosis" (the pathologic cause). In this example, each of the alternatives has a ICD-9-CM code that determines a different DRG. Chest pain patients are assigned to DRG 143, while angina patients are assigned to DRG 140, and atherosclerosis belongs in DRGs 132 and 133. Yet the patients within these groups may be very similar, differing only in level of diagnosis and codes selected for the case. This clinical analysis is buttressed by Iezzoni and Moskowitz's finding that average Part B costs are very similar across these DRGs, but it is not consistent with the differences in Part A costs, which are reflected in different DRG weights.

The problems of diagnostic accuracy, diagnostic style, and physicians' "varying use of the International Classification of Diseases" (Wennberg, McPherson, and Caper, 1984) may reduce the predictive capabilities of DRGs. The literature on quality of care suggests that diagnostic standards may be developed, but the methods for devising these must be carefully formulated. A full discussion of this literature is outside the scope of this review;¹ nevertheless, variation within DRGs begins with variation in the diagnoses that DRGs classify.

Structure of the Coding Scheme

The current coding scheme, the ICD-9-CM, has three features that contribute to within-DRG variation: the diagnosis classification, the procedure classification, and the use of "catch-all" categories.

Diagnosis. In the coding scheme, diagnoses are classified based on anatomy, and rarely provide a way to code clinical severity or acuity (Gertman and Lowenstein, 1984). For example, acute and chronic congestive heart failure can be represented by only a single code, as are both minor and massive heart attacks. Furthermore, important qualifying phrases are lost in coding (Mullin, 1985): both "rule-outs" and "probable" diagnoses are coded as if they were certain. Wirthschafter (1984) aptly summarizes these concerns as follows: there

¹See, for example, Brook (1973); Brook et al. (1977); Donabedian (1978); Greenfield et al. (1975, 1981); Lyons and Payne (1975); McAuliffe (1978); Nobrega et al. (1977); Williams and Brook (1978); Williamson (1971).
is a "loss of specificity from clinical judgment to the ICD-9-CM codes, because the coding system itself is less specific than the nuances of clinical description and diagnosis."

**Procedures.** Additional problems with the ICD-9-CM are encountered for the coding of procedures. In ICD-9-CM, procedures are grouped according to anatomy and what was accomplished, rather than surgical method (Mullin, 1985; Smits, Fetter, and McMahon, 1984). For example, Mullin notes that endoscopic examinations and open operations of the digestive tract receive the same code. The use of coding conventions can group within a single code both injection into an intravertebral disk and laminectomy with disk excision.

**Catch-alls.** As a classification system, ICD-9-CM relies on catch-all categories for rare or imprecisely described diseases or procedures. As medical science advances, these catch-alls may be used for coding newly developed procedures. These "not" categories ("not otherwise specified," "not elsewhere classified") must be used in any classification system that attempts to be entirely inclusive, especially within a rapidly changing field. Yet the lag time between revisions of the system means that both medical advances and nonspecific information will be grouped together in these "not" categories. When used for reimbursement the resulting variation implies an erratic relation between costs and payments (Smits and Watson, 1984).

These problems with ICD-9-CM affect not only the DRGs but also other case mix adjustment systems that use UHDDS data. The developers of Disease Staging and Patient Management Categories have found the imprecision of the system limits the accuracy of their case mix adjustment systems. For example, Louis et al. (1983) point out that ICD-9-CM uses a single code for both irreducible and strangulated hernias, so that the Disease Staging classification software that uses UHDDS data cannot make this clinically relevant distinction. And Young (1985) notes that PMCs must utilize procedure codes (such as dialysis) to distinguish between different levels of patient condition within a single diagnostic code (such as renal failure).

Possible solutions to this problem include revising the ICD-9-CM (Mullin, 1985; Smits and Watson, 1984) and reintroducing Current Procedural Terminology, Fourth Edition (CPT-4) for coding procedures
(Smits, Fetter, and McMahon, 1984). Work on a successor to ICD-9 (the internationally accepted basis for ICD-9-CM) has already started, but ICD-10 will not be ready for use until the mid-1990s; revision of the procedure codes, which are an American creation, could proceed more rapidly. The Mayo Clinic (Nobrega et al., 1985) has recently completed a study of the feasibility of basing surgical DRGs on CPT-4 and found the task feasible but not entirely straightforward. Either strategy would also require DRG revision. A national coding committee, an interim solution proposed by Smits and Watson (1984), could resolve issues of coding conventions, coding for new procedures, and possible introduction of special interim codes for new procedures. Such a committee could direct more attention to the problem and produce greater consistency of practice.

Early National Studies of Coding Reliability

The reliability of coding is also proposed as a reason for unexplained variation in the DRGs. A nationally representative study of the ICD-9-CM coding scheme has not been done, so the degree to which coding reliability may affect DRG variation and reimbursement cannot be assessed. The classic studies on coding reliability are those sponsored by the Institute of Medicine (National Academy of Sciences, 1977a and b, 1980). These separate reports are augmented by summaries in the literature (Demlo, Campbell, and Brown, 1978; Demlo and Campbell, 1981). The coding systems in use during the studied years (1974 and 1977) were different from the currently used ICD-9-CM. The studies constitute the only available national assessment of reliability, however, and the results are sobering.

The purpose of the original Institute of Medicine (IOM) study was to identify an existing data source as a reliable standard for assessing the influence of Professional Standards Review Organizations (PSROs). Dissemination of the results brought requests from HCFA and the National Center for Health Statistics (NCHS) for similar studies. The three studies examine reliability in three different data sources: The PSRO study was conducted on data submitted by hospitals to private abstract services (such as the Professional Activity Study--PAS--of CPHA), the HCFA study on submitted Medicare hospital claims; and the NCHS study on
data collected for the National Hospital Discharge Survey (NHDS). In each study, a team of Registered Record Administrators reabstracted a sample of patient abstracts, which were then compared with the original codes.

As noted in Table 9, agreement between the original data source and the IOM team ranged between 57 and 65 percent over all coded diagnoses. Over all cases, agreement on the principal procedure ranged from 71 to 79 percent. Because not all patients actually have procedures, the statistics for principal procedure are further analyzed, depending on the presence or absence of a coded procedure. When a procedure was coded on the original data document, there was agreement on 60 percent of the cases. When the original data indicated there was no procedure performed, agreement ranged from 76 to 90 percent.

The complexity of the diagnosis influenced reliability markedly. Agreement on chronic ischemic heart disease was lowest in all three studies, with agreement ranging between 30 and 37 percent. Coders also showed considerable disagreement for diabetes, with the two studies for which data are comparable indicating only 50 percent agreement for Medicare patients and almost 61 percent agreement for the private abstract service. Substantially higher agreement was achieved for simple diagnoses such as cataracts (approximately 95 percent overall) and inguinal hernia without obstruction (approximately 92 percent overall).

Table 9

<table>
<thead>
<tr>
<th></th>
<th>Private Abstract</th>
<th>Medicare Record</th>
<th>NHDS Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal diagnosis</td>
<td>65.2</td>
<td>57.2</td>
<td>63.4</td>
</tr>
<tr>
<td>Principal procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All cases</td>
<td>73.2</td>
<td>78.9</td>
<td>71.4</td>
</tr>
<tr>
<td>Procedure coded</td>
<td>66.0</td>
<td>56.6</td>
<td>60.1</td>
</tr>
<tr>
<td>No procedure coded</td>
<td>86.7</td>
<td>89.7</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Disagreement was attributed to coding errors, sequencing errors, and ambiguity. Coding errors consisted of assigning the incorrect code. Sequencing errors involved incorrect selection of the principal diagnosis or procedure when there was more than one of each. In general, sequencing errors were more frequent for the complex medical diagnoses than were coding errors. Ambiguity indicates legitimate (and unresolved) disagreement over correct coding or sequencing. In the Medicare study, ambiguous cases constituted 4.6 percent of all diagnoses and 1.7 percent of all procedures. The majority were sequencing ambiguities.

The effects of coding imprecision may be mitigated somewhat by examining results at higher levels of generality. In the old ICD-8 coding scheme, four digits were the maximum available precision, but it was possible to examine accuracy generalized to the three digit level, and in the Medicare study, aggregated to the DRG level as well.

DRG classification lessens the overall effects of coding errors, but the effect is smaller in complex diagnoses. In the Medicare study, a subset of groupable principal diagnoses agreed in 71.7 percent of cases at the DRG level, compared with 61.9 percent agreement at the four digit level and 68.2 percent at the three digit level. As with the diagnoses themselves, the level of agreement varied depending upon the level of complexity. Chronic ischemic heart disease "improved" from 36.8 percent at the four digit level to 38.6 percent at the DRG level, while diabetes improved from 49.7 percent to 56.2 percent.

These studies show that, independent of reimbursement issues, coding reliability is not high; and it is worse for complex diagnoses and for complex patients with multiple problems. Sequencing discrepancies and ambiguity account, together, for substantial variation in the reliable assessment of the patient's principal reason for admission.

However, the degree of correspondence between current coding methods and those utilized at the time of these studies is low. The Medicare study used 1974 data, in which diagnoses were coded in ICD-8 as adapted by HCFA, and procedures were coded according to Surgical Current Procedural Terminology (CPT), also modified by HCFA. There are no data
available from the national study of coding reliability for the ICD-9-CM coding system currently in use, which has five digit coding precision (Office of Inspector General, Department of Health and Human Services).

Studies of the Effect of ICD-9-CM Reliability on Medicare Reimbursement

Several studies have examined limited samples to investigate the effect of coding error on Medicare reimbursement. Early studies (Barnard and Esmond, 1981; Corn, 1981; Doremus and Michenzi, 1983) used I-8 DRGs and are omitted from this review. More recent studies do not demonstrate an improvement in coding reliability under ICD-9-CM (Johnson and Appel, 1984; Lloyd and Rissing, 1985; Zuidema, Dans, and Dunlap, 1984). The data assessed in these studies were gathered before PPS, however, so that studies do not tell us about coding quality now that the incentives have changed.

The purpose of Johnson and Appel's study was to assess the effect of error on reimbursement with ICD-9-CM DRG assignment, using over 138,000 cases from 26 hospitals submitted in 1980 and 1981. They compared DRG assignment based on Medicare claims data with that based on medical record abstracts and found DRG assignment agreed in 53 percent of the 1981 cases on the average, but tertiary care hospitals were consistently lower than the average. The average difference was a 4 percent gain in revenue under the medical records case mix index. Examining winners and losers under a "budget neutral" reimbursement adjustment (i.e., one in which total payments were held constant) on the hospital case mix indexes calculated from the medical records data showed almost 77 percent of the hospitals changing reimbursement by no more than 2 percent, but the largest percent distribution was a 7.5 loss. Johnson and Appel note that the tertiary hospitals were least likely to gain revenue under their simplified reimbursement calculation. They conclude that claims data in the MEDPAR file understates the case mix index, yielding an average 4 percent gain for hospitals under medical records data. Data quality for complex cases was the most deficient, systematically reducing payments for tertiary care hospitals and complex care.
Zuidema, Dans, and Dunlap (1984) report on errors in coded data that may be attributed to physicians. This was not a study, but a followup on submitted data that had served as the basis for a public policy report on unnecessary permanent pacemaker insertions and high mortality rates for cholecystectomy. The codes had originally been submitted in 1979 and 1980 in Maryland's Guaranteed Inpatient Revenue program. On reexamination of the cases, physicians found that 95 percent of the 610 cases labeled unnecessary permanent pacemaker insertions could be attributed to problems in coding either the diagnosis or the procedure. The problems stemmed from incorrect sequencing of the diagnosis justifying the permanent implant, or from improper wording of the procedure causing temporary pacemakers to be coded as permanent. Similarly, a portion of the high mortality rate for cholecystectomy could be explained by erroneous ordering of the procedures or, in one case, a miscoded death.

Lloyd and Rissing (1985) directly assess the role of physician error in the reliability of data reporting. Physicians were accountable for correct statements of all procedures and diagnoses, including determination of the principal diagnosis. Coding and keypunch errors were also assessed. Over 1,800 medical records from five VA hospitals in 1981 and 1982 were compared with the abstracted data form.

Lloyd and Rissing found 82 percent of the records were discrepant with the abstract in at least one data field: 62 percent of the errors were physician-caused, 35 percent were coding, and 3 percent were keypunch errors. Of the physician errors, 46 percent were missed procedures, 43 percent missed diagnoses, 5 percent inappropriate principal diagnoses, 4 percent inadequate terminology, and less than 1 percent inactive diagnoses labeled as active. (This is the only report on the frequency with which inactive diagnoses are recorded as secondary diagnoses.) Very few of the missed procedures were operating room procedures that changed DRG assignment, but these influenced revenue markedly. Missed diagnoses, however, more often affected assignment to a more complicated DRG.
The fiscal effect of all errors was examined across services and hospitals by calculating the difference in weighted work units, a measure of cost for the average VA patient in a DRG. Changes were not consistent throughout services or hospitals, but in general, those on the medical service yielded low mean changes spread over a large volume, while those on the surgical services showed the greatest differences per change. Overall Lloyd and Rissing found that more accurate coding could have increased payments from $250,000 to $980,000 per hospital.

Mitchell et al. (1984) also noted missed procedures in reconciling hospital and physician claims for Medicare patients in two states. Part B surgical bills totalling more than $400 were examined: 2 percent of New Jersey cases and 8 percent of North Carolina cases were reclassified from medical to surgical DRGs.

Assuming that these studies represent baseline reliability, they suggest that improved data reporting will contribute to an increased number of patients in complex DRGs and, to a lesser extent, to a higher proportion of surgical DRGs under PPS.

Quality and Sensitivity of Data in the MEDPAR File

With the introduction of PPS, issues of data quality became refocused. There were good reasons (in addition to indications of general low coding reliability) to question the quality of data in the MEDPAR dataset, which was subject to particular concern because it consisted of claims submitted to HCFA for reimbursement. Frequently, the claims were submitted by personnel in hospital billing departments based on incomplete charts. Such personnel are not usually as well equipped as medical records staff to abstract clinical information. Translation of narrative to codes by fiscal intermediaries or HCFA created the potential for further inaccuracy. Finally, the fiscal intermediaries were not required to submit specific codes for secondary diagnoses, which influenced the case mix calculation.

Pettengill and Vertrees (1982) tested the 1979 MEDPAR file in simulations of the effects of error. They found that error up to 30 percent caused differences in the case mix indexes no greater than 10 percent above or below the original indexes. However, 10 percent error
in the CMI could result in a large difference in reimbursement. This
test, moreover, was conducted on a dataset with secondary diagnoses coded
only as present or absent, and we might expect larger differences when
specific secondary diagnoses can be used to assign cases to the DRGs.

CODING UNDER REIMBURSEMENT INCENTIVES
Concerns with Current Data Quality: Sequencing and Accuracy

Concern with the effects of prospective payment on data quality has
primarily centered around sequencing diagnoses and procedures, either
because of appropriate ambiguity (Connell, Blide, and Hanken, 1984;
Simborg, 1981) or because of inappropriate manipulation in response to
the incentives to increase reimbursement (Simborg, 1981). The degree of
error identified in reliability studies to date indicates that it will
be difficult to distinguish between improvement in data quality and DRG
Creep.

The clinical and hospital industry literature underlines the
difficulty in making the distinction between appropriate ambiguity and
manipulation. Although some industry literature simply recommends
maximizing reimbursement through "sound coding policies" (Meadors and
Wilson, 1985), others are more explicit regarding the particular
approaches being used. Wirtzschacter (1984) indicates that the focus
within his hospital is on avoiding sequencing errors; after the
principal diagnosis, all subsequent diagnoses are to be listed by the
institution is devising a program to teach physicians accuracy in the
use of ICD-9-CM descriptors. D'Orazio and Goldschmidt (1985) indicate
that their hospital employs concurrent analysts to sequence diagnoses
for physician review on a DRG worksheet during the patient's stay,
apparently with the goal of optimizing Medicare reimbursement.
Sequenced diagnoses are subsequently reviewed and modified by the
physician as necessary.

These observations suggest that hospitals are probably coding more
completely but do not show whether the data are more accurate. Now that
reimbursement depends on the completeness of diagnosis and procedure
coding, we would expect coding to improve because of heavier reliance on
appropriately trained personnel to perform coding, better dissemination
of coding rules, and increased importance attached to the activity. Independent of incentives to upcode, any of these reasons would result in improved coding reliability.

Studies of Changes in Coding under PPS

There is some evidence that coding of complications occurs more frequently under prospective case-based payment. The effects of prospective payment may be evident in the different coding styles found by Iezzoni and Moskowitz (1984) between New Jersey and North Carolina 1982 Medicare data. For example, diabetics (DRG 295) in North Carolina (no prospective payment) are coded showing 35.2 percent have complications; in New Jersey (under prospective payment), 46 percent are complicated diabetics. However, these differences could reflect nothing more than regional differences in coding, admission practices, or health status.

Changes in coding practices have been indirectly documented between 1981 and 1984 as a result of PPS. In a statistical study of the reasons for the change in the Medicare CMI, Carter and Ginsburg (1985) estimate the CMI was 8.4 percent higher in fiscal year 1984 than in calendar year 1981, only 2.8 percentage points of which were attributable to changes in coding practices in response to PPS incentives. This is less change than studies by Johnson and Appel, and Lloyd and Rissing, suggested should occur simply from making coding more accurate.

In Table 10, the overall coding practice changes of 6.2 percent are broken down into two categories, representing the response of coding for PPS as opposed to differences in the quality of the Medicare files used in the comparison. The 1981 MEDPAR file, as indicated earlier, contained data of questionable quality. In contrast, the 1984 Patient Bill (PATBILL) file was constructed from more comprehensive data for DRG assignments including secondary diagnosis codes. The contribution to the increase attributed to the differences between these two files is 3.3 percentage points, the residual difference between other factors explaining the increase and the overall difference.

This study also showed, however, that the CMI on data submitted to Medicare in 1984 was 3 percent lower than the case mix index calculated from data submitted to The Professional Activity Study of CPHA in the
Table 10
DECOMPOSITION OF CMI INCREASE
(In percent)

<table>
<thead>
<tr>
<th>CY1981-FY1984 CMI increase</th>
<th>8.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical practice changes</td>
<td></td>
</tr>
<tr>
<td>Pre-PPS trend</td>
<td>1.4</td>
</tr>
<tr>
<td>PPS-associated shifts to outpatient</td>
<td>0.7</td>
</tr>
<tr>
<td>Setting for lens procedures</td>
<td>0.7</td>
</tr>
<tr>
<td>Other outpatient substitution</td>
<td>0.0</td>
</tr>
<tr>
<td>Total, medical practice changes</td>
<td>2.1</td>
</tr>
<tr>
<td>Older patients</td>
<td>0.0</td>
</tr>
<tr>
<td>Coding practice changes</td>
<td></td>
</tr>
<tr>
<td>PPS-induced</td>
<td>2.8</td>
</tr>
<tr>
<td>MEDPAR/PATBILL/inconsistency</td>
<td>3.3</td>
</tr>
<tr>
<td>Total, coding practice changes</td>
<td>6.2</td>
</tr>
</tbody>
</table>

NOTE: Numbers are multiplicative rather than additive. For example, 1.028 × 1.033 = 1.062.

The same year (Carter and Ginsburg, 1985). This finding may indicate that some concurrent billing codes are still being reported to HCFCA, while the PAS data is submitted later on more complete records. It certainly does not show a response to overreport to Medicare.

Despite the early emphasis on data quality issues, the literature is inconclusive regarding the effects of data on variation in costs not explained by the DRGs and on the potential for manipulating reimbursement through coding. Reporting true complexity could improve the ability of DRGs to predict Medicare cost per case and at least a portion of DRG Creep (Grimaldi, 1981). Others believe that reporting artificial complexity could have a detrimental effect on the predictive power of the DRGs (Finley, 1981; Stern and Epstein, 1985). A study in progress under contract to the Office of the Inspector General of the Department of Health and Human Services will reabstract a national sample of records to identify the rates of inaccuracy in coding ICD-9-CM and the effects of errors. Studies that emphasize coding accuracy
alone, however, are unlikely to yield definitive conclusions regarding DRG variation because coding is entwined with the issue of diagnostic accuracy or "style."
V. VARIATION EXCLUDED FROM DRGS: PRACTICE PATTERNS

INTRODUCTION

Some of the variation in health resource use that is not explained by DRGs can be attributed to variations in physician practice patterns. These variations can take two forms:

- the ways in which physicians strive to accomplish the same goal for the same patient during a hospitalization;
- either the overall goal (e.g., palliation versus cure) or the goal for the particular hospitalization (e.g., staged procedures versus complete treatment).

In the former case, the DRG classification system has not omitted measurement of an important variable. Rather, practice pattern differences unrelated to patient condition and treatment goals were purposely excluded from case mix adjustment and reimbursement formulas. DRGs were based on the costs of "average" practice patterns, thus allowing a range of discretion in physician practice style. In the second case, differences in the goal of a specific admission are recognized primarily by use of surgical procedures in the DRG system (for example, the new DRG for multiple major joint procedures).

For PPS, there are two issues concerning practice pattern variation. First, current studies of DRG performance are confounded by variations in patient management practices (Gertman and Lowenstein, 1984; Smits, Fetter, and McMahon, 1984). Second, some critics question whether hospitals and physicians can distinguish discretionary or inappropriate practices from needed care well enough to effect greater efficiency without sacrificing quality (Stern and Epstein, 1985).

Practice pattern variation confounds DRG performance studies because regional and local variations in appropriate utilization, quality of care, and efficiency have been incorporated in the classification scheme and in the measures of health resource use (costs, length of stay) used to test DRG performance (Frank and Lave, 1985;
Gertman and Lowenstein, 1984; Lave, 1985b; Ommen and Conrad, 1984; Smits, Fetter, and McMahon, 1984; Stern and Epstein, 1985). Issues of unmeasured differences in patient condition and of data quality compound the problem: If patients differ in ways undetected by the case mix adjustment system or imprecisely described by the coding system, then appropriate practice pattern variation may be deemed otherwise.

In an ideal system, DRG performance studies would help to identify practice variations that are unrelated to patient characteristics (Smits, Fetter, and McMahon, 1984) because the classification system would group together patients with identical treatment requirements. The literature, however, yields only studies that accept as given the observed variations in admission practices and use of inpatient services, or cost and length of stay. No studies explicitly distinguish inappropriate variation in resource use from all other unexplained DRG variation.

The second issue thus remains open. Without studies to define standards of inappropriate resource use, it is not possible to distinguish among beneficial, neutral, and harmful cutbacks in service utilization. HCFA has only recently made grant awards to evaluate PPS effects on quality and access to care, and no results are available (U.S. Congress, 1985).

A rich literature documents variations in practice patterns, an issue widely discussed before the onset of PPS. A full review of this literature is outside the scope of this review, but a short list of authors includes Barnes et al. (1985); Chassin et al. (1986); Griffith et al. (1982, 1985); Lembcke (1952); McPherson et al. (1982); Roos (1984); Rosenblatt and Moscovice (1984); Wennberg and Gittelsohn (1973, 1982); and Wennberg et al. (1975).

The studies in this section document substantial practice pattern variations within DRGs. Both local and regional variations in practice patterns are documented. In these research reports, patterns of medical and surgical admissions (Wennberg, McPherson, and Caper, 1984), costs (Horn, Horn and Sharkey, 1984), length of stay (Mitchell et al., 1984, 1985), and other resource utilization (Mitchell et al., 1984, 1985) vary widely. This section is divided into studies of variations in admission practices, charges, length of stay, and use of other resources.
ADMISSION PATTERNS

The implementation of PPS inspired an early focus on hospital admissions for two reasons. First, when PPS was implemented, there was concern that per-case reimbursement created the incentive to increase admissions (Anderson and Steinberg, 1984; Omenn and Conrad, 1984; Stern and Epstein, 1985). Second, it was suspected that DRG performance studies would be confounded by using data that included inappropriate admissions (Gertman and Lowenstein, 1984; Lave, 1985b). Citing Restuccia et al. (1984) and SysteMetrics (1983), Gertman and Lowenstein observe, "Part of the mysterious variation in days of care per 1,000 Medicare beneficiaries across the country may represent variation in inappropriate admissions."

In an examination of hospital admission patterns in Maine, Wennberg, McPherson and Caper (1984) conclude that physicians have so much discretion concerning hospitalization and the assignment of diagnoses that DRGs cannot be expected to be homogenous. Reflecting the concern that PPS would inspire increased admissions, Wennberg also suggests that PPS would be ineffective at controlling overall costs unless hospital admissions were closely monitored.

Wennberg and his colleagues have studied physician practice patterns for many years. Their report examines all nonobstetrical admissions in Maine for the period 1980 through 1982, categorized by DRG, to define the magnitude of variation among 30 different area markets. Admission rates were compared based on standards derived from previous studies by the author (Wennberg, 1983; Wennberg and Gittelsohn, 1982). For example, inguinal hernia repair was selected as the standard to define low variation medical and surgical admissions because the highest rate among Maine areas for inguinal hernia repair is only 1.5 times the rate in the lowest area. The standard for high variation was hysterectomy, the rates for which vary 3.5 times among the areas. The highest variation standard was tonsillectomy, which varies by 12 times in the rates with which it is performed. In some analyses, the individual DRGs were compared for variation. To reduce the number of analytic classes and the effects of variation in coding practices, Wennberg also combined related DRGs based upon similarity of principal diagnosis or MDC, to yield 77 "modified DRGs."
Wennberg found substantial variation in admission rates, using both individual and modified DRGs. Only three DRGs showed low variation, and only 24 had moderate variation. All the rest ranged from high to very high. Of the 46 modified medical DRGs, there were none with low variation, only three with moderate variation, and 43 showed high to very high variation. Of the 31 modified surgical DRGs, 25 showed high to very high variation.

CHARGES

In a study of four university, one community teaching, and one non-teaching hospital, Horn, Horn, and Sharkey (1984) found that identifying individual physicians accounted for 20 to 40 percent of variation in charges after controlling for DRG. The authors cite a followup study (Horn, Horn, and Moses, 1984) comparing physicians' deviations in charges from expected norms by means of DRG-adjusted and SOII-adjusted charges. This analysis produced greatly disparate results in some cases. Horn concludes that further detailed review is required to determine whether the variations are due to efficiency, patient condition or quality of care. She also cautions that identifying inappropriate resource use may depend upon the method of analysis.

LENGTH OF STAY

Chassin (1983) has comprehensively analyzed regional variations in length of stay. These are important to understanding variation within DRGs in national data but are unlikely to be significant at the state or hospital level. Local variations were noted (but not documented) by the Yale team working on development of the I9 DRGs. Differences in length of stay patterns for simple surgical procedures were found that did not depend upon the characteristics of the patients but rather on "the practice patterns of the particular physicians in the particular setting in which that care is provided" (Fetter et al., 1982).

In studies by Mitchell et al. (1984, 1985), average lengths of stay are shorter in Washington within four illustrative DRGs than they are in North Carolina, Michigan, and New Jersey. For example, length of stay for heart failure and shock (DRG 127) is 7.8 days in Washington, but
11.0 in North Carolina, 11.6 in Michigan, and 13.3 in New Jersey. For lens procedures (DRG 39), patients average 3.0 days in Washington, while on average patients in New Jersey stay 4.1 days. North Carolina and Michigan patients average 3.7 days and 3.9 days, respectively.

USE OF OTHER RESOURCES

Findings from the Mitchell et al. studies (1984, 1985) include, in addition to length of stay variations, other evidence of practice pattern variation. The data for these studies include both part A and part B claims on Medicare patients, which permit analysis of physicians' use of discretionary services. In these data it is assumed that the use of physician consultation implies greater use of nonphysician resources as well.

Mitchell and colleagues found the use of consultations was two to three times greater among attending physicians in New Jersey than in North Carolina in 25 high volume DRGs. For example, physicians of patients with heart failure and shock (DRG 127) ordered consultations on 41 percent of cases in New Jersey but only 19 percent in North Carolina. Michigan physicians order 36 percent consultations, while Washington physicians are closer to North Carolina at 22 percent. For this DRG, diagnostic surgery shows a different alignment: Michigan is highest at 12 percent, with New Jersey at 11 percent, North Carolina at 10 percent, and Washington the lowest at 8 percent.

Even within DRG 39 (Lens Procedures), which contains less unexplained variation than most other surgical DRGs, considerable practice pattern variation exists. In New Jersey, ophthalmologists used assistant surgeons for 75 percent of the cases in lens procedures, while North Carolina ophthalmologists almost never used assistant surgeons. Michigan and Washington surgeons used assistants in 28 and 36 percent of the cases, respectively. Nonsurgeon attending physicians in New Jersey made routine visits for 60 percent of cases in New Jersey (in addition to normal surgical followup), while only 5 percent of surgical cases in North Carolina had claims for routine visits.

These studies begin the process of assessing the role of physician practice variation within DRGs. State variations in length of stay documented in the Mitchell studies reflect patterns that are unlikely to
be influenced markedly within an individual facility. It is too early to expect these studies to yield information concerning the separate effects of variations in efficiency of care, goals of treatment, patient condition, and data quality.
VI. CLINICAL AND HOSPITAL RESPONSE TO PPS

INTRODUCTION
This section examines the variety of hospital responses to PPS that have been reported in the literature. The articles reported in this section may shed some light on the perceived problems with PPS in the field. The types of responses reported may also flag areas where problems may arise in the future because of distortions created by these newly adopted strategies.

RESPONSE TO PERCEIVED UNCOMPENSATED VARIATION IN PATIENT CONDITION
In certain clinical specialties, severity measures have already been developed, tested, and accepted. These measures depend on data not currently included in the UHDDS, and their usefulness is narrowly defined within specific clinical problems. Yet their existence indicates that generic severity scales applied across all DRGs may not perform well. For example, trauma severity indexes such as Champion's Trauma Score and Baker's Injury Severity Score were used in Jacobs's (1985) analysis of the trauma DRGs. Other means of subgrouping patients in a DRG are also noted. Munoz, Margolis, and Wise's (1985) transfusion "identifier" helped to clarify where efficiencies (such as curtailing redundant laboratory testing) were possible.

The approach selected by Munoz and Jacobs to characterize a group of patients within DRGs may indicate the need for "severity" refinement to assist clinicians to analyze their efficiency experience with patients. However, the "severity" stratifications are highly specific to particular diagnoses. It may be that, so long as reimbursement to the hospital is fair, appropriate communication within a hospital would be better served by ad hoc "severity" adjustments at the hospital level than by an adjustment across DRGs. If adjustments for fair reimbursement are required, particular adjustments to particular DRGs may be more accurate, but also more administratively cumbersome, than applying uniform severity adjustments across all diagnoses.
Management Use of Severity Indicators

Experience under PPS is also being analyzed using alternative case mix adjustment systems. Although some hospitals find that formal severity measures are necessary for effective communication with physicians, their widespread use has been discouraged by some experts. In a review of formal severity adjustment measures, Nathanson (1985b) cites three experts who question the utility of these measures within the typical acute care hospital. One of the experts cited by Nathanson, Harold Dickson of Baptist Memorial Hospital in Tennessee, indicates that no more than 10 DRGs are so variable as to require adjustment, while Gonnella observes that there is no evidence that they improve patient management. Gertman adds that their cost does not justify the limited yield of improved information, nor are they able to detect large differences among physicians.

Nevertheless, formal severity measures are being leased and sold and may be put to creative use. At the New England Medical Center, SOII analysis is combined with clinical standards devised to specify the expected volume and type of tests associated with a given DRG and severity level (Nathanson, 1984). With these analyses, the medical center has found that the majority of their surgical cases are in fact simple. Comparison of actual with expected resource use has yielded savings of 20 percent in surgical DRGs. Patients who have been identified as less severely ill are moved out of intensive care earlier and discharged earlier.

Discrimination Against Certain Classes of Patient

Several observers have identified classes of Medicare beneficiaries that could be discriminated against if hospitals develop selective admissions policies. These patients include alcoholics or patients with no home support (Ommen and Conrad, 1984); patients likely to incur higher costs based on general medical, socioeconomic, or demographic characteristics; or the practitioners associated with these admissions (Stern and Epstein, 1985). Other vulnerable subgroups of Medicare beneficiaries include the poor or infirm elderly (Dans, Weiner, and Otter, 1985) and the frail elderly (Berenson and Pawlson, 1984).
General characteristics of patients likely to become outliers could be examined (Smits, Fetter, and McMahon, 1984). Patient severity has often been suggested as a patient attribute that puts the hospital at financial risk.

The industry literature documents only one recommendation of this nature; hospital actions are instead directed at admissions that may not be warranted. In a publication directed to hospital administrators, Meadors and Wilson (1985) recommend using preadmission testing to identify cases "too complicated" for the institution and selectively encouraging admissions of the "young old" (patients aged 65 to 70) because they are less likely to have severe illnesses. Hospitals, in turn, report intensifying utilization review programs to control "unjustified" admissions for which insurers might later refuse payment (Gillock and Smith, 1985; Wallace, 1985).

Per case reimbursement has directed increasing attention to the relationship between clinical severity and the use of services that could lead to more appropriate patterns of utilization. The success of PPS will depend upon the extent to which this alternative occurs rather than inappropriate manipulation of the system or the patients included in its charge.

CHANGE IN PRACTICE PATTERNS TO INCREASE EFFICIENCY

Hospitals attempting to encourage efficiency must influence the practice patterns of physicians who admit patients. The literature documents a variety of methods for improving communication between hospital managers and physicians. Omenn and Conrad (1984) believe that PPS can encourage positive responses from physicians, citing "favorable precedents that indicate the potential power of good information among physicians," including the work of Wennberg and Gittelsohn (1982) in reducing excessive surgical and admission rates in Maine.

Developing Practice Standards

The hospital industry literature reflects numerous strategies for influencing physician behavior. These include:
• Developing standards of care within DRGs, including standards for admission, diagnostic evaluation, treatment, monitoring response, and criteria for diagnosis and treatment of complications (Clifford and Plomann, 1985; Nathanson, 1984).

• Developing specific guidelines for same day surgery and for length of stay at different levels of care (Nathanson, 1985a).

• Instituting retrospective review, apparently without explicit standards (Gillock and Smith, 1985; Richman, 1985; Wallace, 1985).

• Increasing utilization review activities. In one hospital, teams of physicians review cases approaching outlier status. In this hospital, no comparative reports of physicians' performance are distributed to physicians (Wallace, 1985).

• Grand rounds or other group meetings on the subject of treatment profiles (Wallace, 1985) or cost containment (Richman, 1985).

These cost containment strategies are apparently being followed, as length of stay statistics recently indicate. In a recent discussion of the results of the 1984 Arthur Anderson and Co./American College of Hospital Administration Survey, Traska (1985) noted that changes being instituted in more cost-effective care for Medicare patients were being applied to other patients as well, thereby producing larger changes in admission and length of stay statistics for the nation than would otherwise be possible.

Some physicians may be concerned about the risks they are assuming under shortened lengths of stay, however. One member of a "panel discussion" in Healthcare Financial Management (1985) attributes an increased demand for HMO and PPO information to physician recognition that "they are being asked by hospitals to assume increased medical risk (for example, through earlier discharges) without any increase in compensation." Some HMO prepaid plan models allow both hospitals and physicians to share risks and rewards.
Many of these examples reflect appropriate cost-savings strategies that are likely to preserve or improve patient care. However, we can assume that the literature reflects only a small proportion of the methods by which hospitals are influencing physicians, and concern that needed care will also be cut is still warranted. The dissemination of cost-savings strategies in the literature may encourage their more widespread adoption. However, within generalized guidelines and cost-savings strategies, there is still considerable room for variation in practice style.

Other Methods for Influencing Physician Behavior

In contrast to the numerous articles on influencing physicians to contain costs, only one article recommends influencing physicians to maximize hospital reimbursement in a publication directed to hospital administrators. A preadmission program, according to Meadors and Wilson (1985), would allow "medical review personnel time to identify medical cases that could be scheduled for minor elective surgery concurrent with their medical admission, allowing more favorable reimbursement under DRG 468 (discharges with operating room procedure unrelated to a principal diagnosis)." It seems barely plausible that such a strategy could be proposed to physicians without jeopardizing their continued relationship with the hospital.

Indirect methods for influencing physician behavior are being recommended. Meadors and Wilson (1985) suggest that hospitals should offer cooperative physicians the best operating room schedules and preferred rooms for their patients. Staff recruitment is another indirect method, although Stern and Epstein (1985) point out that recruitment within a given specialty is likely to increase the numbers of patients within both profitable and unprofitable DRGs at the same time (for example, in cardiology).

In discussing the severity issue, we cited a number of studies that suggested DRGs are not sensitive to certain kinds of severity. For the most part, these studies do not make necessary adjustments in reimbursement calculations and thereby fail to demonstrate true underpayment. In this case, however, the perception that drives these
studies may be more important than the validity of the results. To the extent that hospitals perceive consistent underpayment for some DRGs, some patients, or some hospitals, hospitals may attempt to manipulate the patient classification, admissions, or even therapies to maximize reimbursement.
VII. EFFECTS OF WEIGHTS AND PAYMENTS

INTRODUCTION

Medicare reimbursement depends not only on the classification system, but also on the DRG weighting structure and reimbursement formula determining payment. For PPS, the fundamental payment issue is the same as that resulting from unexplained variation in the DRGs: Hospitals may be over- or underpaid for Medicare services.

Designers of PPS intended to distribute Medicare reimbursement equitably among hospitals based on the resource requirements of patients. Medicare reimbursement is based on several factors in addition to DRG case mix adjustment (for example, teaching, urban/rural location, and local wage adjustments). Payments to individual hospitals or classes of hospitals may thus be inappropriate because of features of the DRG classification system (as reviewed in preceding sections) or because of any one or the combination of other adjustments made in the reimbursement formula. Conversely, the reimbursement formula as a whole may compensate for deficits in the DRGs or other adjustment factors. The system as designed is intricate.

Two features of that designed system have attracted special attention in the literature: the possibility that DRG weights are "compressed," and the redistribution of payments associated with moving to national payment rates. This overview reviews discussions aimed at PPS as it was designed. Recent revisions in DRG logic and weighting structure, as well as other recent regulatory and legislative actions, were designed to address some of these problems, which may, in some cases, no longer be pertinent.

COMPRESSION IN THE DRG WEIGHTS

Beginning with Pettengill and Vertrees (1982), it was acknowledged that the relative price weights for the DRGs were likely to be compressed. Compression means that the highest weighted DRGs (more costly, complicated cases) are underpaid relative to the rest of the DRGs, while the lowest weighted DRGs (less costly, simple cases) are
overpaid relative to the rest of the DRGs. The effect of compression is that hospitals treating proportionately more of the highly weighted DRGs are likely to be reimbursed inadequately, and hospitals treating proportionately more of the DRGs with low weights are likely to experience windfalls.

The problem is exacerbated if the compression is nonlinear. Nonlinear compression affects hospitals increasingly as they deviate from the "average" case mix index. For example, a hospital with a CMI 10 percent above average could have costs 12 percent above average, while a hospital with a CMI 20 percent above average may have costs 26 percent greater than average. This means that even if payment rates are adjusted so as to correct for both the average hospital and for the hospital with a CMI 10 percent above average, they will still be too low for a hospital with a CMI 20 percent above average.

This section reviews three studies. The first describes derivation of the original DRG weights, the second examines whether those weights were compressed, and the third describes derivation of the weights used for recalibration in October 1985.

Derivation of the Original DRG Weights

Pettengill and Vertrees (1982) report the methodology for weighting the DRG classification system and for testing its reliability and validity. After selecting the DRG classification system as the case mix measurement device, they estimated a relative cost weight for each DRG and then derived the hospital case mix indexes. The CMI was tested for its ability to predict average Medicare cost per case, controlling for local area wages, intensity of teaching activity (number of residents per bed), hospital bed size, and city size (small, medium, and large). The authors then examined the predictions for many possible sources of bias. This discussion concentrates on features of Pettengill and Vertrees's methodology that relate specifically to the compression issue, which involved testing for possible inappropriate aggregation of the weights.

In deriving the relative weights, Pettengill and Vertrees analyzed the hospital's cost of each case in three steps. First, the hospital's allowable per diem costs for routine and special care days were
multiplied by the patient's length of stay in each unit to give the
total room and care costs. Second, ancillary charges for the case were
converted to costs by multiplying departmental charges by department
ratios of cost to charges. Third, the total cost for the case was found
by adding ancillary costs to room and care costs. Hospital costs were
then adjusted to eliminate the effects of indirect teaching costs and
local wages.

Pettengill and Vertrees derived the weights by first assigning each
case in each hospital to a DRG. Next, outliers were excluded, and all
the hospital's cases in each DRG were divided by the number of cases in
that DRG to derive the hospitals' average cost of caring for a patient
in the DRG. Third, the average costs of each DRG in each hospital were
added and divided by the number of hospitals to determine the national
average cost of caring for patients in each DRG. (Because the values
are not weighted by the number of cases in the hospital, the average is
a hospital average and not a per case average.) The national mean costs
for each DRG were divided by the average cost over all DRGs to derive
the relative weights, often referred to as the HCFA weights.

Pettengill and Vertrees examined two sets of correlations to search
for possible bias introduced by assuming the relative costs were similar
across all hospital types. The first compared the national DRG weights
and case mix indexes with weights and indexes calculated for hospitals
grouped by bed size and urban/rural locations. Correlation coefficients
of national versus hospital weights were lowest for small urban (.87)
and small rural (.91) hospitals. The correlations of rural hospital
weights with national weights increased with increasing bed size: The
correlation coefficient for hospitals with 100 to 169 beds was .95, and
that for larger rural hospitals was .97. The highest correlations among
urban hospitals occurred for bed size between 100 and 404 (.99) and 405
and 684 (.98). Hospitals larger than this had coefficients of .97. The
small number of cases included for small hospitals may mean that the
effect of cost structure cannot be tested by these methods. The second
set of correlations compared regional with national weights and indexes.
The lowest regional with national weight comparison yielded a
correlation coefficient of .97. All coefficients between national and
regional case mix indexes were .98 and .99.
Having shown that the case mix index was proportional to the average cost per Medicare case, providing reasonable evidence of its validity, Pettengill and Vertrees tested it against known sources of error. Random error in DRG assignment would cause errors in either the proportions of various DRGs in the hospitals or the calculated weights (or both). In simulations of different degrees of error, Pettengill and Vertrees showed that increasing amounts of error caused compression in the weights, in the proportions, and in both simultaneously. To the extent that error can be shown to exist, then, we can expect "the relative costliness of the hospital's case-mix is overstated for hospitals with low values and understated for hospitals with high values."

**Are the Weights Compressed?**

Pettengill and Vertrees' original observation has been recently investigated by Lave (1985a). The purpose of Lave's study was to determine if compression exists and, if so, to what extent. Lave believes that compression is probable for three reasons:

* The calculations of costs per DRG were based on the assumption that routine and special care unit costs were the same over all DRGs, yet it is unlikely that nursing intensity within either level of care is the same for complex DRGs as for simple DRGs.
* Anecdotal evidence that hospitals cross-subsidize within ancillary services.
* The MEDPAR file contains data that omit specific secondary diagnoses, error and lack of specific secondary diagnoses in the MEDPAR file, and coding errors.

Lave compared four sets of DRG weights: the 1981 MEDPAR price weights, 1979-81 Maryland Medicare price weights calculated in the Coffey and Goldfarb study (1984), and New Jersey price weights calculated by the author from New Jersey adjusted cost data separately for teaching and nonteaching hospitals. Both the Maryland and New Jersey costs were thought to be based on diagnostic and procedural data
reflecting fewer coding errors and better cost information than the MEDPAR data. New Jersey’s data was expected to be best because considerable precision in costing was available (although this assumption was not tested). The price weights were standardized so that DRG 135 had a weight of 1.00 in all four sets.

Coefficients of variation calculated on each set of normalized relative prices showed that HCFA weights had the lowest variation (67 percent), followed by Maryland (80 percent), New Jersey nonteaching (90 percent), and New Jersey teaching hospitals (95 percent). Regressions of HCFA normalized prices against each of the other prices yielded coefficients of less than 1 in all cases, ranging from .69 in New Jersey nonteaching hospitals through .83 in Maryland Medicare patients. These results confirm the compression of HCFA weights relative to the comparative datasets.

Lave believes that, of the sources of the compression problem, diagnostic and cost data problems will be corrected by the incentives of PPS so that recalibration will produce less compressed weights. She recommends that HCFA study the degree of bias introduced because of the special and routine care assumptions, and refine the algorithm for costing if the bias is significant. She also suggests alternative approaches for calculating price weights:

- Determining prices for appropriate treatments as recommended by an expert panel, instead of empirically. This is similar to the strategy employed by Young (1985) in PMCs.
- Determining relative prices based on charges as reported to HCFA rather than using the existing costing algorithm. The advantage of this method is that recent data may be used. Cost report data is available only two to three years after it has been reported. This method was investigated by Cotterill, Bobula, and Connerton (1985), reported below. The weights were recalibrated for FY1986 using this method.
- Determining relative prices by intensive microcosting studies in selected hospitals. This is the most costly but likely to be the most accurate of the methods proposed.
Changing policy so that relative DRG payments may vary with the cost experience of hospitals. This approach is recommended by the American Hospital Association as a method to compensate for the heterogeneity of DRGs and the presumed likelihood that hospitals selectively treat patients of greater or lesser severity.

To evaluate Lave's study, one has to consider other possible explanations for the results besides compressed HCFA weights. For example, her comparison datasets were from two eastern states. If the regional differences in length of stay (Chassin, 1983) are the result of selective differences within certain DRGs, the eastern datasets would contain a wider spread of costs based on longer lengths of stay alone than the national average costs. In addition, the national relative price weights are normalized at the hospital level, while the price weights calculated by Coffey and Goldfarb (1984) for the Maryland data are normalized at the case level (Frank and Lave, 1985). Hospital level normalization reduces the amount of variation (Kominsky et al., 1984). Further, the New Jersey all-payor weights may reflect differences between Medicare and non-Medicare patients. These factors must temper Lave's conclusions that HCFA relative price weights for the DRGs are inappropriately compressed.

Finally, Lave does not address a source of compression that is clearly present and cannot be accurately corrected. Because practice patterns are not uniform, weights are not uniform across hospitals, cities, states, and regions, a point illustrated by the very imperfect correlations between the weights she examines. When imperfectly correlated numbers are aggregated, they regress toward the mean--that is, their values become more alike or compressed. This is a necessary consequence of using national weights in a country where practice is not homogeneous.
Derivation and Testing of New Relative Weights

The use of historical cost data to derive the relative weights is perceived as one of the potential reasons for discrepancy between actual costs and reimbursement. The relative weights could be recalibrated in a more timely manner if Medicare charges were used for the calculations instead of derived costs. Typically, cost report data are two to three years old before they become available for use in weight calculations, and the time and resources required for deriving costs from charges create additional delay.

Cotterill, Bobula, and Connerton (1985) investigated the correspondence between weights based on costs and those based on charges to determine if the charges could replace costs as the basis for weight calculation. They derived charge based weights from the 1981 MEDPAR file following the methods related in Pettengill and Vertrees (1982) as closely as possible, with two exceptions. First, they did not remove capital and medical education costs (both direct and indirect) from the charge weights. (Cotterill reported that a special analysis eliminating these factors showed very high correlation with charges but did not report the coefficient.) Second, they did not use supplementary data to increase sample sizes for small volume DRGs; hence, they restricted the comparison to 358 DRGs.

Cost and charge data yield very similar estimates of the relative weights; 45 percent of the DRGs (representing 36.6 percent of the cases) increase no more than 5 percent in weight, while 38 percent of the DRGs (53 percent of the cases) decrease no more than 5 percent. Both Spearman (rank) and Pearson (value) correlation coefficients were greater than .99, indicating a high degree of correspondence between charge- and cost-based weights. Coefficients of variation within individual DRGs show somewhat greater variation in the charge weights, perhaps because of variations in markup among hospitals. Finally, there is less compression in the charge weights than in the cost weights, as indicated by a slightly larger standard deviation for the charge weights. These result in higher CMIs for large, urban, and teaching hospitals, and lower CMIs for small, rural, and nonteaching hospitals.
Other analyses gave support to the cross-subsidization theory as a source of compression, but patterns emerged that were different from the expected direction of subsidization effects. High cost services such as special care units have lower charge-to-cost ratios than those for ancillary services such as pharmacy and laboratory, as expected. But Cotterill and colleagues found that DRGs with high proportions of special care charges also tend to have high proportions of ancillary charges. The contribution of ancillary to total charges is substantial enough that the overall effect is that charge-based weights have less compression than cost-based weights.

Tests of the CMI based on charges indicate that it was linearly related to the Medicare cost per case. The case mix coefficient and the teaching coefficients are slightly lower than those in the cost equation but are not significantly different at conventional levels. The regression coefficients are presented in Table 1 of this review. Based on these results, HCFA selected the charge-weight method for FY1986 recalibration.

An independent analysis by the Prospective Payment Assessment Commission--ProPAC--staff also indicated that the use of charge data would have little effect on the DRG weights (1985c). In a summary of weight changes after recalibration, they noted:

- Over all DRGs, the greatest weight changes occurred for DRGs that required supplementary data to achieve sufficient cases for the original weight calculations.
- Weight changes range from a 22.8 percent increase for DRG 410 (Chemotherapy) to a 55.4 percent decrease for DRG 125 (Circulatory disorders excluding acute myocardial infarction, with cardiac catheterization, without complex diagnosis). (This observation excludes the supplementary data DRGs above and those with significant logic adjustments.)
- The degree of change within the surgical DRGs differed from that for the medical DRGs: the average weight for surgical DRGs increased 0.5 percent, and the average weight for medical DRGs decreased 3.4 percent. (This observation also excludes the DRG groups mentioned above.)
The ProPAC report speculates that the differential change between medical and surgical DRG weights could be due to improved coding or shifts to the outpatient setting, and plan to study this finding further.

For fiscal year 1987, ProPAC recommends another recalibration, using charge data alone (ProPAC, 1985a). A schedule of annual recalibration was deemed optimal for keeping pace with practice pattern changes while enabling hospitals to make fairly stable financial forecasts. The Commission will also study alternatives to historical cost and charge data for determining relative weights.

NATIONAL RATES

A potentially more serious concern has to do with the effect of national rates. Hospitals currently doing well, or breaking even, on a DRG are aware that today's strategies may not be sufficient under national rates.

Their concerns are also expressed by other analysts of the prospective payment system (Lave, 1984; Vladeck, 1984a). As Lave observes, "The effect of a national rate is to reallocate Medicare payments from hospitals that have relatively high costs to those that are relatively low cost. The majority of the saving from prospective payment come from the overall limit on the rate of growth of the average payment rate, not from the establishment of a national rate."

Vladeck (1984a) adds:

Movement to uniform national rates produces no net savings to the trust fund whatsoever. For every hospital or group of hospitals that is severely and unfairly penalized by the inherent arbitrariness of a single national standard, there is a symmetrical hospital or group of hospitals that receives an unmerited windfall. A uniform national standard of efficient and effective production of care is certainly needed in the determination of Medicare payment rates, but to make that standard the sole basis for the rates, in light of the enormous variations in cost patterns from one part of the country to another, reflects a preference for abstract principle over simple equity or even common sense.
In an analysis of the effects of the transition to national rates, ProPAC (1985b) notes that redistribution of PPS payments regionally and across hospital groups could occur, as well as an increase in PPS payments overall. Specifically, urban hospitals and those in New England would gain under the national rates, while rural hospitals and hospitals in the East North Central and several Western census regions would lose under national rates. ProPAC also reports that average payments for each case are 4.3 percent higher based on national rates than on hospital-specific rates. The difference is probably due to recent data used to calculate national rates, while hospital-specific rates depended upon data from 1981.
VIII. EFFECTS OF PPS: TWO EXEMPLARY POLICY ISSUES

INTRODUCTION

In this section, we review literature that illustrates the combined effects of DRG performance, possible sources of variation, and the DRG weighting structure on two policy issues. The first example concerns teaching and public hospitals, where the issue of case mix adjustment is complicated by other adjustments in the PPS reimbursement formula. The second issue concerns change in health care technology and practice as related to PPS, where concerns have been expressed regarding both the intended and unintended changes prompted by PPS, as well as the ability of PPS to respond to the rapidly changing medical field.

TEACHING AND PUBLIC HOSPITALS

For teaching and public hospitals, two issues stand out:

- The major issue is whether treatment needs unmeasured by the DRGs result in systematic underpayments to these hospitals. In the belief that DRGs do understate the needs of patients served by these hospitals, Congress set the indirect teaching adjustment at double the statistical estimate of the indirect costs of medical education programs. (The estimate was derived in an equation that included other variables benefitting teaching hospitals. When these variables were not included in the PPS reimbursement formula, doubling the indirect teaching adjustment also offset their exclusion.)

- The second issue is the relative contribution of such factors as the costs of education, unmeasured need for treatment, specialized technology, and efficiency to the higher costs of teaching hospitals.

Resolving these issues requires studies that have not been performed. As indicated earlier, no studies in the literature document definitively whether the distribution of patients' treatment needs are
unmeasured by DRGs. Adequate assessment of both the first and second
issues requires this information. The second issue is complicated by
the need for still more information. For example, the costs of teaching
programs were statistically estimated because no direct measure exists.
As a result, the current estimate may be confounded with factors such as
inefficient practice patterns, use of expensive technology, hospital
size, or urban location, some of which cannot be factored out.

Section III of this review examined several studies by Coffey and
others that shed considerable light on the differences between teaching
and nonteaching hospitals as part of analyzing the differences between
case mix methods. In this section we review several studies that
compare the costs and case mix distributions in teaching and public
hospitals, using DRGs to control for case mix differences between these
hospitals and their comparison groups. We then review a single study of
the effect of PPS on teaching and municipal hospitals. Finally, we note
commentary on the design of the indirect medical education adjustment in
the Prospective Payment System.

Studies of Cost and Case Mix Differences

Even before the introduction of PPS, expensive hospitals frequently
asserted that their higher costs were the inevitable result of their
complex case mix. As Simborg (1981) has noted, "hospitals whose costs
exceeded those allowed began to claim that they had a costlier case mix.
These hospitals tended to be teaching hospitals, which have
characteristics other than complex case mix that could account for
increased costs." Now the argument has shifted to whether the DRGs
capture enough of case mix complexity, yet Simborg's observation is
still true. The studies described below begin to discriminate between
case mix complexity and other factors associated with the costs of
teaching and public institutions.

Several recent studies have used the DRG classification system to
elucidate differences in case mix between teaching and nonteaching
settings, and to clarify other differences which are not measured in
case mix but contribute to higher costs of teaching institutions
(Cameron, 1985; Frick, Martin, and Schwartz, 1985; Garber, Fuchs, and
Silverman, 1984; Jones, 1985). Public hospitals' case mix have also
been described (Shwartz, Merrill, and Blake, 1984).
Of these studies, Garber, Fuchs, and Silverman (1984) are unique in analyzing outcomes as well as case mix and remaining cost differences between teaching and nonteaching patients in the same major university hospital. Garber and colleagues studied 12 ICD-9-CM DRGs that met two criteria: at least 20 admissions to each of the teaching and nonteaching services, and 10 or more deaths per DRG. In 1981, these DRGs accounted for 16.2 percent of all admissions and 29.5 percent of all costs of patients 45 years and older. There were 1,007 teaching patients and 1,018 nonteaching patients. Charges rather than costs were used as the dependent variable after determining that charges adjusted by cost-to-charge ratios yielded less than 1 percent difference in the results. For some analyses, patients were stratified into groups depending upon the predicted probability of death.

Garber presented results in several different ways. Before adjusting for case mix using DRGs, charges were 59.6 percent higher for teaching than nonteaching patients. Adjusting for case mix reduced the difference to 10.8 percent. Eliminating outliers and controlling for sociodemographic characteristics of the patients had little effect. Because length of stay was equal between the teaching and private services, cost differences were attributed to different cost per day. The smallest difference in costs between the two services was for patients with the lowest probability of death; the largest difference—up to 70 percent higher after case mix adjustment—occurred for patients with the highest predicted probability of death.

At time of discharge, the death rates among the groups were different. Death rates on the nonteaching service were 34 percent higher than those on the teaching service, but the relationship between costs and subsequent mortality was variable across the DRGs. For five DRGs, there was no discernible cost difference but a substantial difference in mortality. Three DRGs showed differences in both mortality rates and cost. A third group of four DRGs had large cost differences that were not reflected in mortality rate differences.

The authors matched 51 pairs of high risk patients from the two services at admission and followed them to nine months after discharge. The average inpatient costs of care for these patients was twice as high
on the teaching as on the nonteaching service, with half the death rate. However, by nine months following discharge, the death rates for the 48 pairs who could be followed were equivalent, with less than 20 percent of the patients alive.

Garber and colleagues conclude that cost differences between the teaching and nonteaching services may be explained by differences among physician practice patterns and systematic differences unaccounted for by the case mix adjustment method. Patient preferences for or against heroic measures may also account for some of the difference. This study contributes a dimension that is lacking in all other studies of costs of care under DRGs; yet, as the authors indicate, neither the quality of life nor costs of subsequent care for the survivors after discharge was studied. There is not yet sufficient information to answer the difficult question of whether the intensive treatment was worth its cost.

In this study, case mix adjustment by DRGs lowered the difference between teaching and nonteaching service patients from 60 to 11 percent. In patients with the highest risk of death, however, the difference remained high, even after case mix adjustment. This suggests differences in practice patterns related more to differences in goals than to treatment needs.

Three small studies used 18 DRGs to describe differences between teaching and nonteaching case mix (Frick, Martin, and Shwartz, 1985; Jones, 1985) or between public and private hospitals (Shwartz, Merrill, and Blake, 1984). The combined effects of use of 18 DRGs, small sample sizes, and the descriptive nature of the work limits their usefulness here. Yet each contributes some information regarding the likely composition of case mix in these institutions.

Jones (1985) conducted a smaller study at the same hospital as Garber. Differences among the services were examined, controlling for case mix, patient demographics, and other measures labeled "severity." Four surgical 18 DRGs were studied after homogeneity was increased by omitting patients with certain principal diagnoses. For example, patients with simple breast biopsies and no further treatment were removed from DRG 23: Cancer of the Breast with Surgery. Among the variables abstracted from the medical record were: health status at
admission (the APACHE chronic health indicator, modified by the author), number of complications, number of medical and nonmedical consultations, and type of anesthesia. The dependent variable was patient charges, which included a $16 per day surcharge for patients on the teaching service. Regressions were performed across all patients, using DRGs and other independent variables as categorical variables. In the model consisting only of case mix measured by DRGs and teaching services, total charges for the faculty group were 23 percent higher than charges for nonteaching patients. A separate case mix effect was not determined for this model. Teaching status was less important at predicting total charges than DRG, type of anesthesia, and number of medical and nonmedical consultations (which is not surprising because this measure is also a billed service that contributes directly as well as indirectly to the bill). Health status at admission was not an important predictor in total charges, although it did predict diagnostic and routine charges.

Frick, Martin, and Shwartz (1985) provide insight into the case mix differences between teaching and nonteaching hospitals. Eleven teaching hospitals were compared as to case mix and cost per case, using the 18 DRGs. Comparisons of the 30 most frequent DRGs show the surgical DRGs prominent in the teaching, but not in the nonteaching institutions, a finding similar to those of Coffey and colleagues. Yet there was also a high degree of overlap: Both hospital types had large proportions of routine deliveries, gynecologic surgery, and mental disorders. There were also many DRGs with small numbers of patients common to both types of institutions. Differences in costs (including an estimate of nursing intensity for each DRG) showed higher costs per patient within the same DRG for teaching institutions, as well as a higher proportion of patients in the more costly DRGs.

Shwartz, Merrill, and Blake (1984) report comparable differences and similarities in case mix descriptions between public and nonpublic hospitals in New York City.

Cameron (1985) measured indirect costs of hospital teaching programs, using Medicaid claims data from 1978-79 fiscal year from nearly all the California hospitals. The five Los Angeles County hospitals were excluded because their per diem rates do not permit
adequate cost analysis. Case mix adjustment was accomplished using the I8 DRGs. Physician services, including both separately billed and hospital compensated portions, were controlled, enabling separate analysis of cost savings due to substituting residents for fee-for-service physicians in teaching programs. Hospitals were identified as university, major teaching, minor teaching, and nonteaching.

Table 11 shows the percentage by which costs per case for the hospital groups exceeded those for nonteaching hospitals, before and after adjusting for case mix and full physician costs. The case mix effect is largest within the university hospitals, and controlling for physician costs always reduces the difference further. Because the majority of MediCal patients were admitted for uncomplicated DRGs, the case mix effect might be higher in a Medicare population. Moreover, because MediCal is rarely billed separately by physicians treating patients in county hospitals, the physician effect might be higher in studies with a smaller proportion of public hospitals. The effect of using I8 DRGs rather than ICD-9-CM DRGs is not known. Cameron's adjustment for physician costs, however, helps to clarify cost savings within teaching programs.

Review of PPS Impact

Schwartz, Newhouse, and Williams (1985) examine teaching function and public control simultaneously. Their study examines the effects of PPS, as well as those of other cost containment strategies, in a coherent analysis of the effect related to the traditional sources of revenue for these institutions.

Schwartz and colleagues assert that the higher costs of teaching hospitals in general--estimated by Sloan et al. at 13 percent, and by Cameron (1985) at 14 percent--are probably somewhat lower than these figures for two reasons. First, the case mix adjustments made in these studies are unlikely to account for the entire complexity and costliness of teaching hospitals' patients. Second, the housestaff services are not billed, thus contributing to lower medical costs (if not to lower hospital costs) than the same services provided in nonteaching hospitals by private physicians. At present, they argue, PPS has acknowledged the costs of teaching hospitals amply, by exempting housestaff salaries from
Table 11

PERCENT DIFFERENCES ABOVE NONTEACHING HOSPITALS
IN DIRECT PATIENT CARE COSTS, ADJUSTING FOR
CASE MIX AND PHYSICIAN COSTS
(387 nonteaching hospitals)

<table>
<thead>
<tr>
<th></th>
<th>University (n=8)</th>
<th>Major (n=15)</th>
<th>Minor (n=51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before case mix adjustment</td>
<td>108</td>
<td>36</td>
<td>20</td>
</tr>
<tr>
<td>After case mix adjustment</td>
<td>33</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>After case mix adjustment full physician costs</td>
<td>26</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

NOTE: Reference group is nonteaching hospitals (n=387).

DRG limits and through the indirect medical education adjustment. Teaching hospitals as a group, then, are not currently more vulnerable than their nonteaching counterparts.

But public hospitals that are also the center of medical schools' teaching and research activities (the municipal "flagship" hospitals) are at special risk because of the combined effect of PPS and other cost containment schemes. Schwartz and colleagues project that complex patient loads will increase as other hospitals develop active transfer or other cost-saving patient selection policies, and as increasingly strict eligibility requirements are applied to other health care financing programs. In addition, these hospitals already carry a disproportionate share of bad debt and charity care. Flagship municipal hospitals carry almost 22 percent of their total revenue as combined "free" care. Other municipal teaching hospitals have combined bad debt and charity care of 12 percent. (By contrast, nonteaching, nonmunicipal hospitals have 3.6 percent.) Further, the proportion of Medicare patients in these hospitals is low compared with that of other
hospitals, so that the benefits of Medicare reimbursement policies cannot have a substantial protective effect. For example, Medicare patients account for 18 percent of the flagship municipal hospital revenue, 21 percent of revenue for other teaching municipal hospitals, and 34 percent for nonteaching, nonmunicipal hospitals. Finally, the other sources of revenue for public flagship hospitals are unlikely to be good sources of increasing revenue to cover losses. Neither state and local taxes (which represent 23 percent of total revenue for these hospitals) nor third party insurance patients (already low proportionately) are likely to increase.

The Schwartz and colleagues study provides insight into the hospitals where both major teaching and public care commitments have been made. The authors suggest that the teaching adjustment is adequate at present, but the problem of uncompensated care remains.

**Commentary on the Indirect Medical Education Adjustment**

As Schwartz and colleagues indicated, the amounts paid to teaching hospitals under the indirect medical education adjustment are apparently currently adequate; however, Lave (1985b) notes that the manner in which the size of the adjustment was derived is flawed. Consequently, payments for "indirect medical education" include explicit and implicit compensation for other factors omitted from the design of PPS. Because payments are greater than the statistical estimate of indirect teaching costs, the adjustment is likely to be perceived as overcompensation, making it vulnerable to cost-trimming efforts. Lave argues that the size of the adjustment should not be cut unless simultaneous changes are made in other factors making up the PPS formula.

According to Lave, the estimate could be biased for several reasons: error in the 1981 MEDPAR file, compression in the DRG cost weights, unmeasured need for treatment in the DRGs, incomplete specification in the PPS formula of wages and other factors that vary by location, and incomplete specification of the indirect teaching factor itself. These inaccuracies cause bias because estimating the effect of the indirect costs of teaching programs on Medicare operating cost per case depends in part on which other independent variables are included in the equation and how well they are measured. (An example of this
principle in operation occurs with the indirect medical education estimate. The estimate of 5.79 was derived in an equation that included bed size and SMSA size; excluding these variables and including the simple urban/rural adjustment increases the estimated relationship between costs and each .1 increase in residents per bed to 9 percent.)

Lave proposes several steps to improve the estimate and PPS, many of which HCFA has already done: recalculating case mix indexes with improved data, modifying the structure of the DRGs, improving the method for calculating relative cost weights, accounting for size of urban areas in rate setting, and studying the influence of other market factors in addition to wages.

As predicted by Lave, proposed changes in the medical education adjustment are perceived as simple cost-cutting mechanisms when unaccompanied by other adjustments in PPS. HCFA's proposal to exclude interns and residents delivering outpatient care from the hospitals' count that determines the resident-to-bed ratio has brought strong criticism from ProPAC (1985c). The Commission charges that this modification undercuts the intended purpose of the indirect medical education adjustment.

With each new adjustment to the indirect medical education adjustment, the relationship between its purpose and its reality becomes increasingly questionable. It is unlikely that the adjustment currently reflects the "true" indirect costs of medical education, and it is possible that it does not restore equity to hospitals disadvantaged by other aspects of the PPS formula (such as the urban/rural adjustment). Debate over the purpose and appropriateness of change will no doubt continue until the adjustment is explicitly redesigned.

RESPONSIVENESS TO CHANGE IN TECHNOLOGY

Several authors have suggested that DRGs are unresponsive to technological innovation and may stifle the introduction of new technology. Four issues are of concern: weight recalibration, coding, the DRG structure, and the incentives of per-case reimbursement. Some authors have also suggested that the exemption of capital costs from the DRGs (a feature of PPS) could produce inappropriate introduction of new technology.
Weight Recalibration

The recalibration of weights for DRGs is the primary way of recognizing changes in technology that make individual conditions less or more expensive to treat. Recalibration has been criticized because it will occur infrequently, using old data. As Wirtshafter (1984) noted, "the built-in four-year administrative lag period" almost guarantees "a fossilized nomenclature for clinical medicine." The recent changes in DRG logic and weight recalibration, as well as announced changes in the strategies for updating and revising, render some of the complaints obsolete. Weight recalibration was viewed as especially unwieldy because of the assumption that cost data would be the basis of calculations. Now, more recent data may be used, thus making recalibration more closely congruent with current practice.

Technologic development is the issue that has been most often mentioned in conjunction with concerns about PPS' response to change (Davis, Anderson, and Steinberg, 1984; Iglehart, 1982; Lave, 1984; Saksena, Greenberg, and Ferguson, 1985; Smits and Watson, 1985; Stein and Epstein, 1985; U.S. Congress, 1983). Among the concerns are the incentives for adopting inappropriate cost-saving technology (Davis, Anderson, and Steinberg, 1984; Lave, 1984; Stern and Epstein, 1985), the disincentives for developing potentially effective new technology (Saksena, Greenberg, Ferguson, 1985; Smits and Watson, 1985), and the likelihood that medical specialties may be pitted against each other in competing for hospital resources to be allocated among new technologies (Iglehart, 1982).

Some observers think the necessity to reduce costs per case may stimulate inappropriate use of technology. For example, Stern and Epstein speculate that capital equipment (such as kinetic tables) may be purchased and utilized to reduce nursing costs. Davis, Anderson, and Steinberg mention that low cost pacemakers (with short lifespan) could be used instead of more expensive pacemakers with longer functioning time. In this case, the ultimate cost to PPS would be greater (assuming the pacemakers with short duration are replaced in time) despite reductions in cost per case.
The Coding System

Smits and Watson (1985) identify the ICD-9-CM coding structure as requiring monitoring and revision for technologic breakthroughs, and also mention the effect of price weights on innovation. The ICD-9-CM coding structure, the DRG classification structure, and the price weights must all reflect the recently developed (cost-saving, cost-effective) advance for the effect to occur. Conversely, the price weights in effect at a given point in time will help to determine which advances are stimulated and which discouraged. They recommend a national committee to monitor technologic change and to standardize and disseminate needed concurrent changes and coding rules within ICD-9-CM.

The DRG Classification System and New Technology

Lave (1984) predicts that there will be pressure to devise new DRGs to accommodate new, expensive technologies. A case in point appears in the literature. Saksena, Greenberg, and Ferguson (1985) argue for developing a new DRG for a technologic advance not accounted for in the DRGs. The technique—electrophysiologic evaluation of patients with arrhythmias—allows physicians to determine appropriate drug therapy in a single admission, while the empiric approach may require several hospitalizations to stabilize drug dosage. The costs of the single admission for electrophysiologic monitoring exceed the reimbursement allowed for the DRG based on the empiric treatment, but empirically determined therapy will ultimately be more costly. Whether this particular technologic advance warrants DRG evaluation or not, the appearance of this type of literature prompts consideration of Smits and Watson's strategy or others to account for technologic change within a fairly stable classification system.

Incentives to Adopt New Technology Under Per-Case Payment

Smits and Watson (1985) believe that PPS may stimulate a larger regulatory role over health technology diffusion, but the limited evidence on diffusion alone is contradictory. Comparing the diffusion rates for CAT scanning and magnetic resonance imagers (MRI), Steinberg, Siske, and Locke (1985) cite greater elapsed time in HCFA's
decisionmaking regarding coverage for the MRI as one of the factors contributing to the slower rate of diffusion for this technology. Such coverage decisions are not part of PPS, but per-case payment might also give hospitals incentives either to adopt or not adopt a new technology. However, in interviews of physicians and administrators considering MRI purchase, Hillman et al. (1985) found consensus that competitive market forces would prompt purchase regardless of HCFA's decision. Hillman et al. observe that incentives to contain costs are not being applied to this specific technology.

New technology that may become cost-saving may still be expensive at the outset (Davis, Anderson, and Steinberg, 1984; Stern and Epstein, 1985). Stern and Epstein recommend that hospitals be designated and reimbursed for testing promising technologic developments. These hospitals would conduct trials to determine cost effectiveness, while overall PPS costs would be spared capital equipment costs while the testing is restricted to these hospitals.

Technology also directs attention to the exclusion of other payors and physicians from PPS. Overall system costs may not be reduced if costs of new technology are simply shifted to other payors or to ambulatory settings (Stern and Epstein, 1985; Davis, Anderson, and Steinberg, 1984). As Davis and colleagues also note, "Physician reimbursement still encourages the performance of procedures, especially technologically sophisticated procedures."
IX. CONCLUSIONS

In part because of the impetus of the Medicare Prospective Payment System, there is an extensive and rapidly growing literature on Diagnosis Related Groups. Despite its size, for several reasons the literature rarely engenders unequivocal answers to questions about the performance of the DRG case mix adjustment within PPS:

- **The issues surrounding case mix adjustment are broad and often lack clear definition.** The concept of case mix itself means different things to different people. The boundaries between DRGs and PPS are blurred. So too are the boundaries between case mix and practice patterns, coding practices, quality of care, and appropriateness of care.
- **PPS was implemented after a fairly short public debate, catching the health services research community somewhat unprepared.** As a result of the high level of interest in the topic of DRGs, there was pressure to report quickly on whatever was known, but early reports lacked hard data. Many theoretical or philosophical pieces appeared, but without the benefit of a database adequate to test the relevance of the theory. A large number of fragmentary studies appeared, in which a few cases from one or two hospitals were used to address a narrowly focused question. Even now, as national, representative data become available, the pace with which various parameters within PPS are being modified condemns researchers to reporting on historical rather than current systems.
- **A "definitive" study would require such a large sample size and such extensive data collection on each case that the cost would be prohibitive.** Studies with large, representative samples of cases are limited to a few, easily collected measures about each case. Studies in which more complete information is collected on each case, from detailed medical records or
patient observation, must restrict the size of the sample and limit the number of hospitals involved.

Despite these limitations, a more settled, coherent literature on DRGs is beginning to appear. But as this literature arises, it becomes clear that even apparently sophisticated researchers publishing in reference journals do not understand important features of PPS. This suggests that for the bulk of hospital administrators and practicing clinicians the system may be even more mysterious and that the incentives it produces may reflect perceptions that are often mistaken rather than the actual workings of this complex system.

The conceptual issues are difficult to resolve and will undoubtedly be with us for many years. However, as the policy issues crystallize, the studies can direct attention toward these issues in a more useful way. Over the next two or three years, a clearer evaluation of the strengths and weaknesses of DRGs should emerge.
REFERENCES


Brook, R. H., Quality of Care Assessment: A Comparison of Five Methods of Peer Review, Department of Health, Education and Welfare (DHEW publication No. HRA 74-3100), Rockville, Maryland, 1973.

Brook, R. H., A. Davies-Avery, S. Greenfield, L. J. Harris, T. Lelah, N. E. Solomon, and J. E. Ware, "Assessing the Quality of Medical Care Using Outcome Variables: An Overview of the Method," Medical Care, 15(9):1-163 (September Supplement), 1977.


Coffey, R., and M. G. Goldfarb, DRGs and Disease Staging for Reimbursement Medicare Patients, Working Paper Number 1, Hospital Studies Program, National Center for Health Services Research, Rockville, Maryland, October 1984.


Roos, N. P., and L. L. Roos, "Surgical Rate Variations--Do They Reflect the Health or Socio-Economic Characteristics of the Population?" *Medical Care*, 20(9):945-958, September 1982.


SysteMetrics, Inc., The Measurement of Inappropriate Hospital Utilization in a Nationally Representative Sample, Santa Barbara, California, January 1983.


