United States General Accounting Office

GAO

Report to the Ranking Minority Member, Special Committee on Aging, U.S. Senate

June 1988

# MEDICARE

# Improved Patient Outcome Analyses Could Enhance Quality Assessment


RELEASED





GAO/PEMD-88-23

GAO
United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-229397

June 27, 1988

The Honorable John Heinz
Ranking Minority Member
Special Committee on Aging
United States Senate

Dear Senator Heinz:

This report responds to your November 4, 1986, letter asking us to examine the Health Care Financing Administration's (HCFA) analysis and use of existing administrative data to monitor patient outcomes under the Medicare program. As you requested, we compared the approaches that HCFA has employed in its internal analyses of Medicare outcomes to analogous approaches developed by HCFA contractors and independent researchers. We also examined the feasibility of using Medicare administrative data to assess the effects of the prospective payment system (PPS) on patient outcomes.

As we agreed with your office, unless you publicly announce the contents of this report earlier, we plan no further distribution of it until 30 days from the date of the report. We will then send copies to interested congressional committees, the Secretary of Health and Human Services, the Administrator of HCFA, and other interested parties, and will make copies available to other persons upon request.

Sincerely yours,

Eleanor Chelimsky,
Director

# Executive Summary

## Purpose

In March 1986 and again in December 1987, the Health Care Financing Administration (HCFA) identified specific hospitals having mortality rates that were substantially higher or lower than expected given the mix of Medicare patients they treated. These analyses attracted widespread interest as well as concerns about misinterpreting the results.

At the request of the ranking minority member of the Senate Special Committee on Aging, GAO examined HCFA's approach to analyzing Medicare patient outcomes. The primary question was whether HCFA could obtain more or better information to guide Medicare quality assurance activities using the administrative data on individual patients that it already collects. The five study objectives were to (1) describe the approaches HCFA employs to analyze existing Medicare administrative data on mortality and morbidity as an indicator of the quality of hospital care, (2) examine the uses that HCFA has made of these outcome analyses to guide quality assurance in the Medicare program, (3) identify other relevant approaches that could be applied to Medicare administrative data, (4) assess the relative strengths and limitations of HCFA's and other approaches in terms of their substantive focus and technical quality, and (5) determine the feasibility of analyzing administrative data to assess changes in Medicare patient outcomes associated with the introduction of the prospective payment system in 1983.

## Background

Primary responsibility for ensuring quality care for Medicare hospital patients rests with the 54 state-level Peer Review Organizations (PROs). They fulfill this function through reviews of medical records by nurses and physicians for selected cases. Patient outcome analyses based on Medicare's administrative data files provide a useful complement to the PRO reviews because the uniform billing data on every Medicare patient permits an efficient and consistent examination of all cases.

One difficulty confronting outcome analyses based on administrative data files is the restricted range of clinical data generally included in such files. Analysts need clinical data to adjust for differences among patients in "severity of illness"; that is, their intrinsic risk of dying or experiencing other adverse outcomes, independent of the quality of care received. With adequate adjustments, typically based on differences in diagnosis and general health status, comparisons of health care outcomes may provide a credible indication of quality of care.

## Results in Brief

A comparison of the 1986 and 1987 hospital mortality analyses shows that HCFA has strengthened the technical quality of its analyses. However, it could make additional improvements in the key area of patient severity adjustment. To make future analyses of Medicare patient outcomes more credible and useful, HCFA should more fully validate the analytical approaches selected, systematically check its data for accuracy and completeness, and analyze outcomes from several years to reduce the effect of random variation. HCFA's application of Medicare patient outcome analyses has so far been limited, and not notably effective in identifying quality problems.

## Principal Findings

The 1987 hospital mortality analyses improve on the 1986 analyses in their use of patient-level data, clinically coherent diagnostic groups, information on comorbidities, and more appropriate techniques to adjust for severity of illness. HCFA also maintains ongoing monitoring systems that compare outcomes over time and across a limited number of patient subgroups.

HCFA's major use of its outcome analyses was to require organizations bidding to remain or become PROs in 1986 to examine the hospitals identified in HCFA's 1986 analyses. GAO found that the PROs confirmed only a handful of these hospitals as having definite or likely quality problems. The data cannot answer why this occurred, but GAO believes that a careful investigation of this issue should precede any future use of similar outcome analyses to target PRO reviews.

GAO identified six distinct approaches to analyzing Medicare patient outcome data, in addition to the three employed by HCFA. Four emerged from HCFA's extramural research program, and two were developed independently.

GAO found that several of the approaches developed independently or by HCFA contractors adjusted for differences in patient severity in ways that took greater advantage of the clinical data on principal diagnoses and procedures available in administrative files than did HCFA's approaches. HCFA could potentially achieve similar results by incorporating comparable risk variables into future mortality analyses.

Several approaches that analyze patient subgroups demonstrate the potential for identifying types of cases with unusually favorable or adverse outcomes. HCFA has primarily compared mortality rates among

individual hospitals. If HCFA were to expand its analyses of patient subgroups, rather than hospitals, using more sophisticated adjustments for patient severity, it could then test whether outcome analyses focusing on patient subgroups defined by demographic or diagnostic characteristics would usefully supplement, or partially substitute for, hospital-based analyses as a way of targeting PRO quality reviews.

Certain limitations apply to all nine analytical approaches. First, none has yet been adequately validated for effectiveness in targeting cases for quality review. Adequate testing would involve systematic comparison of outcomes using these approaches to outcomes derived from a detailed review of medical records or other available evidence of quality of care. Second, all of the approaches are vulnerable to missing and inaccurate data in Medicare's administrative files. Until HCFA establishes the nature and magnitude of such problems for each data element used by these approaches, the effect of such deficiencies on analyses of Medicare patient outcomes will remain unknown.

Third, all the approaches must contend with the uncertainty that random variation introduces to analyses of mortality rates, especially those that involve small numbers of cases. In its 1987 hospital analyses, HCFA took account of random variation by calculating a range of expected mortality for each hospital. The breadth of these ranges increased as the number of cases analyzed declined; thus observed mortality for smaller hospitals had to deviate more markedly from expected mortality to fall outside the predicted range. This made the HCFA analysis less capable of detecting relatively poor outcomes for smaller hospitals. One solution would be to combine Medicare patient data from several years. Hospitals with larger numbers of Medicare patients could still be analyzed yearly to monitor short-term trends in outcomes.

Finally, existing analytical approaches using Medicare administrative files provide little capability for analyzing outcomes other than mortality. HCFA has addressed this problem in its most recent extramural grant solicitation.

An analysis of changes in Medicare patient outcomes associated with the shift to prospective payment could be conducted using existing administrative files. However, the results would be open to challenge, owing to the likelihood of major systematic error in the diagnostic information needed to adjust for patient severity, as well as the difficulty of distinguishing PPS-induced changes from other changes likely to have occurred over the lengthy period of phasing in prospective payment.

Two ongoing HCFA studies may produce much of the information about the effects of PPS that is feasible to derive, given the limitations of the available data.

## Recommendations to the Secretary of HHS

GAO recommends that the Secretary of HHS direct the Administrator of HCFA to (1) strengthen HCFA's outcome analyses by adopting specific improvements identified in this report, such as taking greater advantage of available diagnostic data in adjusting for patient severity of illness, employing data for several years when analyzing outcomes involving small numbers of cases, and expanding HCFA's analysis of comparative outcomes among demographic and diagnostic subgroups of patients (see pp. 96, 97, and 99); and (2) improve outcome analyses more generally by actions outlined in this report, such as periodically assessing the relative strengths and limitations of available approaches for analyzing Medicare patient outcome data in terms of substantive focus, technical adequacy, and degree of validation (that is, their overall effectiveness in identifying patterns of patient care with quality problems). Further, HCFA should evaluate the completeness and accuracy of the data elements that are used to analyze Medicare patient outcomes. The assessment should be based on a nationally representative sample of Medicare patients. The results should be published and appropriate corrective actions taken. (See pp. 102 and 103.)

## Agency Comments

While HHS found the report "thorough and scholarly" and generally concurred with GAO's recommendations, its comments do not always address the specific points raised in those recommendations. For example, the Department described its longer term efforts to expand the clinical data in its administrative files, but did not comment on GAO's proposals to strengthen patient severity adjustment in HCFA's interim analyses using its existing data sets. Overall, the GAO recommendations would both strengthen HCFA's analyses in the near term and facilitate more fundamental improvements by establishing procedures for validating analytical approaches and assessing data accuracy. HHS' comments pertaining to the recommendations and GAO's responses are presented in chapter 6; technical comments are addressed in appendix II.

# Contents

**Abbreviations**

| | |
|---|---|
| CPHA | Commission on Professional and Hospital Activities |
| DRG | Diagnosis-related group |
| GAO | General Accounting Office |
| HCFA | Health Care Financing Administration |
| HSQB | Health Standards and Quality Bureau |
| HHS | Department of Health and Human Services |
| ICD-9-CM | International Classification of Diseases, 9th Revision, Clinical Modification |
| ICDA-8 | Eighth Revision, International Classification of Diseases, Adapted for Use in the United States |
| MEDPAR | Medicare Provider Analysis and Review File |
| PPS | Prospective payment system |
| PRO | Peer Review Organization |
| RAMI | Risk-Adjusted Mortality Index |

# Introduction

In March 1986, the Health Care Financing Administration (HCFA) released a list of hospitals its study revealed as having 1984 mortality rates that were significantly higher or lower than expected, either overall or for patients with one of nine specific medical conditions. HCFA published a second analysis of Medicare hospital mortality rates—based on a revised methodology—in December 1987. Meanwhile, another HCFA analysis showed an increase in the proportion of Medicare hospital patients who died in 1985 compared to 1984, though the death rate for the beneficiary population as a whole remained unchanged.

The high level of interest in these analyses, and the concerns of hospitals and others that the results could be misinterpreted, underscore the importance of employing the best available methodology when using outcomes of hospital treatment such as mortality as indicators of the quality of care received by Medicare patients. At the request of Senator John Heinz, the ranking minority member of the Senate Special Committee on Aging, we have examined HCFA's approach to analyzing existing Medicare data on individual patients to monitor the outcome of inpatient hospital care. Specifically, we have assessed HCFA's current methods for interpreting and using such data and suggested possible improvements.

## Background

Improving health status and preventing untimely death represent the ultimate goals of health care. This is the main rationale for using information on patient outcomes as an indicator of the quality of care provided. However, success in forestalling adverse outcomes is often not feasible, even with the best available care, for patients who are very sick. Therefore, comparisons of health care outcomes should be based on an analytical approach that takes account of initial differences in patient condition or "severity of illness."[1] Moreover, any limitations in adjusting for severity or other methodological problems need to be recognized in interpreting the results of such analyses.

The limitations of outcome analyses tend to be particularly acute when they rely exclusively, as does each of the HCFA analyses cited above, on information recorded in Medicare's computerized administrative data files. The kind of detailed clinical information required to make sophisticated assessments of patient condition can generally be found only in a patient's original medical record. Since abstracting medical record data

[1] In this report, the term "severity of illness," or "patient severity," refers to the full range of demographic (e.g., age and sex) and clinical factors, including both principal diagnosis (the main reason for admission to a hospital) and comorbidities (diagnosed problems that are not related to the principal diagnosis), that could affect a patient's prospect for recovery.

is costly, it typically is done only for relatively small groups of patients for research purposes. Therefore, HCFA has based its overall outcomes monitoring on the few data elements that Medicare routinely records for each patient as part of its administrative process to confirm eligibility and pay bills. These include information about diagnoses and major procedures performed; dates of admission, discharge, and death; and certain demographic characteristics such as age and sex.

The interest in analyzing Medicare patient outcomes derives in part from a concern that Medicare's adoption of a prospective payment system (PPS) for hospitals has led to compromises in the quality of care that patients receive. This concern stems from the financial incentives created by PPS to limit the amount and complexity of services provided. Until the introduction of PPS in 1983, Medicare had reimbursed hospitals for discrete services rendered to patients, such as laboratory tests, hospital days, and so on. Because it perceived that this encouraged hospitals to maximize the number of services delivered, Congress enacted the prospective payment system, which provides a fixed fee for each patient. The fee varies with the patient's diagnosis or major procedure performed, but (with few exceptions) is not increased for patients whose hospital stay exceeds a set number of days or who receive more ancillary services. As a result, hospitals now profit from a minimization, rather than maximization, of the number and complexity of services provided to the patients in their care. This raises the question of whether, and to what extent, PPS leads hospitals to refrain from furnishing medically necessary or useful services to ensure that their own costs in providing care do not exceed the Medicare payment for that care.

The Medicare program relies on Peer Review Organizations (PROs) as the main safeguard against inadequate treatment for individual patients. The 54 PROs are private organizations, under contract to HCFA, that review the appropriateness and quality of care provided to selected groups of Medicare patients in a particular state or territory. These reviews are typically conducted by trained medical personnel, nurses and doctors, who examine the full medical record. All hospitals, as a condition of payment under PPS, are required to supply these records for the cases that the PROs designate for review. Hospitals and physicians whose reviewed cases demonstrate a significant pattern of quality problems can, if they do not improve their performance, lose their eligibility to participate in the Medicare program.

Outcome analyses based on administrative data are potentially useful as a complement to the basic process of PRO review. Since outcome data are

collected and merged into comprehensive computer files for all Medicare beneficiaries, a wide variety of analyses can be conducted relatively efficiently. Thus, outcome analyses offer an inexpensive means for monitoring general patterns of care for all Medicare patients, including those that the PROs do not examine, and may provide a check on some aspects of the criteria that HCFA and the PROs use to select cases for review. Moreover, outcome analyses can treat each case in a systematic and uniform fashion, which facilitates both an aggregation of the results to the national level and flexible disaggregation to a large number of potential subgroups.

# Objectives, Scope, and Methodology

## Objectives

As requested by Senator Heinz, and agreed to with his staff, our objective was to examine HCFA's analysis and use of existing administrative data to monitor the outcome of care received by Medicare beneficiaries and compare HCFA's analyses to possible alternative approaches. Our purpose was to see if HCFA could obtain more or better information to guide Medicare quality assurance activities such as PRO reviews.

To address this overall objective, we formulated five subsidiary objectives. They were to:

- describe the approaches HCFA employs to analyze existing Medicare administrative data on mortality and morbidity as an indicator of the quality of hospital care,
- examine the uses that HCFA has made of these outcome analyses to guide quality assurance in the Medicare program,
- identify other approaches for conducting outcome analyses that could be applied to Medicare administrative data,
- assess the relative strengths and limitations of HCFA's and other approaches in terms of their substantive focus and technical quality, and
- determine the feasibility of analyzing administrative data to assess changes in Medicare patient outcomes associated with the introduction of the prospective payment system in 1983.

## Scope

This study addresses the potential for improving HCFA's analysis and use of outcome data derived from administrative data files. The focus on outcome indicators of quality rather than process of care measures—such as judgments on the appropriateness of diagnosis and treatment—reflects the nature of the information currently available in Medicare administrative files. It does not imply a judgment on our part that outcome measures alone are sufficient for quality monitoring and assessment. On the contrary, most of the approaches we review are explicitly intended to serve as screening devices for targeting intensive, process of care reviews of medical records.

The emphasis on inpatient hospital care similarly reflects the greater availability of HCFA data on treatments in that setting. We examine both mortality and morbidity outcomes of care for Medicare patients, but in practice, most of the work in this area has concentrated on analyses of mortality. For the purposes of this study, we take the current content of HCFA's central administrative data files as a given, and leave to others to study the question of what additional data elements might be useful if HCFA were to expand the information assembled on individual patients.

We identified nine separate analytical approaches that we judged relevant to analyses of Medicare patient outcomes using existing administrative data, even though these approaches may not have been designed with that intent. Each approach represents a specific combination of measures and statistical analytical techniques. Measurement issues relate to individual variables or factors. Some variables pose few measurement problems (e.g., sex), but for others there may be a wide variety of ways to categorize and thereby structure individual cases. Statistical techniques such as logistic regression or recursive partitioning are used to establish relationships among multiple variables; for example, to make adjustments for variations in patient severity. The assumptions and limitations inherent in these statistical techniques constrain the ways they can be applied appropriately using particular data elements and measures.

The nine approaches to analyzing and using outcome data that we investigate include those HCFA has applied in its intramural analyses of the Medicare program, as well as approaches developed under its aegis through extramural research. We also examine approaches developed independently from HCFA that employ similar data elements and that have been empirically tested to some degree. We excluded from our analysis the independent approaches that require data elements not

available in the existing Medicare Statistical System (such as specific clinical findings or physician ratings of patient condition).

## Methodology

We based our findings and conclusions on information drawn from official HCFA documents, interviews with HCFA and PRO officials and other substantive and methodological experts, and an in-depth review of the quality of care literature. First, we describe the data sources and analytical methods used for each of our five study objectives. Next, we outline the seven specific criteria employed in our comparison of the strengths and limitations of the nine analytical approaches we examined.

### Data Sources and Analytical Methods

To determine HCFA's current practices regarding the analysis and use of health outcome data (objectives 1 and 2), we interviewed HCFA officials responsible for different aspects of the agency's quality assurance activities, including those in the Health Standards and Quality Bureau, the Office of Research and Demonstrations, and the Bureau of Data Management and Strategy. Wherever possible, we supplemented these interviews with a review of HCFA documents that outline changes in the analysis and use of these data over time. We collected available documentation on ongoing research on outcome analyses funded by HCFA, including grant proposals and draft reports, and interviewed many of the principal investigators. Finally, we attended conferences organized by HCFA to discuss its intramural and extramural activities relating to quality of care.

We also had extensive discussions with representatives of the Peer Review Organizations and attended relevant professional meetings. We interviewed officials in each of the 51 PROs that were required to analyze HCFA's 1986 hospital mortality outlier lists as part of the process of negotiating contracts with HCFA for the 1986-1988 contract cycle.[2] These interviews allowed us to characterize the primary example of HCFA's use of outcome analyses (objective 2): the PROs' examination of the 2,313 outliers identified by HCFA. Wherever possible, we spoke to PRO officials directly involved in that effort, but there may be some imprecision in their recollection of the results of that analysis. We therefore focused on the broad patterns that emerged across the 51 PROs regarding the procedures they followed and the results they obtained.

[2]The outlier hospitals were those identified in HCFA's statistical analyses as having mortality rates that were significantly higher or lower than expected. HCFA's 1986 analyses did not include American Samoa/Guam, Puerto Rico, or the Virgin Islands.

We identified potential alternatives to HCFA's current practices (objective 3) by reviewing various technical literature relating to quality of care. We drew as much as possible from the previous efforts of other researchers,[1] supplemented with our own iterative searches through written and computerized bibliographical sources and consultations with experts inside and outside HCFA.

By using multiple sources, we hoped to maximize the probability of identifying relevant approaches for review, including research conducted abroad. Nevertheless, the rapid development of this field and the wide spectrum of academic disciplines involved made it difficult to ensure that we included everything of possible interest.

Next, we obtained descriptions of all potentially applicable approaches to analyzing outcome data identified in these searches and screened them to see if they met the following criteria: (1) applicability to an elderly, or at least adult, population; (2) reliance on computerized administrative data such as hospital discharge abstracts, rather than medical record reviews; (3) applicability to inpatient care; and (4) focus on outcome as opposed to process of care measures of quality. If an approach did not meet all four criteria, we did not consider it further.

For each approach, we obtained information pertinent to our criteria from published or unpublished written sources. Where these were not adequate, we contacted the principal investigators associated with the approaches and asked them to provide that information. All investigators had an opportunity to review and comment on the characterization of their approaches in this report.

We based our assessment of the strengths and limitations of the HCFA intramural and extramural and non-HCFA approaches (objective 4) on a standard set of criteria reproduced in appendix I and discussed below. They derive from a synthesis of the critical commentary that has

[1]See, for example, Mark S. Blumberg, "Risk Adjusting Health Care Outcomes: A Methodologic Review," Medical Care Review, vol. 43, no. 2, (Fall 1986), pp. 351-93; and Arlene Fink, et al., "The Uses and Misuses of Hospital Outcome Data: What Does the Literature Say? A Literature Handbook" (Presentation to the American Medical Review Research Center Executive Training Program on Peer Review Outcome Data, Washington, D.C., April 21-22, 1987).

emerged with respect to the competing strategies for conducting outcome analyses of this type.[1] Taken together, the criteria attempt to specify what conclusions about quality of care can and cannot be drawn using these alternative approaches.

Our evaluation of the potential for assessing PPS effects on quality using administrative data on outcomes (objective 5) derives from a combination of interviews with relevant HCFA officials and contractors together with material collected through our literature review.

We performed our review in accordance with generally accepted government auditing standards.

## Comparisons Across Approaches

We assessed the strengths and limitations of the nine selected analytical approaches (objective 3) in terms of seven key issues:

1. General Purpose. Policymakers can obtain information relevant to three major aspects of quality assurance from outcome analyses using Medicare data: assessing the performance of individual hospitals, monitoring of changes in Medicare patient outcomes over time, and identifying differences in outcomes across subgroups of patients. Each of these policy concerns leads to somewhat different analytical issues and program decisions. Some approaches may be flexible enough to address several concerns, while others may be limited to just one.

2. Substantive Focus. The analytical approaches we examined differ in the aspects of care they study and the portion of the Medicare beneficiary population they encompass. This includes both the specific types of outcomes monitored (mortality, readmissions), the range of medical or surgical conditions considered, and any other patient or provider subgroups (e.g., demographic) that have been separately analyzed.

3. Severity Adjustment. Outcomes per se can be reported without adjustment; however, comparing outcomes to make inferences about quality of care requires an assumption that differences in outcomes reflect differences in the care received, and not inherent variation among patients in

[1]For example, American Medical Review Research Center, "Draft Statement on Public Release of Mortality Data" (Presentation to the Center's Executive Training Program Apr. 21-22, 1987); Blumberg, "Comments on HCFA Hospital Death Rate Statistical Outliers," Health Services Research, vol. 21, no. 6 (Feb. 1987), pp. 715-39; Blumberg, "Risk Adjusting Outcomes"; Fink, et al., "Uses and Misuses of Hospital Outcome Data"; Harold S. Luft and Sandra S. Hunt, "Evaluating Individual Hospital Quality Through Outcome Statistics," Journal of the American Medical Association, vol. 255, no. 20 (May 23-30, 1986), pp. 2780-84.

their responsiveness to treatment. As noted, patient "severity of illness," as used in this report, connotes the amalgam of demographic factors (such as age and sex) and clinical factors (specific diagnosis, stage of illness, and presence of preexisting comorbidities) that influence the probability of favorable or adverse outcomes for individual patients independent of the treatment received.

Given existing limitations in medical knowledge, as well as constraints in information available in administrative data files, any assessment of severity of illness is necessarily imprecise. However, we can compare different analytic approaches to see whether they attempt to adjust for patient severity, which specific factors they take into account, whether there are specific problems with the measures or data employed, and how they deal with recognized pitfalls. For example, the way Medicare codes patient diagnoses makes it difficult to distinguish between comorbidities present at admission and complications that develop during the course of a hospital stay. We consider how, if at all, the nine approaches try to overcome this problem.

4. Technical Adequacy. The level of confidence that can be placed in the results of these approaches depends on the properties of the measures and analytical techniques employed, as well as the relationship between them. We considered the validity, reliability, and sensitivity of the measures used.

Validity—how accurately a measure or indicator represents what it purports to represent—is a particular problem because the data available in administrative files often relate only indirectly to the main subject of concern. We critically examined the limitations of the measures employed in the nine approaches we studied to assess the risk of misleading inferences about outcomes or quality of care. Similarly, we looked for evidence of reliability—the consistency with which measures, as they are applied in a given approach, produce uniform results in equivalent situations (e.g., two different raters of the same case). Finally, we examined the sensitivity of the indicators used to see if we found significant variation among individual cases on a particular measure. Mortality rates in and of themselves are generally recognized to be insensitive measures of quality of care for most medical and surgical conditions precisely because most patients survive, whether or not they receive quality care.

Another technical issue concerns the appropriateness of the application of specific statistical procedures. Most of the analytical approaches we

examined employ statistical procedures to make adjustments for patient severity. Typically, these procedures generate the estimates of "expected" outcomes; that is, the anticipated rate or probability of certain outcomes (such as death) occurring for selected groups of patients, taking only specified predictor variables (such as age or diagnosis) into account. The techniques of multiple regression, logistic regression, recursive partitioning, and Cox proportional hazards modeling are all used for this purpose. All are based on certain assumptions and have defined limitations. Confidence in the results of outcome analyses depends in part on the degree to which these techniques have been applied in ways that are consistent with their assumptions and limitations. In addition, we looked for evidence of any bias in the results produced by these techniques; that is, systematic overestimation or underestimation of the predicted outcome for any specific category of cases.

Of particular concern was the use of statistical tests of significance. Such tests perform a somewhat different function in these approaches compared to their usual application in quantitative analyses. Normally, a researcher uses significance tests to assess how certain one can be that one or more "explanatory" variables are associated with the outcome of interest. By contrast, in the outcome analyses we examined, the set of predictor variables, or "model," used to adjust for severity of illness is not intended to explain all variation in the outcome (mortality), but just that component of the variation related to patient condition at hospital admission.

For each individual case (hospital, patient), these models generate an "expected outcome" indicating the likely result of treatment based on patient condition alone—as best the models can determine that. The magnitude of the difference between this "expected outcome" and that actually observed for a hospital or patient represents, conceptually, the effect of variation in the quality of care. However, the models adjust for severity on the basis of probabilities that intrinsically cannot accurately predict outcomes for each individual patient. Thus an element of random variation remains, even under the best of circumstances, particularly over short periods of time or among small groups of cases.

Limitations in medical knowledge and in the data available from administrative files make the severity adjustments provided by these models even more inexact. Therefore, some part of the difference between expected and observed outcomes does not reflect real differences in quality. In this context, significance tests provide a means of identifying

cases where that difference is sufficiently large that it exceeds what would be anticipated, on average, just on the basis of incomplete severity adjustment and random variation.

A variety of specific test procedures are available, which also have limitations and assumptions. We examined the extent to which the application of these statistical tests properly reflects their characteristics and requirements, as well as the appropriateness of inferences drawn from these results.

5. Data Quality. Since missing or inaccurate data could distort the results produced by these approaches, we considered their procedures for checking data and compensating for problems uncovered, as well as any available information on the likely impact of uncorrected data problems. For the most part, such information is relatively sparse. Consequently, our analysis tends more to raise questions and suggest general cautions about all the approaches than to distinguish clearly among them.

6. Validation. Ultimately, the value of these approaches derives from their capacity to identify, using administrative data, genuine quality problems. Validation refers to evidence on the overall effectiveness of these systems in accurately locating such problems. Some indication may be inferred from checking the consistency and logic of the results produced by the approaches under study. However, a more complete validation effort requires separate sources of evidence independent of the administrative data being analyzed. Different data sources will provide validating evidence that is more or less persuasive depending on the particular quality issues in question. Thus, a retrospective review of medical records can usually determine the medical condition of patients at hospital discharge, but generally will provide little insight on the skill with which a given surgical procedure was performed.

7. Overall Assessment. Finally, we summarized the most salient strengths and limitations for each approach under the individual criteria and highlighted the specific contribution the approach could make to analyses of Medicare outcomes.

We analyzed each issue in turn for each approach. Within each, we focused on particular aspects, either positive or negative, that distinguish that approach from the others. For example, we did not reassess

for each approach the measurement characteristics of standard variables such as age and sex. Table 6.1 in chapter 6 summarizes the characteristics of each approach with respect to these substantive and technical criteria.

## Report Organization

Chapter 2 examines the approaches HCFA has employed in its intramural analyses of Medicare patient outcomes. HCFA's use of those approaches in Medicare quality assurance is reviewed in chapter 3. Chapter 4 assesses the relevant approaches developed through HCFA's extramural research program, and chapter 5 evaluates those approaches developed independently of HCFA that met our criteria for inclusion. Chapter 6 summarizes the information from chapters 2, 4, and 5, systematically compares the nine approaches to each other along the dimensions defined by our criteria, and presents our conclusions and recommendations for improvements in HCFA's current practices. Readers less interested in detailed examinations of each analytical approach may wish to proceed directly to chapters 3 and 6. Chapter 7 considers the special question of how much could be learned from existing outcome data about any changes in quality of care associated with Medicare's shift to prospective payment.

Chapter 2

# HCFA's Intramural Analyses of Outcome Data

This chapter examines the intramural efforts within HCFA's Health Standards and Quality Bureau (HSQB) to use Medicare administrative data to monitor the quality of care provided to beneficiaries. Two activities predominate. Most widely known are the analyses of individual hospital mortality rates: the list of hospital mortality "outliers" released in March 1986 and the revised analysis of hospital-specific mortality released in December 1987. The second activity conducted by HSQB is the ongoing monitoring of patient subgroups and trends in aggregate outcomes over time. These approaches are analyzed and evaluated according to the seven criteria specified in chapter 1. Table 6.1 (p. 84) summarizes the most important findings with respect to these approaches.

## Analyses of Hospital Mortality Rates

On March 10, 1986, in response to a Freedom of Information Act request, HCFA released a list of hospitals with mortality rates in 1984 that exceeded or fell short of their estimated "expected" mortality rate by a statistically significant margin, either for all conditions or for nine clusters of diagnosis-related groups (DRGs—the payment categories established under Medicare's prospective payment system).[1] Originally intended for internal use in directing Peer Review Organization activities, the list attracted widespread attention in the media and generated considerable concern among hospitals, particularly those identified as aberrant by their inclusion on the lists. HCFA subsequently used the lists in negotiating new contracts with the PROs.

In December 1986, HCFA announced its intention to publish new hospital-specific mortality analyses by the end of 1987. HCFA then consulted with outside clinical and statistical experts, as well as representatives of provider and consumer groups, on ways to improve on the approach used in the first set of analyses. This second set of mortality analyses was released to the public on December 17, 1987.[2] It differed substantially from the 1986 analyses both in the technical details of the approach and the procedural steps that were followed.

In light of this sequence of developments, we chose to combine our review of the 1986 and 1987 analyses of hospital-specific mortality.

[1]Memorandum, Office of Medical Review, Health Standards and Quality Bureau, Health Care Financing Administration, Mar. 10, 1986; see also The New York Times, Mar. 12, 1986, p. A-1.

[2]Health Care Financing Administration, Medicare Hospital Mortality Information: 1986, HCFA Pub. No. 01-002, 7 vols. (Washington, D.C.: U.S. Government Printing Office, 1987); see also The New York Times, Dec. 18, 1987, p. B-5.

This facilitates an appreciation of how the new approach resolves some concerns raised about the original analysis. It also highlights limitations that continue to apply to both approaches.

## General Purpose

HCFA's 1986 hospital mortality analyses were designed to identify specific hospitals that, based on the outcomes of patients in their care, were more likely to have experienced quality of care problems than others. Therefore, only those hospitals whose observed mortality was outside statistically defined confidence intervals of "expected mortality" were listed.

The 1987 analyses of hospital mortality had the somewhat broader purpose of providing outcome-based information on all 5,971 Medicare-certified hospitals. Thus, rather than listing only those hospitals that are identified in the analysis as statistical outliers, the observed mortality rate and expected mortality rate (expressed in terms of a range representing the 95-percent confidence interval) are provided for every Medicare-certified hospital.

## Substantive Focus

The 1986 hospital analyses took inpatient mortality as their key outcome measure. HCFA performed separate analyses for all Medicare discharges in 1984 and for nine clusters of diagnosis-related groups. They included four medical conditions and five surgical procedures.[3] Together, these diagnostic clusters include about 30 percent of all Medicare discharges. No other patient subgroups were analyzed.

The outcome of interest changed for the 1987 analyses of hospital mortality during 1986. Instead of inpatient mortality, HCFA counted deaths in or out of the hospital if they occurred within 30 days of admission. For Medicare beneficiaries with multiple hospital admissions, HCFA only included the outcome of the last hospital stay completed in 1986. HCFA adopted this approach as a way of assigning deaths to a particular hospitalization, when multiple admissions occurred within 30 days of death.

[3]The medical conditions analyzed were congestive heart failure (DRG 127), acute myocardial infarction (DRGs 121-123), pneumonia (DRGs 89-90), and gastrointestinal hemorrhage (DRGs 174-175); the surgical procedures were cholecystectomy (DRGs 195-198), major joint procedures (DRG 209), transurethral prostatectomy (DRGs 336-337), coronary artery bypass surgery (DRGs 106-107), and pacemaker implants (DRGs 115-116).

Shifting the analysis from all hospital stays to the last discharge in a given year necessarily increased the observed mortality rate for hospitals, since patients had to survive all the excluded admissions in order to be admitted on a subsequent occasion. Some hospitals probably had a higher proportion of admissions excluded than others, depending on their mix of patients. Particularly where the earlier excluded admissions involved different diagnoses and hospitals, we question the rationale for dropping these hospital episodes from the analysis. A focus on the ultimate outcome for an individual patient over the course of a year may make sense in analyzing outcomes for the Medicare program overall, but this analysis was specifically concerned with the performance of individual acute care hospitals, each of which may treat only a portion of a lengthy series of illnesses.

As before, one overall mortality analysis encompassed all medical and surgical conditions. Separate analyses of certain specified diagnostic groups were also provided, including 10 higher risk and 6 lower risk diagnostic clusters, plus a 17th residual category for cases with other principal diagnoses. The 16 diagnostic clusters accounted for 70 percent of Medicare patients. They were defined by ICD-9-CM codes rather than DRGs.[1]

## Severity Adjustment

### The 1986 Analyses

In its 1986 mortality analyses, HCFA calculated an "expected" or predicted mortality rate for all Medicare-certified acute care hospitals (5,750), using multiple linear regression procedures. It conducted separate regression analyses for overall mortality and the nine DRG clusters. For each of these analyses, HCFA used this expected rate as its standard for assessing each hospital's actual observed mortality. It represented an estimate of what that hospital's mortality rate would be if its rate corresponded to the average for all hospitals having a comparable mix of patients.

The description of the analyses released by HCFA did not characterize the calculation of expected mortality as a severity adjustment—and in fact, noted some of its limitations in that regard. However, the procedure did

[1]ICD-9-CM stands for the International Classification of Diseases, 9th Revision, Clinical Modification, the internationally recognized classification scheme for specific diagnoses and surgical procedures.

make adjustments for differences in patient characteristics such as sex, age, and race, factors that could be expected to affect patient outcomes.

There are, however, three broad concerns raised by the particular methods used to generate these estimates of predicted mortality. They relate to the aggregation of data to the hospital level, the adjustment for differences in outcome by race, and the way in which case-mix—the variations across hospitals in the mix of diagnoses they treat—was handled.

Aggregating Data to the Hospital Level. Because these regression equations treated hospitals as the unit of analysis, the characteristics of individual patients could not be included in the analysis as such. Rather, they were aggregated to characterize the hospital as a whole (e.g., average age, proportion of patients who were male). This process of aggregation—analyzing the relationship of average characteristics to overall mortality rates—clearly sacrificed information relevant to predictions of inpatient deaths.

The alternative would be to analyze the relationship of individual patient characteristics to the outcomes experienced by those individuals. This approach does exact higher data processing costs, especially for very large data sets encompassing all Medicare hospitalizations. Expected mortality for any given hospital could then be estimated based on the cumulative expected mortality of the patients discharged from that hospital.

Adjusting for Race. The inclusion of race (proportion black and proportion neither black nor white) in the statistical analysis of expected mortality raises a separate caution. Its insertion presupposes that differences in outcomes across racial groups primarily represent differences in their inherent vulnerability to adverse outcomes. A counterargument is that blacks and other nonwhites could be receiving a disproportionate amount of poor quality care. To the extent that the former assumption is true, inclusion of the race variable protects those hospitals treating racial minorities from undue disadvantage, whereas to the extent the counterargument is justified, adjustment for this factor shields those hospitals from identification.

Case-mix Adjustments. An area of particular concern with HCFA's 1986 mortality analyses was its adjustment for case-mix. Variation in principal diagnosis (i.e., the nature of the medical problem that motivates the admission of a patient to the hospital) represents a major part of the difference among hospitals in the severity of the patients they treat. The

intensity of a given illness plays a role, but some diagnoses typically involve a much higher risk of mortality than others. Hospitals vary in their proportion of patients with low-risk and high-risk diagnoses, thus this difference needs to be taken into account when comparing patient outcomes for institutions overall.

HCFA based the adjustment it used for its analyses of overall mortality on the DRG assignment of each patient discharged. A regression equation took into account the proportion of cases each hospital had falling into 81 specific DRGs, plus a residual category of other, cancer-related DRGs. Two difficulties, one highly specific and the other more general, were raised by this procedure.

First, among these 81 DRGs were three for acute myocardial infarction (heart attack). One of these three codes is exclusively assigned to patients who die in the hospital and the other two only to patients discharged alive, thus in effect, the separate statistical adjustment for each of the three DRGs perfectly "predicted" whether or not the patient would live or die. Because acute myocardial infarction is a fairly common cause of inpatient deaths for elderly patients, this misspecification of the regression model could have substantially reduced the number of hospitals identified as outliers.[5]

More generally, DRGs provide only limited information about the specific diagnoses that led a patient to be admitted to the hospital. Patients are assigned to a given DRG in large part on the basis of their "principal diagnosis"[6]; however, each DRG combines a number of different diagnoses that have similarities in expected resource utilization, rather than similarities in patient condition or expected outcome. And patients with comparable diagnoses can often be assigned to different DRGs if their conditions are treated differently; for example, medically rather than surgically. Moreover, comorbidities, which represent additional medical problems besides the condition identified as the principal diagnosis,

[5]Mark S. Blumberg, "Comments on HCFA Hospital Death Rate Statistical Outliers," p. 721.

[6]Defined as "the condition established after study to be chiefly responsible for occasioning the admission of the patient to the hospital for care." See National Committee on Vital and Health Statistics, "Uniform Hospital Discharge Data Set, Final Summary Report of Activities, 1975-1978," reprinted as appendix X in Robert B. Fetter et al., The New ICD-9-CM Diagnosis-Related Groups Classification Scheme, HCFA pub. no. 03167 (Baltimore: Department of Health and Human Services, Sept. 1983), p. 456.

sometimes affect a DRG assignment, but often do not.[7] As a result, patients within a given DRG can often be clinically quite heterogeneous.[8]

## The 1987 Analyses

The 1987 mortality analyses differed in their approach to severity adjustment in several important respects, which effectively eliminated the three concerns with the 1986 analyses discussed above. Moreover, when releasing the results of these analyses to the public, HCFA took care to emphasize the limitations in their adjustment for severity.[9] We nevertheless raise several questions about HCFA's severity adjustment in order to suggest further areas of potential improvement. These questions relate to the use of two new variables and the 17 new diagnostic clusters that HCFA introduced into its 1987 analyses.

Improvements Over the 1986 Analyses. First among the improvements over 1986, the 1987 hospital analyses derived predicted mortality from patient-level data. Results for the individual hospitals reflected the cumulative expected mortality of their patients, rather than estimates based on hospital-wide averages. This should yield a more precise estimate of expected mortality. Second, race was not included as a predictive factor for mortality.

Third, the new analyses adjusted for case-mix by employing ICD-9-CM codes to define the 17 diagnostic clusters. These clusters were designed to be clinically coherent categories, based on the advice of a panel of physicians. The 1987 analyses also adjusted separately for the occurrence of specified comorbidities. The same eight comorbidities were considered for all patients. However, because each of the 17 diagnostic clusters had its own regression equation, the magnitude of the effect attributed to each of these comorbidities on expected outcome varied for each diagnostic group. The comorbidities selected by HCFA are relatively

[7]For example, there are many pairs of DRGs, where the "higher" or more severe alternative is assigned if the patient either has a complication or comorbidity or is over 69-years-old. For these DRGs, comorbidities become irrelevant for all patients aged 70 and above.

[8]See Mark C. Hornbrook, "Hospital Case Mix: Its Definition, Measurement, and Use, Part II: Review of Alternative Measures," Medical Care Review, vol. 39, no. 2 (Summer 1982), pp. 83, 89-91; and Paul Gertman and Steven Lowenstein, "A Research Paradigm for Severity of Illness: Issues for the Diagnosis-related Group System," Health Care Financing Review, Annual Supplement (Nov. 1984), pp. 80-82.

[9]HCFA's comments referred to a lack of any "direct measurement" of severity of illness in the analyses. But the analyses did include information on demographic and diagnostic factors as well as comorbidities. Thus, HCFA evidently employed a narrower definition of severity than we did in this report, focusing on the intensity or stage of illness for the principal diagnosis.

prevalent chronic conditions that would usually predate any given hospital admission.[10]

New Concerns. While avoiding most of the problems associated with the 1986 analyses, the 1987 hospital analyses raised some new concerns regarding adjustments for patient severity. First, we have questions about the two new factors that HCFA added to the adjustment formula: prior hospitalizations within the year and whether or not the patient was transferred from another hospital. HCFA included these factors in order to take account of prior hospital experience, given that the analyses focused on the outcome of the last recorded admission. However, these particular variables, as specified for the HCFA analyses, have several limitations that should be noted.

Transfers, in particular, do not have a clear conceptual relationship to variations in patient condition at admission. Unlike age, sex, diagnosis, and the presence of comorbidities, whose logical connection to patient severity is fairly direct and well understood, transfers could be associated with variations in patient severity in several different and inconsistent ways. For example, patients could be transferred from another hospital either to receive more intensive care, such as at a tertiary referral center, or to complete recuperation at a community hospital closer to home.

Multiple prior admissions may bear a more direct relation to patient condition, indicating a more advanced stage of an illness.[11] Nevertheless, in some instances patients with multiple admissions may have a chronic condition with a low probability of death within 30 days. Moreover, restricting the analyses to admissions within calendar year 1986 means that the amount of information on prior admissions will vary greatly among patients. For patients whose last admission occurred in December, the analysis will count all admissions in the preceding 12 months; whereas, for patients whose last admission occurred in January, it will only count prior admissions in the preceding few weeks or days. In fact, patients whose previous admissions occurred in 1985 and who died early in 1986 would have fewer admissions noted and therefore appear to be less severely ill when they died than patients who had two or more

[10]They include cancer, chronic liver disease, chronic renal disease, chronic cardiovascular disease, chronic pulmonary disease, cerebrovascular degeneration/chronic psychosis, hypertensive disease, and diabetes. (See HCFA, Medicare Hospital Mortality Information: 1986, Technical Appendix, p. ii.)

[11]More on readmissions as an outcome measure appears in our discussion of the technical adequacy of HSQB monitoring systems at the end of this chapter.

admissions in 1986 and died later in the year. HCFA could have avoided this problem had it calculated the number of prior admissions in the 12 months preceding the last recorded admission.

Our second concern with the 1987 analyses involves the 17 diagnostic clusters. Although derived independently from DRGs, they too have a substantial degree of clinical heterogeneity. That is, each cluster contained patients whose inherent probability of dying varied considerably. This was due to the great breadth of these diagnostic clusters, with just 16 clusters encompassing close to 70 percent of Medicare patients, plus a 17th residual category that is heterogeneous by definition and accounts for 30 percent of patients. Some critics have already pointed to major differences in the death rates of the principal diagnoses incorporated in these clusters.[12]

The central problem is that the analyses made no further adjustment for case-mix beyond the grouping of cases into these 17 broad categories. They therefore did not take full advantage of the more detailed clinical information conveyed by the principal diagnosis codes.[13] Even though the separate adjustment for the eight comorbidities brought additional diagnostic information to bear on the severity adjustment, this could not replace the information imparted by each patient's specific principal diagnosis. Incorporating that information into the analyses should be relatively easy. In all likelihood, HCFA could retain the 17 diagnostic clusters, maintaining the advantages of analyzing relatively larger groups (see the discussion of random variation below), but add one or more predictor variables that capture the differences in observed mortality rates among the principal diagnoses included within the clusters.

## Technical Adequacy

### Measurement Issues

One advantage of the fairly rudimentary model used to estimate expected deaths for the 1986 analyses is that variables such as age and sex pose few measurement problems. However, as noted, DRGs have limited validity as indicators of case-mix for purposes of evaluating out-

[12]Comments by the Commission on Professional and Hospital Activities, submitted Sept. 11, 1987, in response to the Federal Register notice of Aug. 17, 1987.

[13]Several of the approaches reviewed in chapters 4 and 5 demonstrate the feasibility and utility of incorporating information on separate principal diagnoses into the severity adjustment. See pp. 57 and 63.

comes since they were originally defined on the basis of expected resource utilization. Similar questions were raised above about the probable clinical heterogeneity of the 17 diagnostic clusters employed in the 1987 hospital analyses.

Substantial controversy also exists over the validity of the outcome measure used in the 1986 analyses: inpatient mortality. Critics of this measure point to its sensitivity to variations in average lengths of stay across hospitals and regions. Hospitals with relatively short lengths of stay are more likely to have discharged a patient before he dies. Therefore, two hospitals could treat an equivalent mix of patients equally successfully and still be rated very differently using inpatient mortality as the outcome measure. HCFA acknowledged this problem to a degree by including the state's average length of stay as a control variable in its regression equation, although this adjustment would provide only a rough correction for individual variation among hospitals.

Alternatively, analysts can assess mortality on the basis of a fixed time interval following hospital admission, whether or not the death takes place in the hospital. This, in fact, is what HCFA did in its 1987 mortality analyses, using an interval of 30 days from admission. This approach raises different issues. Whatever the length of that interval, some portion of hospital care will drop out of the analysis at the point where hospital stays extend past the set number of days. In addition, some proportion of patients will die outside the hospital during that time interval for reasons unrelated to their hospital treatment, while others whose hospital-related deaths following discharge occur after the cutoff will be counted as survivors.

Shortening the time interval decreases the likelihood of unrelated deaths outside the hospital but increases the amount of hospital care lost to the analysis as well as the number of hospital-related deaths that are missed. Lengthening the interval has the opposite effect. HCFA plans to assess the relationship of the results obtained for 30-day postadmission mortality to those revealed using longer intervals as data permitting follow-up of the 1986 discharges over a longer period become available.

## Analytical Techniques Employed in 1986

The specific analytical technique HCFA used to estimate an "expected" mortality rate for each hospital in its 1986 analyses, both overall and for the specified DRG groups, was multiple linear regression. Each hospital's residual—the difference between observed and expected mortal-

ity—was then tested for statistical significance to exclude differences reflecting chance variation in mortality rates over a given time period. Those hospitals that had residuals large enough to meet this statistical test were identified as outliers.

We discuss three issues related to HCFA's application of multiple linear regression in its 1986 analyses: the limited degree to which this statistical technique reduced variation among individual hospitals—particularly in the DRG-specific analyses, the potential for biased findings, and the appropriateness of the significance test used to identify hospital outliers.

Limited Reduction in Observed Variance. The regression equation establishes a level of "expected deaths" for each hospital (which serves as the standard for assessing each individual hospital's observed mortality) based on the relationship of the specified predictive factors to variation in mortality among all the hospitals under investigation. The credibility of these estimates for individual hospitals depends in part on the strength of the statistical relationship between the predictive factors included in the regression equation and hospital mortality overall. This relationship is usually assessed through a "goodness-of-fit" measure called "R-squared," which represents the proportion of variation in mortality "explained" or accounted for by the regression equation.[11]

Regression equations of this sort need not account for a large proportion of the variation in hospital mortality to provide a plausible severity adjustment. No regression model can be expected to perfectly predict observed hospital mortality rates, given the effect of random variation and other factors not included in the equation. Moreover, for these analyses, the regression equation is not intended to explain all variation in hospital mortality, but only that portion representing differences in patient severity. The remaining differences in hospital mortality rates presumably reflect, in addition to random variation, the differences in quality of care, which the analyses were designed to highlight.

Nonetheless, as the proportion of variation in hospital mortality rates accounted for by such regression models gets very small, severity adjustments made on the basis of these models become increasingly suspect. This is particularly an issue for the nine DRG-specific analyses conducted by HCFA, all of whose R-squared figures were quite low, ranging

[11]R-squared can range from 0 to 1.0, with 1.0 signifying that the regression equation accounts for all the observed variation in mortality and 0 indicating that the equation accounts for none of it.

from 0.003 to 0.068. In other words, HCFA's regression analyses of these conditions left between 93.2 and 99.7 percent of the variation in hospital mortality unexplained.

In this situation, the low R-squared has one of two explanations. If the HCFA regression equations in fact did a good job of adjusting for patient severity, that would indicate that severity of illness has little to do with differences in hospital mortality rates for the conditions being analyzed and can be safely ignored. If patient severity contributes substantially to variation in hospital mortality, adjustments based on a regression analysis with a very low R-squared would necessarily reflect only a small part of the relevant differences among hospitals in the condition of patients that they admit. In either case, the regression analysis provides little improvement over a simple comparison of the observed mortality rates of individual hospitals.

HCFA's analysis of overall mortality had a much higher R-squared—0.59—than any of the DRG-specific analyses. This, in part, reflected the inclusion of a much larger number of factors in the regression equation for overall mortality, most notably the 80-plus DRG variables. In particular, the adjustment for DRG 123 (myocardial infarction, discharged dead), in itself artificially boosted the model's predictive power.[15] In short, only in the overall mortality analysis did the regression account for a substantial proportion of the variance among individual hospitals. And there, the magnitude of that relationship remains uncertain because the adjustment was inappropriate for heart attack patients.

The amount of variation accounted for by a regression equation can be enhanced both by changing the variables introduced into the equation and by reducing the amount of random variation in the data being analyzed. The first strategy would focus on improving the severity adjustment. Random variation, however, represents the fluctuations in outcomes that occur by chance. Its effects are accentuated in analyses based on relatively small numbers of cases. Since deaths tend to be rare events for most hospital patients, small hospitals in particular are likely to have few deaths in any given time period—especially within a restricted diagnostic category. On a purely probabilistic basis, some hospitals experience one or two deaths while comparable hospitals do not. The solution to this problem is to analyze more cases by expanding the analysis to include a wider range of diagnoses or a longer period of time.

[15]See Harry M. Rosen and Barbara A. Green, "The HCFA Excess Mortality Lists: A Methodological Critique," Hospital and Health Services Administration (Feb. 1987), pp. 126, 123.

Thus the low R-squared of the DRG-specific analyses probably derives, at least in part, from the limitation of those analyses to hospital cases from one year.

Potential for Bias. Another basic criterion for judging regression models is the degree of bias. A biased model is one that makes systematic errors, either overestimating or underestimating predicted values, for particular categories of cases. For example, Blumberg found evidence that HCFA's 1986 overall mortality regression equation tended to predict higher mortality rates than actually occurred at hospitals where the expected rates were relatively low, and to predict lower mortality rates than actually occurred at hospitals expected to have relatively high death rates.[16]

However, when we replicated Blumberg's test for the outliers identified under DRGs 121-123, we found no similar pattern of bias.[17] It may be that the bias derives from the much more complicated and problematical adjustment for case-mix and state average length of stay made by HCFA in the calculation of expected death rates for overall hospital mortality, but not the DRG-specific analyses. Thus, the evidence of bias in the 1986 analyses is mixed. These results underline the value of careful testing to help ensure that such models do not make biased comparisons across the major groups being examined.

Significance Tests. A third area of concern in the application of the 1986 regression analysis has to do with the significance test employed to identify outlier hospitals. HCFA used a t-test, which was not appropriate for this analysis. A t-test assumes a "normal" (essentially bell-shaped) distribution of observed values above and below the predicted value. As several critics of the HCFA analysis have noted, that assumption applies poorly to numbers of deaths in individual hospitals, especially smaller hospitals. Deaths can only occur in integers (e.g., 0, 1, or 2 deaths) and they cannot fall below zero, while a normal curve centered on a relatively low value would predict both fractional deaths and deaths fewer than zero.

[16]Blumberg, "Comments on HCFA Outliers," pp. 722-23.

[17]We chose this diagnostic cluster because it had by far the largest number of negative or low outliers, though high outliers still clearly predominated. Blumberg's test requires a reasonable number of both types of outliers because it compares the proportion of high to low for different levels of predicted mortality.

For this reason, a Poisson or binomial distribution, which takes these characteristics into account, is more appropriate. By using a t-test of significance, hospitals having relatively few Medicare discharges were identified as outliers even when the actual probability that the number of deaths they experienced could have occurred by chance was substantially greater than the intended 5 percent.[18] The results of the 1986 mortality analyses were therefore distorted, to the disadvantage of smaller hospitals.

## Analytical Techniques Employed in 1987

For its 1987 mortality analyses, HCFA employed an entirely different set of statistical procedures than for its 1986 analyses. Except for the question of potential bias, which remains unexplored by HCFA, these changes either corrected or made irrelevant the specific concerns raised about the earlier analyses. Nevertheless, the 1987 analyses continue to confront some of the same basic problems, most notably the impact of random variation.

The 1987 analyses were based on logistic rather than linear regression. This selection, which follows from the shift of the analysis from hospital-level to patient-level data, is generally considered more appropriate for an outcome such as mortality, which is categorical in nature (i.e, alive or dead, not some increment or amount). The 17 separate logistic regression equations (one for each of the diagnostic clusters) generated coefficients from which the predicted mortality of all patients were calculated. These were aggregated to produce a predicted mortality rate for each hospital.

The predicted mortality rates for each hospital were then converted into a range of values representing HCFA's estimate of the uncertainty inherent in those estimates. These ranges took the place of the significance test used in the 1986 analyses to designate certain hospitals as outliers, but are generally more informative in that they show by their size the magnitude of the uncertainty for specific hospitals. Moreover, one can readily note which hospitals clearly fall within or outside of their expected mortality range and which barely miss or exceed those thresholds.

The formula that HCFA employed to construct these ranges took account of both expected random variation in the patients that each hospital

[18]See Blumberg, "Comments on HCFA Outliers," pp. 725-30; American Medical Review Research Center, "Draft Statement on Public Release of Mortality Data," pp. 7-8.

admitted in a given year and systematic differences among hospitals related to their mortality rates. The latter include any differences that were not specified in the regression model, including, HCFA was careful to note, differences in quality of care.[19] Thus, those hospitals whose actual mortality rates fell outside their range of predicted mortality rates diverge from the predicted by a margin that is substantially larger than one would expect given typical differences among hospitals as well as random fluctuations from year to year.

HCFA's formula for generating the range of predicted mortality rates appears to avoid the disproportionate risk to smaller hospitals of being designated outliers, which derived from the use of the t-test in the 1986 analyses. The new formula builds on the conventional statistical benchmark of two standard errors above and below the mean. This does assume a normal distribution of the residuals between observed and predicted mortality. However, these values were transformed using a logit scale, which prevented the range of predicted death rates from falling below zero or exceeding 100 percent. It also made the effective distribution quite asymmetrical.

This was true especially for smaller hospitals, for which the calculated impact of random variation was relatively large. As a result, the predicted ranges for smaller hospitals, or diagnostic categories with very low death rates or few patients within larger hospitals, tended to be quite wide, even as much as 0-100 percent. Few hospitals had observed mortality rates that fell outside such broad ranges. Nevertheless, it is not clear how closely the logit transformation approximated the results that would have been obtained if the confidence interval were based directly on a Poisson or binomial distribution.

Statistical Tests. Unlike linear regression with its R-squared, there are no "goodness-of-fit" measures for logistic regression that show the extent to which a given equation reduces total observed variance in the dependent variable—in this case, mortality. This renders moot any questions comparable to those raised about the small amount of variance explained by the linear regression equations used in the 1986 DRG-specific analyses. However, the potential for bias in the results, and the consequent desirability of testing for such bias, remains an issue. Unfortunately, as in 1986, HCFA did not examine the results of its 1987 analyses to see if they systematically underestimated or overestimated mortality for specified categories of hospitals.

[19]HCFA, Medicare Hospital Mortality Information: 1986, Technical Appendix, p. iii.

Instead, HCFA applied several other statistical tests to the results obtained from its logistic regression equations. These included cross-validation analyses designed to assess the potential impact of sampling error on the relative rankings of hospitals. They indicated that as long as the diagnostic mix of patients for each hospital is fixed, the rankings among the hospitals are highly stable. However, taking into account the impact of random variation in the types of patients that hospitals will treat in a given year made the rankings much less stable. Other tests showed that individual hospital rankings remained fairly consistent using different statistical procedures; that is, indirect standardization or multiple linear regression rather than logistic regression.

Implications of Relying on a Single Year of Data. The 1987 analyses, like those in 1986, reflected the outcomes of Medicare patients discharged from hospitals in a particular 12-month period. In March 1986, HCFA did not have much choice. Data after 1984 were not yet available, and diagnostic information on each Medicare patient had not been recorded prior to 1984. However, the decision to continue focusing on a single year in the 1987 analyses was more questionable, given that the analyses encompassed all Medicare hospitals across the country and that data were available from prior years.

The difficulty with focusing exclusively on one year's results stems from the particular problem noted above in overcoming the impact of random variation when analyzing the outcomes of smaller hospitals. Many of the nearly 6,000 Medicare-certified hospitals treat so few Medicare patients that they will normally experience only a handful of deaths in any diagnostic category over the course of a year. Under these circumstances, one or two extra deaths can produce a large discrepancy between observed and expected mortality rates.

The formula HCFA used to construct its range of expected mortality in the 1987 analyses addressed this problem by taking account of the number of cases involved. This results in wider intervals for smaller hospitals than larger ones. That compensated for the greater impact that random fluctuation would have on the observed mortality rates of smaller hospitals. By the same token, however, this formula reflected the limited capacity of the 1987 analyses to identify individual smaller hospitals with genuinely poor outcomes. This in turn reduced the utility of the results for purposes of quality assurance.

The way to deal with the problem of small numbers is to assemble more information on each institution, that is, collect outcome data on more

cases. This could involve aggregating cases into larger diagnostic clusters, accepting greater clinical heterogeneity for increased statistical stability. However, for analyses of overall mortality, the number of cases analyzed can only be increased by expanding the time period under investigation. HCFA has already constructed comparable data files for 1984 and 1985 discharges, so little additional effort would be needed to add these cases to the analysis.

Of course, many hospitals treat substantial numbers of Medicare patients in a given year. For these hospitals, there is no reason to exclude separate analyses for individual years, or possibly even shorter intervals. This would enable HCFA to monitor changes in outcomes over fairly short time periods, as long as there were sufficient numbers of cases to analyze effectively.

Thus, an alternative to HCFA's exclusive reliance on one year's data would be a more flexible approach that combined single-year and multi-year analyses. For smaller hospitals and diagnostic categories with low mortality rates, cases from several years could be pooled to provide an observed mortality rate less influenced by random variation than that derived from data for any single year. With more cases, HCFA's formula for calculating the range of predicted hospital mortality rates would produce a narrower interval. This, in turn, would make it more likely that the analysis would identify as outliers those hospitals whose observed mortality differed markedly from expected as a result of poor-quality care rather than random fluctuation.

Alternatively, for larger hospitals and more risky diagnostic groupings (with larger numbers of deaths to analyze), results for separate years, or other intervals, could be computed and compared to see if trends or patterns had developed over time. This approach has the advantage of highlighting what the current or most recent performance level has been, while placing it in the context of the institution's previous experience.

## Data Quality

HCFA performs a variety of data edits as bills are processed for payment. To a large extent, these involve checks for consistency and logic.[20] PROs also assess the accuracy of DRG assignments for the cases that they review. However, we found no attempt within HCFA to measure and

[20]For more detail, see our report on strategies for assessing quality of care in Medicare: Improving Quality of Care Assessment and Assurance, GAO/PEMD-88-10 (May 2, 1988).

record the specific percentage of cases in their administrative files for which specific data elements were accurate and complete, particularly in comparison to such primary data sources as patient medical records. The only information of this type that we have located derives from the studies of the reliability of hospital discharge abstract data performed by the Institute of Medicine (IOM) on 1974 hospital discharges. It found a high level of completeness and accuracy in Medicare records for such items as sex and admission or discharge dates, but substantially lower levels for diagnoses and procedures.[21] How much the current situation differs from what the study found more than a decade ago, given the introduction of a new payment system based more directly on diagnostic and procedural data, has been the subject of much speculation but, to our knowledge, of no empirical study within or outside of HCFA.[22]

In addition, evidence of undercounting of inpatient deaths in the Medicare files has raised particular concerns about the use of those data to measure patient outcomes, as HCFA did in its 1986 analyses.[23] HCFA has since added a new data element, date of death as reported to the Social Security Administration, to the data file typically used for hospital outcome analyses.[24] This new element not only permits analyses including deaths outside the hospital—and therefore was used in the 1987 mortality analyses—but also benefits from Social Security's administrative incentives to validate this information. However, as Social Security benefits are paid monthly, the accuracy of the reported day in the month is not known. Some empirical data on the relative accuracy of different data sources for date of death should be available in the next few years

[21]Institute of Medicine, Reliability of Medicare Hospital Discharge Records (Washington, D.C.: National Academy of Sciences, Nov. 1977), pp. 23-24.

[22]The HHS Inspector General has a study underway examining medical records for a large, representative sample of Medicare patients discharged in 1984 and 1985. However, this study focuses more on the accuracy of DRG assignments than on diagnostic codes per se (many inaccurate diagnostic codes would not change the DRG). Thus, it is not yet clear how much this study will clarify the issue of diagnostic coding accuracy. In addition, Rand researchers have recently reported a high level of accuracy in the coding of three specific procedures in Medicare files based on physician, rather than hospital, claims for payment. These data involved procedures performed in 1981, so they also predate the implementation of PPS. See Jacqueline Kosecoff et al., "Obtaining Clinical Data on the Appropriateness of Medical Care in Community Practice," The Journal of the American Medical Association, vol. 258, no. 18 (Nov. 13, 1987), pp. 2540-42.

[23]Blumberg, "Comments on HCFA Outliers," pp. 716-18; see also, California Medical Review, Inc., Premature Discharge Study: Final Report, Typescript (San Francisco: Dec. 10, 1986), which found a 23-percent error rate in the recorded "patient disposition" for the Medicare patients examined in this study. Most of these errors involved patients recorded as discharged alive who actually died in the hospital.

[24]MEDPAR—the Medicare Provider Analysis and Review file, which consolidates information on all hospitalizations for individual patients over a 3-year period.

from the Nonintrusive Outcomes Study currently funded by HCFA. (See chapter 4.)

## The 1986 Analyses

HCFA undertook no separate efforts to assess the completeness and accuracy of the data from Medicare administrative files used in the 1986 outlier analyses and did not make adjustments for possible data deficiencies. One indication that missing data could have affected the analyses comes from the PROs, which were charged with validating the HCFA outliers in their state as part of their contract negotiations for the 1986-1988 cycle. In our interviews with PRO officials about that process, many stated that they were unable to confirm the results of the HCFA analyses using their own data bases. Moreover, several observed that in some cases their data showed substantial differences in the raw numbers of discharges or deaths in the outlier hospitals compared to what HCFA reported. This occurred even though HCFA had analyzed 1984 hospital discharges, all of which should have been completed well over a year before the analyses were conducted.

## The 1987 Analyses

For the 1987 mortality analyses, HCFA asked hospitals to comment individually on the data it used to assess their own outcomes before those analyses were released. Each hospital received an analysis of its actual mortality rate and the range of predicted mortality rates, plus detailed data on each Medicare patient included in that analysis. HCFA did not attempt to correct its data base using the comments it received from about 2,600 of the 5,971 hospitals analyzed. Instead, all comments submitted within the allotted 30-day period were published together with HCFA's results. Presumably, readers interested in individual hospitals could weigh the hospital's assessment of data accuracy as well as other factors noted by the hospital in interpreting HCFA's figures for that institution. However, such comments do not provide the kind of systematic and consistent checking of data elements needed to judge the overall accuracy of the data on which the analysis was based.

In the introduction to its release of the 1987 hospital analyses, HCFA noted the potential for data inaccuracies but generally minimized their probable importance. For example, while noting "the possibility" of coding errors in the bills submitted by individual hospitals, from which the MEDPAR data base was derived, HCFA stated that "the assumption is that hospitals do not make errors on billing forms which ultimately affect payment to them." HCFA also indicated that the MEDPAR file did not fully reflect corrections in "diagnosis assignment" made by the PROs in the

cases they reviewed, but it did not assess the extent of error thereby introduced into the analyses. Finally, HCFA estimated that the MEDPAR file contained information on 95 percent of the cases discharged in 1986.[25]

## Validation

HCFA did not attempt to analyze the overall results of the PROs' review, as part of their contract negotiations, of the outlier hospitals identified in the 1986 analyses. (See chapter 3). However, the cumulative assessment of the separate PRO evaluations of those outliers could be considered tentative information from independent data sources on the effectiveness of the approach in identifying patterns of genuine quality problems. As we note in chapter 3, the lack of specific guidance from HCFA to the PROs resulted in considerable variability in the way PROs conducted their reviews. Nevertheless, they substantiated only a small number of the HCFA outliers as having real quality problems.

The 1987 hospital analyses received less validation then the 1986 analyses, even though HCFA had planned in advance to release the results to the public. Moreover, whatever HCFA may have learned about the validity of the 1986 analyses had limited relevance for these new analyses, since in most essential respects the 1987 analyses diverged markedly from the previous year's efforts. The 1987 analyses used different factors to adjust for patient severity and different statistical techniques to perform that adjustment and control for random variation. In addition, the 17 diagnostic clusters were entirely new, never used in any outcome analyses prior to these.

Many of the changes represent improvements over the 1986 approach; however, by the time all the various components of the 1987 analyses were decided upon in August 1987, there was little time left to validate the approach as a whole. To give hospitals a chance to comment and to enable HCFA to include these comments in the report scheduled for release in December 1987, HCFA decided to send the preliminary results to the hospitals in September. In the period between August and September, HCFA's validation efforts primarily involved the statistical analyses described above and the solicitation of comments through the Federal Register notice outlining its planned approach.[26]

[25]HCFA, Medicare Hospital Mortality Information: 1986, "Information Sources and Notes," p. 5.

[26]52 Fed. Reg. 30741-745 (1987).

The comments from the hospitals could provide some information about how the technique actually works in practice. However, hospitals that lacked the necessary methodological expertise would find it difficult to critically examine the statistical analyses that HCFA employed. For the most part, their comments focused on specific instances of inaccurate data for individual cases and general observations about factors such as the presence of an oncology ward or a high proportion of admissions from nursing homes that might affect the results. In any case, these comments came too late to influence the way HCFA conducted its analyses and serve primarily to guide interpretation of individual hospital ratings.

## Overall Assessment

The 1987 hospital mortality analyses represent a substantial improvement over those HCFA released in 1986. Although the 17 diagnostic clusters used to adjust for differences in case-mix were relatively broad and did not take into account substantial differences in expected mortality among the principal diagnoses they contain, these clusters were clinically coherent and were accompanied by a separate adjustment for a number of prevalent comorbidities. Moreover, the application of logistic regression to patient-level data to make the severity adjustment within the diagnostic categories allowed greater analytical precision than in the 1986 analyses, which were based on hospital averages for patient age, sex, and so on.

The ranges of expected mortality reported for each hospital in the 1987 analyses both avoided the problems created by the inappropriate use of the t-test in the earlier analyses and provided more information than a simple designation of a hospital as an outlier or not. However, the decision to analyze just one year of data accentuated the impact of random variation on the results and limited the ability of the approach to detect genuinely poor outcomes among smaller hospitals. The approach also received only limited validation prior to its application in a national assessment of individual hospital performance.

## HSQB Monitoring Systems

HCFA has analyzed the same data sets it used to investigate mortality rates in individual hospitals to compare different patient populations and program trends along a series of outcome dimensions. Extramural researchers funded through HCFA's Office of Research and Demonstrations have performed some of this work, particularly on changes over

time. (See chapter 4.) In this section, we describe the analyses performed by HCFA staff within its Health Standards and Quality Bureau (HSQB), which administers the PRO program.[27]

## General Purpose

Unlike the hospital mortality analyses, which focused on the performance of individual providers, these analyses address questions relating to variations in outcomes among patient subgroups and trends over time for the Medicare program as a whole.

## Substantive Focus

Among the patient subgroups specifically compared in the HSQB monitoring systems are those defined by sex, race, and age, as well as cases with certain principal diagnoses or comorbidities.[28] Many of the analyses done to date have examined the relative risks of mortality (and other outcomes) among these groups. Those analyses show that the relationship of demographic and clinical factors to the probability of adverse outcomes varies for the nine conditions and procedures examined (the same ones used for the 1986 hospital mortality analyses), which represent just over 30 percent of total Medicare discharges.

The HSQB analyses of trends over time focus on the issue of whether mortality rates for all Medicare discharges (and to a lesser extent readmission rates) have changed since 1984, following the introduction of the prospective payment system and the Peer Review Organizations. They also analyze the components of that change. Thus, increases in mortality rates following hospitalization are contrasted with largely unchanged mortality for the Medicare beneficiary population as a whole. HSQB has further analyzed postadmission mortality to see whether changes in the demographic characteristics of the patient population and in the distribution of cases among certain specific major diagnoses are associated with changes in mortality.

[27]Our description relies primarily on an unpublished paper summarizing these analyses: Henry Krakauer, "Outcomes of In-hospital Care in 1983-1985: The Medicare Experience," typescript, Office of Medical Review, HSQB (Baltimore, no date).

[28]Three different sets of somewhat overlapping conditions are analyzed at different points: seven "major conditions" (malignancy, myocardial infarcts and failure, cerebrovascular accidents, sepsis, trauma, pulmonary disease, renal disease), nine "tracer conditions" (congestive heart failure and shock, acute myocardial infarction, pneumonia, gastrointestinal hemorrhage, cholecystectomy, major joint procedures, transurethral prostatectomy, coronary artery bypass surgery, and pacemaker implants), and six "causes of readmission" (pulmonary, cardiovascular, gastrointestinal, musculoskeletal, genitourinary, and complications of treatment). The comorbidities included in the analyses are malignancy, ischemic heart disease and failure, chronic pulmonary disease, chronic renal disease, chronic liver disease, degenerative cerebral disease and psychosis, hypertensive disease, diabetes, and autoimmune disease.

## Severity Adjustment

A variety of factors are used to adjust for patient severity, including age, sex, race, and the presence of particular comorbidities. For the comparisons across patient subgroups, the results are expressed in terms of the relative effect of each factor (holding the effect of other factors constant) on outcomes within the given disease categories. For example, everything else being equal, 80-year-old men are 2.7 times as likely to die from a transurethral prostatectomy as 65-year-olds. Only the analysis of aggregate trends over time for the entire Medicare population simultaneously adjusts outcomes for both case-mix (i.e., changes in the relative frequency of different diagnoses in the Medicare population as a whole) and a range of demographic variables.

The cautions raised earlier regarding statistical adjustments for race should be considered for both the analyses across years and across patient subgroups. However, this concern would apply only to the extent that treatment of specific ethnic groups changed systematically across the nation over a period of a few years or varied systematically for different medical conditions. This seems somewhat less likely than systematic variation across individual hospitals.

The conditions considered as comorbidities are largely chronic in nature (possibly excluding some of the diagnoses included under ischemic heart disease and failure), thereby supporting the inference that they describe debilitating conditions that probably occurred prior to the hospital admission of interest. However, these comorbidity variables were only incorporated in the analysis of factors affecting the mortality rates of patients with one of the nine selected conditions or procedures. The comparison of overall mortality between the 1984 and 1985 groups takes account of changes in case-mix; that is, in principal diagnoses, but not comorbidities

## Technical Adequacy

With respect to measurement issues for the outcomes of interest, all these analyses investigate mortality and readmission in terms of time periods following admission, rather than, for example, inpatient deaths. These survival analyses use the actuarial or life-table method to calculate death and readmission rates across a range of time intervals, generally 30 to 360 days. This approach has the advantage of displaying the actual pattern of mortality or readmission over an extended period of time. By avoiding a single threshold (e.g., 30 days) for assessing post-admission outcomes, survival analyses ameliorate some of the disadvantages described earlier for using fixed time intervals to define the

outcomes. Nevertheless, the conceptual link between the quality of hospital treatment and the outcome becomes increasingly tenuous as the length of time from hospital discharge to death grows longer.

This set of analyses also attempts to overcome the limitations of mortality in general as an outcome (notably its lack of sensitivity) by examining morbidity and disability outcomes as well. However, the measures judged to be feasible using the data elements available in the Medicare Statistical System raise additional issues, primarily related to validity. In other words, what is actually examined in the analyses may not closely correspond to the core concepts represented by the terms "morbidity" and "disability."

For example, the analyses assess morbidity in terms of hospital readmissions and the costs of ambulatory care subsequent to the hospitalization of interest. The analyses do distinguish between readmissions for "related causes" (involving the same organ system) and other readmissions, and by using the actuarial method, avoid relying on a single threshold (e.g. within 30 days) to define relevant readmissions. Nonetheless, readmissions intrinsically reflect medical or administrative decisions as well as patient need. They may serve as an indicator of patient health status, which if related to a prior hospitalization can be considered an outcome.

However, health status is only one of several factors affecting the decision on whether to hospitalize a patient. For many conditions there are wide variations among physicians in their proclivity to hospitalize patients with comparable health problems and wide differences among patients along nonclinical dimensions that affect the appropriateness of treatment alternatives outside the hospital. Moreover, multiple admissions may constitute a preferred course of treatment for certain types of cases, as opposed to a single prolonged hospital stay. In short, the fact that one patient was readmitted to a hospital and another was not provides only limited information about their relative health status, and therefore by inference, about the outcome of their previous hospitalizations.

The same point applies even more strongly to comparisons based on costs of ambulatory care. Patients vary widely in their demand for health services, while physicians differ in the intensity of services that they perform for comparable cases. Moreover, the differences in the

charges assessed by physicians for ambulatory services contribute substantial additional error to using Medicare costs as an indicator of morbidity. Similarly, charges to Medicare for days spent in skilled nursing facilities or for home health services reflect as much or more the provisions of Medicare regulations on eligibility for benefits and the availability of such services in a given geographic area as they do an individual's degree of disability.

HSQB employed the Cox proportional hazards model to assess the relative impact of specified factors on the observed pattern of patient mortality. This analytical technique makes three key assumptions: (1) that the separate factors being analyzed have in combination a multiplicative effect on the outcome, (2) that categories within a variable have a proportional effect on the outcome, and (3) that the magnitude of the relationship of these factors to the outcome is constant across the time period under study.

The last assumption may under certain circumstances be somewhat problematic with respect to one of the demographic variables entered into the model: age. As the follow-up period extends in years, the normal probability for an 85-year-old dying in that period increases at a more rapid rate from year to year than does the probability of death for a 65-year-old. This means that the relative risk associated with age is not constant, but increases over time. For the age range representing most Medicare beneficiaries, 65 to 85 years, the risks associated with age change over time, but only marginally, for the 2- to 3-year follow-up period that HCFA so far has used. However, if the follow-up period was extended substantially, or if analyses focused on the oldest cohorts of Medicare beneficiaries whose probability of death increases most markedly from one year to the next, this issue could become more salient.

## Data Quality

HCFA used essentially the same data files for these analyses as for the hospital mortality analyses, which makes our earlier discussion of data quality relevant here as well. Like the 1987 study, this work employed the information on date of death supplied by the Social Security Administration.

## Validation

HCFA has not specifically validated this approach. However, it has underway a pilot project in which eight PROs collect clinical information on a sample of approximately 3,000 cases divided among six different conditions. Since five of these correspond to conditions examined under

the HSQB monitoring systems, this project should provide HCFA at least some preliminary evidence on how the results of these outcome analyses compare to findings based on medical record reviews.

## Overall Assessment

Although the scope of the analyses undertaken to date has been limited, the HSQB monitoring systems demonstrate the potential for examining Medicare outcome data to address issues other than individual hospital performance. One major strength of this approach derives from its use of survival analyses to follow outcomes over an extended period of time. While not entirely free of controversy, since deaths that occur long after discharge from the hospital are less plausibly related to the quality of hospital care, this method avoids the loss of information inherent in setting a single threshold, such as 30 days following hospital admission.

Much more problematical are the efforts to analyze morbidity and disability, which use proxy measures from the Medicare data files that bear little relation to the core concepts involved. Moreover, the severity adjustment used so far in this approach is quite constrained, making only restricted use of information on comorbidities and principal diagnosis to structure comparisons across a limited number of groups. Nonetheless, the mortality analyses represent an important first step in identifying subgroups of Medicare patients who experience substantial variations in outcomes. The results of the ongoing pilot project should provide useful guidance on ways to expand and improve this approach.

Chapter 3

# HCFA's Use of Outcome Analyses

The 1986 hospital mortality analyses constitute HCFA's major use of outcome data to guide its quality assurance activities. Consequently, this chapter focuses almost exclusively on that experience. We also briefly discuss HCFA's use of the HSQB monitoring systems.

## The 1986 Hospital Mortality Analyses

### The PROs' Role in Validating Hospital Outliers

Each organization responding to the 1986 request for proposal to serve as the PRO in one of the 50 states and the District of Columbia was required to "verify" or evaluate the lists of overall mortality and DRG-specific hospital outliers generated in the March 1986 analyses for that state.[1] Each was supposed to investigate not only hospitals with relatively high mortality rates, but also those with statistically significant lower-than-predicted death rates, to see if their low mortality reflected patterns of premature discharges or inappropriate transfers. The results of these analyses then entered into the negotiations between HCFA and the would-be PROs over the formulation of specific contractual objectives to improve quality of care by reducing the incidence of adverse outcomes, both overall and for specific DRGs.[2]

We found, however, that HCFA did not analyze the utility of these negotiations with respect to either the proportion of outlier hospitals "verified" by inclusion in quality of care objectives or the utility of the lists for focusing PRO contract negotiations on quality of care issues. Therefore, we interviewed officials at each of the funded PROs to obtain a description of its validation process and results.[3]

Our analysis focused on the 10 final lists of mortality outliers released by HCFA on March 10, 1986, one for all Medicare patients and nine for specific DRGs.[4] Altogether, there were a total of 2,313 outliers, including

[1]The analyses did not include American Samoa/Guam, Puerto Rico, or the Virgin Islands.

[2]HCFA also asked the would-be PROs to evaluate a list of utilization outliers—hospitals whose number of Medicare discharges per acute care bed, overall and for nine specific DRG clusters, were significantly greater or less than expected based on the same type of regression analysis as the mortality outlier analysis. However, we only examined the evaluation of the mortality outliers.

[3]Interviews were conducted in Aug. and Sept. 1987, with follow-up continuing until Dec.

[4]Some PROs received preliminary lists of outliers from HCFA that diverged somewhat from the March 10 list, but our analyses focused on the final list that the PRO investigated.

1,970 with higher-than-expected mortality rates ("high outliers") and 343 with lower-than-expected mortality rates ("low outliers"). Because 426 hospitals appeared on more than one list (including three hospitals in New York that appeared on 6 of the 10), the total number of hospitals identified by the outlier analyses was 1,714.

The requirements of the request for proposal were not carried out in every case. About 26 percent of the 2,313 outliers were never evaluated, although only 7 percent of the overall outliers were missed. Table 3.1 shows the number of outliers evaluated by the PROs for each of the specific DRGs.

**Table 3.1: Number of HCFA's 1986 Hospital Mortality Outliers Evaluated by the PROs**

| | | Evaluated | | Not evaluated | |
|---|---|---|---|---|---|
| **Outlier category** | **Total** | **No.** | **%** | **No.** | **%** |
| Overall mortality | | | | | |
| High outliers | 142 | 141 | 99 | 1 | 1 |
| Low outliers | 127 | 109 | 86 | 18 | 14 |
| **Total** | **269** | **250** | **93** | **19** | **7** |
| Congestive heart failure | | | | | |
| High outliers | 213 | 131 | 62 | 82 | 38 |
| Low outliers | 55 | 36 | 65 | 19 | 35 |
| **Total** | **268** | **167** | **62** | **101** | **38** |
| Acute myocardial infarction | | | | | |
| High outliers | 180 | 121 | 67 | 59 | 33 |
| Low outliers | 115 | 80 | 70 | 35 | 30 |
| **Total** | **295** | **201** | **68** | **94** | **32** |
| Pneumonia | | | | | |
| High outliers | 214 | 139 | 65 | 75 | 35 |
| Low outliers | 39 | 27 | 69 | 12 | 31 |
| **Total** | **253** | **166** | **66** | **87** | **34** |
| Gastrointestinal hemorrhage | | | | | |
| High outliers | 266 | 197 | 74 | 69 | 26 |
| Low outliers | 4 | 2 | 50 | 2 | 50 |
| **Total** | **270** | **199** | **74** | **71** | **26** |
| Cholecystectomy (all high) | **270** | **202** | **75** | **68** | **25** |
| Major joint procedures (all high | **220** | **155** | **70** | **65** | **30** |
| Transurethral prostatectomy (all high) | **253** | **191** | **75** | **62** | **25** |
| Coronary artery bypass surger | | | | | |
| High outliers | 33 | 27 | 82 | 6 | 18 |
| Low outliers | 3 | 3 | 100 | 0 | 0 |
| **Total** | **36** | **30** | **83** | **6** | **17** |
| Pacemaker implants (all high) | **179** | **147** | **82** | **32** | **18** |
| **Total DRG-specific** | | | | | |
| **High outliers** | **1,828** | **1,310** | **72** | **518** | **28** |
| **Low outliers** | **216** | **148** | **69** | **68** | **31** |
| **Total** | **2,044** | **1,458** | **71** | **586** | **29** |
| **Grand total** | **2,313** | **1,708** | **74** | **605** | **26** |

In addition, the role played by the outlier lists in the PRO contract negotiations varied a great deal. For a number of PROs whose contracts came up for negotiation early in the cycle, agreement had already been reached before they received the list to analyze. In some states, HCFA accepted the PRO analysis with little discussion, while in others, HCFA's negotiating team required multiple revisions of the analysis before agreement on the quality of care objectives could be reached. Thus, some hospital outliers were included in quality of care objectives based on the analysis and judgment of the PRO, while others were included primarily at the insistence of HCFA officials.

HCFA specified no standard methodology for analyzing the outlier lists. The scope of work in the request for proposal simply called on the proposers "to verify which hospitals on the list, on the basis of the offeror's data or knowledge of the medical events and practices in its jurisdiction, merit intensified review." As a result, PROs adopted a number of different analytical strategies, depending on their capabilities and circumstances. Some of the successful proposers had no previous experience in that state; they typically had no data upon which to base an evaluation of the outliers, and so they frequently incorporated all or most of the HCFA outliers in their objectives, at least provisionally. Among the incumbent PROs, some attempted to confirm HCFA's analyses using their own data on hospital admissions and mortality rates (some using HCFA's regression equation and some not), while others focused on the results of past reviews of cases discharged from those hospitals in the previous PRO contract. Still other PROs relied heavily on detailed descriptions of the particular characteristics of the individual outlier hospitals; for example, noting the presence of an oncology treatment unit, a burn center, or a hospice ward, which might tend to explain unusual mortality rates.

## Results of PRO Reviews

Overall, the PROs reported that their review of the hospital outliers identified by HCFA's 1986 analyses revealed very few hospitals with substantial quality problems. Given the diverse analytical approaches taken by the PROs, and the varying scrutiny of those evaluations applied by different teams of HCFA negotiators, the aggregate results of the PRO analyses need to be interpreted cautiously, essentially in terms of broad patterns. Moreover, the nature of our data do not allow us to estimate the extent to which the limited yield from this effort resulted from deficiencies in the PRO review, as opposed to deficiencies in HCFA's mortality analyses or the data that were analyzed. Nevertheless, these tentative results provide the only available evidence on the potential utility of such outcomes analyses for targeting PRO reviews.

The large majority of HCFA outliers examined by the PROs during the contract negotiation process were not included in the PROs' quality of care

objectives, or selected for intensified review. In other words, for these hospitals the PROs determined that their own evidence (derived from past review results, mortality rates in previous and succeeding years, and so on) did not support the inference of suspected quality problems conveyed by the hospitals' outlier status. Table 3.2 presents the cumulative determination of the 51 PROs for each of the 10 lists of outliers. Of the 1,708 outliers examined by the PROs, only 350 (20 percent) were incorporated into an objective. Another 11 outliers were targeted for intensified review without setting specific goals for improved performance.

**Table 3.2: Results of PRO Evaluation of HCFA's 1986 Hospital Mortality Outliers**

| Outlier category | Total evaluated | Rejected for objective | | Included in objective | | Intensified review | |
|---|---|---|---|---|---|---|---|
| | | No. | % | No. | % | No. | % |
| Overall mortality | | | | | | | |
| High outliers | 141 | 76 | 54 | 65 | 46 | 0 | 0 |
| Low outliers | 109 | 90 | 83 | 17 | 16 | 2 | 2 |
| **Total** | **250** | **166** | **66** | **82** | **33** | **2** | **1** |
| Congestive heart failure | | | | | | | |
| High outliers | 131 | 100 | 76 | 31 | 24 | 0 | 0 |
| Low outliers | 36 | 33 | 92 | 3 | 8 | 0 | 0 |
| **Total** | **167** | **133** | **80** | **34** | **20** | **0** | **0** |
| Acute myocardial infarction | | | | | | | |
| High outliers | 121 | 91 | 75 | 30 | 25 | 0 | 0 |
| Low outliers | 80 | 56 | 70 | 22 | 28 | 2 | 2 |
| **Total** | **201** | **147** | **73** | **52** | **26** | **2** | **1** |
| Pneumonia | | | | | | | |
| High outliers | 139 | 98 | 70 | 41 | 30 | 0 | 0 |
| Low outliers | 27 | 25 | 93 | 2 | 7 | 0 | 0 |
| **Total** | **166** | **123** | **74** | **43** | **26** | **0** | **0** |
| Gastrointestinal hemorrhage | | | | | | | |
| High outliers | 197 | 153 | 78 | 44 | 22 | 0 | 0 |
| Low outliers | 2 | 2 | 100 | 0 | 0 | 0 | 0 |
| **Total** | **199** | **155** | **78** | **44** | **22** | **0** | **0** |
| Cholecystectomy (all high) | **202** | **173** | **86** | **24** | **12** | **5** | **2** |
| Major joint procedures (all high) | **155** | **135** | **87** | **20** | **13** | **0** | **0** |
| Transurethral prostatectomy (all high) | **191** | **162** | **85** | **27** | **14** | **2** | **1** |
| Coronary artery bypass surgery | | | | | | | |
| High outliers | 27 | 22 | 81 | 5 | 19 | 0 | 0 |
| Low outliers | 3 | 3 | 100 | 0 | 0 | 0 | 0 |
| **Total** | **30** | **25** | **83** | **5** | **17** | **0** | **0** |
| Pacemaker implants (all high) | **147** | **128** | **87** | **19** | **13** | **0** | **0** |
| **Total DRG-specific** | | | | | | | |
| **High outliers** | **1,310** | **1,062** | **81** | **241** | **18** | **7** | **1** |
| **Low outliers** | **148** | **119** | **80** | **27** | **18** | **2** | **1** |
| **Total** | **1,458** | **1,181** | **81** | **268** | **18** | **9** | **1** |
| **Grand total** | **1,708** | **1,347** | **79** | **350** | **20** | **11** | **1** |

The proportion of outlier hospitals examined by the PROs that ultimately were incorporated into review objectives or subjected to intensified review ranged from 46 percent of the overall mortality outliers with higher-than-expected death rates to none for low outliers from several of the DRG-specific lists. The PROs generally did not incorporate low outliers into their objectives, aside from the overall mortality analysis, where 19 (18 percent) of the HCFA low outliers they evaluated were included in objectives or selected for intensified review. The other main exception was for acute myocardial infarction (DRG 121-123), where a single state—California—was responsible for 18 of the 24 low outlier hospitals selected, producing an overall inclusion rate nationwide of 30 percent. For high mortality outliers from the DRG-specific analyses, rates of inclusion in quality of care objectives or intensified review ranged from 13 to 30 percent.

We also asked the PRO officials we interviewed how the 361 hospital outliers targeted under quality of care objectives or selected for intensified review had fared since the new contract period had begun. The results of the case reviews conducted by the PROs for both groups of hospitals appear in table 3.3. They indicate no quality problems for almost half (49 percent) of the outliers. For another 39 percent, insufficient data had been collected to judge one way or the other.

Across all the PROs and all the outlier analyses, only 13 hospital outliers identified on HCFA's lists were reported to have shown definite quality problems continuing into the new contract period, with another six showing possible problems.[5] Five of the 13 appeared on HCFA's list of outliers for acute myocardial infarction (DRG 121-123). Three appeared on the list of overall mortality outliers, while the other five were scattered among another four DRG lists. There were nearly equal numbers of high (seven) and low (six) outliers with confirmed quality problems.

[5]As noted above, the same hospital could appear on more than one DRG list. Of the 13 confirmed hospital outliers, 3 represent the same hospital for separate DRGs.

**Table 3.3: Hospital Outliers Included in Contract Quality of Care Objectives or Subject to Intensified Review**

| Outlier category | Total reviewed | Definite quality problems | | Possible quality problems | | Likely no quality problems | | Definitely no quality problems | | Undetermined | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | No. | % | No. | % | No. | % | No. | % | No. | % |
| Overall mortality | | | | | | | | | | | |
| High outlier | 65 | 2 | 3 | 3 | 5 | 12 | 18 | 16 | 25 | 32 | 49 |
| Low outliers | 19 | 1 | 5 | 0 | 0 | 2 | 11 | 6 | 32 | 10 | 53 |
| **Total** | **84** | **3** | **4** | **3** | **4** | **14** | **17** | **22** | **26** | **42** | **50** |
| Congestive heart failure | | | | | | | | | | | |
| High outliers | 31 | 0 | 0 | 0 | 0 | 1 | 3 | 16 | 52 | 14 | 45 |
| Low outliers | 3 | 1 | 33 | 0 | 0 | 2 | 67 | 0 | 0 | 0 | 0 |
| **Total** | **34** | **1** | **3** | **0** | **0** | **3** | **9** | **16** | **47** | **14** | **41** |
| Acute myocardial infarction | | | | | | | | | | | |
| High outliers | 30 | 2 | 7 | 0 | 0 | 1 | 3 | 16 | 53 | 11 | 37 |
| Low outliers | 24 | 3 | 12 | 0 | 0 | 1 | 4 | 20 | 83 | 0 | 0 |
| **Total** | **54** | **5** | **9** | **0** | **0** | **2** | **4** | **36** | **67** | **11** | **20** |
| Pneumonia | | | | | | | | | | | |
| High outliers | 41 | 1 | 2 | 0 | 0 | 0 | 0 | 16 | 39 | 24 | 59 |
| Low outliers | 2 | 1 | 50 | 0 | 0 | 0 | 0 | 1 | 50 | 0 | 0 |
| **Total** | **43** | **2** | **5** | **0** | **0** | **0** | **0** | **17** | **40** | **24** | **56** |
| Gastrointestinal hemorrhage | | | | | | | | | | | |
| High outliers | 44 | 1 | 2 | 0 | 0 | 0 | 0 | 19 | 43 | 24 | 55 |
| Low outliers | 0 | | | | | | | | | | |
| **Total** | **44** | **1** | **2** | **0** | **0** | **0** | **0** | **19** | **43** | **24** | **55** |
| Cholecystectomy (all high) | **29** | **1** | **3** | **0** | **0** | **0** | **0** | **23** | **79** | **5** | **17** |
| Major joint procedures (all high) | **20** | **0** | **0** | **2** | **10** | **0** | **0** | **15** | **75** | **3** | **15** |
| Transurethral prostatectomy (all high) | **29** | **0** | **0** | **1** | **3** | **1** | **3** | **15** | **52** | **12** | **41** |
| Coronary artery bypass surgery | | | | | | | | | | | |
| High outliers | 5 | 0 | 0 | 0 | 0 | 2 | 40 | 2 | 40 | 1 | 20 |
| Low outliers | 0 | | | | | | | | | | |
| **Total** | **5** | **0** | **0** | **0** | **0** | **2** | **40** | **2** | **40** | **1** | **20** |
| Pacemaker Implants (all high) | **19** | **0** | **0** | **0** | **0** | **2** | **11** | **13** | **68** | **4** | **21** |
| **Total DRG-specific** | | | | | | | | | | | |
| **High Outliers** | **248** | **5** | **2** | **3** | **1** | **7** | **3** | **135** | **54** | **98** | **40** |
| **Low Outliers** | **29** | **5** | **17** | **0** | **0** | **3** | **10** | **21** | **72** | **0** | **0** |
| **Total** | **277** | **10** | **4** | **3** | **1** | **10** | **4** | **156** | **56** | **98** | **35** |
| **Grand total** | **361** | **13** | **4** | **6** | **2** | **24** | **7** | **178** | **49** | **140** | **39** |

## Analysis of the Results

The substantial proportion of outlier hospitals whose quality rating remained undetermined when we contacted the PROs reinforces the need for caution in interpreting these results. In addition, it is possible that some hospitals may have improved their performance as a result of being designated for special attention by the PROs. However, it seems unlikely that many hospitals would successfully identify and eliminate

the treatment practices responsible for poor outcomes before the PRO reviews could get underway.

Moreover, one would expect that a substantial number of outliers would not demonstrate quality problems, given the statistical nature of the exercise. For the 5,750 hospitals analyzed, a 95-percent confidence interval would, on average, produce about 144 high and 144 low outliers for each of the 10 analyses, based purely on random variation. The problems created by HCFA's use of a t-test to establish outlier status for smaller hospitals (discussed in chapter 2, p. 30) would increase the number of outliers expected by chance alone, particularly high outliers for some of the DRGs with relatively low mortality rates, such as gastrointestinal hemorrhage, cholecystectomy, major joint procedures, transurethral prostatectomy, and pacemaker implants. An examination of the totals reported in table 3.1 suggests that a substantial proportion of the outliers found in each category would be expected on the basis of random variation.[6]

The utility of the analysis therefore rests on the number of outliers with confirmed problems. A useful screening instrument would balance the costs of evaluating and rejecting outliers reflecting random variation (false positives) with the benefits derived from identifying hospitals with genuine quality problems (true positives). Determining a "reasonable" cost depends on the value attached to finding problem hospitals and the relative effectiveness (proportion of total "true" problem hospitals identified)[7] and efficiency (proportion of "true positives" to "false positives") of alternative methods. If a very high value is placed on the identification of each problem hospital, then use of a fairly ineffective and inefficient approach may be justified—in the absence of a better alternative. However, the lower the number of "true positives" relative to the total number of problem hospitals, and the more numerous "false positives" are compared to "true positives," the greater the likelihood that alternative methods could be developed that would identify an equivalent or larger number of problem hospitals more efficiently.

Since this was HCFA's initial effort to apply outcome analyses to quality assurance in the Medicare program, we do not know how much more effective or efficient any alternative approach might have been. But the

[6]The low number of outliers reported for coronary artery bypass surgery reflects the fact that relatively few hospitals perform this procedure.

[7]That is, the number of true positives divided by the sum of true positives and false negatives—the latter representing hospitals with genuine quality problems that were not identified by the analysis.

disproportion between outliers with confirmed or probable quality problems and outliers the PROs determined did not have problems (19 to 178)—even among the outliers which the PROs considered most likely to reveal quality problems—suggests that there may be potential benefits in actively developing and testing alternatives.

In sum, a very small number of targeted hospitals were confirmed by the PROs as having patterns of poor quality care. This suggests that HCFA's application of its 1986 outlier analyses in the PRO contract negotiations did not enable PROs to concentrate their quality assurance efforts on hospitals likely to reveal quality of care problems in subsequent PRO reviews. Our data cannot answer the question of why this occurred, or what alternative approaches might have been more productive. However, we believe that a careful investigation of both these issues should precede any future use of similar outcome analyses to target PRO reviews.

At the time our data collection was completed in December 1987, HCFA had not yet decided how its 1987 mortality analyses would be used, other than to publish them for public information purposes. In particular, their use in the administration of the PRO program remained undetermined.

## The HSQB Monitoring Systems

Different versions of HSQB's analyses of trends in aggregate Medicare outcomes over time and comparisons across selected patient subgroups have been circulated within HCFA's Office of Research and Demonstrations and Office of the Actuary as well as to officials responsible for managing the PRO program in the Health Standards and Quality Bureau. So far, the analyses have served primarily to provide background information to HCFA officials.

# HCFA's Extramural Approaches for Analyzing Outcome Data

Our review of HCFA's intramural analyses of Medicare outcomes identified some important improvements as well as limitations in the approaches that HCFA has used to date. In this section, we begin to examine the potential for further improvements by looking at research that HCFA itself has funded.

Of the numerous extramural research efforts currently supported by HCFA, primarily through its Office of Research and Demonstrations, we found and analyzed four that involved both the development and the application of relatively distinctive approaches for analyzing outcome data as they currently exist in Medicare's administrative data systems.[1] They are evaluated below, following the seven key issues outlined in chapter 1. The results of this review are summarized in chapter 6, table 6.1.

Each of these approaches has its own limitations (some arising from the fact that the purpose of the research may not necessarily be focused directly on the issues of interest here). None has been sufficiently tested to definitively demonstrate its overall strengths and weaknesses. Nevertheless, all have elements with the potential to enhance HCFA's analysis of Medicare outcomes. These positive features are highlighted in the overall assessment sections that conclude the description of each approach.

## Nonintrusive Outcomes Study

### General Purpose

The Nonintrusive Outcomes Study, conducted by the Rand Corporation, aims to establish an empirical link between analyses of outcomes using administrative data and the more detailed and subtle information on the process of care obtainable from medical records. The part of this study that relates most directly to our own is a series of analyses of administrative data designed to characterize different medical and surgical conditions in terms of the extent and distribution of adverse outcomes,

[1] Although other work that HCFA has underway may also contribute in this area, we have chosen to focus on these four studies as indicative of the approaches that HCFA is pursuing for this type of outcome analyses. See our report, Medicare: Improving Quality of Care Assessment and Assurance, GAO/PEMD-88-10 (May 2, 1988), for an extensive compilation of HHS-funded research related to quality of care.

largely mortality. The approach compares observed to expected mortality within specified diagnostic groups for individual hospitals. Those groups are in turn compared, based on variations in the patterns of hospital outcomes found within each. Thus both levels of analysis, within and across diagnostic groups, are grounded on interhospital comparisons.

Rand used this component of its study in its selection of two specific diagnostic groups—congestive heart failure and myocardial infarction—on which to focus the medical record review part of its analysis. However, these medical record reviews (which are still underway) fall outside the scope of our study, except as a potential future source of information on the accuracy of data elements in hospital abstracts and the validity of measures derived from administrative data.

## Substantive Focus

The Rand analysis of Medicare patient files was structured to select a few conditions for in-depth study for the larger Nonintrusive Outcomes project. Beginning with a file containing data on 1984 discharges for elderly Medicare patients (excluding the disabled and end-stage renal disease beneficiaries), Rand dropped some DRGs, consolidated others, and eliminated certain individual diagnoses in an effort to construct clinically homogeneous "candidate conditions" that would be suitable for the validation phase of the study. This process resulted in 48 diagnostic groups representing 48 percent of all Medicare discharges, including a mix of medical conditions and surgical procedures. The remaining 52 percent of Medicare discharges represent conditions that could not be consolidated into "clinically homogeneous" groups. They therefore differ systematically from those that Rand has analyzed in ways that could affect the relationship of outcomes to patient characteristics and the quality of care received.

Rand then analyzed variations in outcomes within each diagnostic cluster across hospitals. Those hospitals whose observed mortality rates exceeded their "expected" rates by a statistically significant margin were designated as outliers. Next, the Rand researchers compared the diagnostic clusters themselves in terms of the relative number of hospital outliers in each and the number of patients affected (i.e., the difference between the observed and the expected number of deaths summed across all the outlier hospitals). Rand examined several different outcomes, including readmissions and "total deaths" (inpatient plus 30 days after discharge), but largely its analysis focused on inpatient mortality.

It chose two conditions for medical record review from among those that demonstrated relatively high numbers of hospital outliers and affected patients. Aside from limited comparisons of hospital performance across states, Rand has not employed this approach to analyze patient or provider subgroups.

## Severity Adjustment

Rand deliberately limited the adjustment for patient condition in its analysis of administrative data to two demographic factors, age and sex, plus case-mix (i.e., inclusion in one of the 48 candidate conditions). Within each diagnostic cluster, Rand assessed each hospital's outcomes by the indirect standardization method. This involved a comparison of each hospital's observed mortality rate to an "expected" rate derived from the experience of Medicare patients as a whole, adjusted for the age and sex distribution of the patients treated by the hospital for that condition.

The study will generate much more detailed information on severity of illness through the medical record reviews of the sample of cases abstracted. Once it completes those reviews, Rand will be able to test the validity and reliability of a wide variety of severity measures (at least for those two conditions) using the data available on those patients in Medicare's administrative files.

## Technical Adequacy

Given the basic simplicity of Rand's adjustment for patient severity, the main technical issues relate to the statistical procedures Rand used to identify hospitals whose mortality rates exceeded that which would be expected through random variation. Rand did this in two stages. First, in each diagnostic cluster, Rand identified individual outlier hospitals by assessing the significance of the difference between expected and observed mortality rates. The test it employed assumed a binomial distribution, which is appropriate for this type of comparison. (See p. 30.)

Second, Rand assessed the overall pattern of outlier hospitals in each cluster to see if the total number of hospitals identified as outliers exceeded that which would be expected by chance. This required a different statistical significance test, the chi-square. In order to meet the assumption of large samples called for by chi-square tests, Rand simulated the observed mortality rates for each hospital 99 times to create an "empiric" chi-square distribution against which to apply the test.

The test showed that only 22 of the 48 diagnostic clusters had more outlier hospitals than were deemed likely to occur by chance 9 times out of 10. In other words, in looking for diagnostic clusters where outlier hospitals were most likely to represent genuine differences in outcomes rather than random fluctuations, 26 of the clusters showed no significant pattern of variation in outcomes among hospitals.

## Data Quality

The Rand study did not examine the impact of potential imperfections in Medicare data on its results. However, the medical record reviews of the two conditions will include verification of the completeness and accuracy of selected data elements in administrative files. For example, Rand will observe the accuracy of the date of inpatient deaths recorded in the hospital abstract compared to that derived from Social Security files.

## Validation

The medical record reviews of sampled cases will likewise provide extensive validation of the outcome analyses based on administrative data, but only for those two conditions. If an analytical approach employing administrative data is thereby validated, a method would be established that could be repeated through similar medical record reviews of other conditions.

## Overall Assessment

Because the Nonintrusive Outcomes approach to analyzing administrative data on Medicare outcomes was not designed to adjust for differences in patient severity, its relevance to our study focuses on several specific elements that might usefully be applied elsewhere. Thus, outcome analyses in which clinical homogeneity within diagnostic clusters was an important issue, but projection to the full spectrum of Medicare patients was not, could employ Rand's 48 candidate conditions. Moreover, the chi-square test Rand developed to test the distribution of outlier hospitals within diagnostic categories could aid HCFA and the PROs in focusing their quality of care reviews to patient subgroups where variation in outcomes represented more than random fluctuation.

Finally, the techniques that Rand has developed for medical record abstraction in the validation phase of its study could prove useful in validation efforts for other analytical approaches.

# The Risk-Adjusted Mortality Index

## General Purpose

As part of an ongoing analysis of Medicare quality of care funded by HCFA, the Commission on Professional and Hospital Activities (CPHA) has developed the Risk-Adjusted Mortality Index, or RAMI, an approach for comparing hospital performance using existing administrative data.[2] As its name implies, RAMI was designed to differentiate among patients on the basis of individual characteristics that increase or reduce their risk of dying in the hospital.

Drawing on CPHA's own data base assembled from 776 member hospitals, Susan DesHarnais et al. determined the overall expected risk of mortality associated with specific diagnoses and other patient characteristics within 310 DRG-based clusters for the total patient population. They then assessed the cumulative effect of these factors on mortality rates using two different statistical techniques. By aggregating the results for individual patients treated by particular hospitals, they evaluated the observed outcomes of those hospitals relative to expected outcomes based on the average performance of all the hospitals, both for specific diagnostic categories and total discharges.

## Substantive Focus

The Risk-Adjusted Mortality Index incorporates all medical and surgical conditions. These are broken down into 310 categories, which CPHA formed by consolidating DRGs that include the same diagnoses but are distinguished by other factors, such as age and presence of comorbidities. CPHA developed RAMI from a patient data base including all types of patients, not just Medicare. It was designed to analyze inpatient mortality, although parallel procedures for analyzing readmissions and surgical complications are under development.

To date, CPHA's analyses have focused primarily on comparisons among individual hospitals, either overall or for particular DRG clusters. However, many other subgroup comparisons would be feasible, since information on expected and observed mortality is derived from data on individual patients that can be aggregated any number of ways.

[2]Described in an unpublished paper by Susan DesHarnais, et al., "The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance" (Ann Arbor: CPHA, Sept. 28, 1987).

## Severity Adjustment

CPHA analysts derived the mortality probabilities used to compute expected mortality from the observed outcomes of over 6 million cases treated in CPHA hospitals in 1983. For DRG clusters with overall death rates of less than 5 percent, which includes 246 of the 310 clusters (and 84 percent of all patients), CPHA analyzed mortality in terms of a "contingency table model." This involved calculating the observed death rates of six subgroups within each cluster defined by three age levels (0-64, 65-74, 75+) and the presence or absence of comorbidities.

CPHA employed a more sophisticated approach for the 64 DRG clusters with a death rate of 5 percent or higher, building on data about each patient's principal diagnosis, major surgical procedures, all recorded secondary diagnoses, age, sex, and race. It constructed several different predictor variables from the diagnostic data. Two reflected the presence or absence of any secondary diagnosis at all and the presence of any secondary diagnosis of cancer (other than skin cancer). Three others involved calculations of risk scores, based on observed mortality rates among the 6 million CPHA discharges in 1983, for: (1) principal diagnosis, (2) principal surgical procedure (if any), and (3) highest risk comorbidity. Finally, CPHA counted the number of high-risk comorbidities. It calculated the risks associated with comorbidities from cases falling within particular classes of DRGs (major diagnostic categories), rather than averaging the effect of particular comorbidities across every medical and surgical condition.[3]

To determine the separate effect of each of these six predictor variables, plus patient age, sex, and race, on expected mortality, CPHA analyzed each of the DRG clusters using logistic regression procedures. The regression equations produced coefficients for each of the predictor variables, from which CPHA calculated the expected mortality rates for individual patients in that DRG cluster. Cumulating the expected and observed mortality of patients treated at individual hospitals produced expected and observed deaths and death rates for those hospitals. CPHA tested the statistical significance of the differences between the hospitals' expected and observed mortality by constructing confidence intervals based on the binomial distribution for DRG cluster-specific analyses and the Poisson approximation of a binomial distribution for the larger number of cases involved in a hospital-wide analysis.

[3]Where there were too few actual cases from which to derive a stable estimate of a comorbidity's mortality risk in a particular major diagnostic category, CPHA applied a Bayesian adjustment to the observed mortality rate based on the mortality experience for that comorbidity in other major categories.

CPHA's use of DRG clusters to structure its analysis raises the issue of DRG clinical heterogeneity. Unlike the 1986 HCFA analyses, which relied exclusively on selected DRGs for their case-mix adjustment, the CPHA approach also adjusts for the mortality risks associated with specific principal diagnoses and comorbidities. Thus, the predicted mortality for individual patients should not be distorted by the clinical heterogeneity within DRGs, at least for the high mortality conditions. However, for the low-risk DRG clusters, no comparable adjustment is made, so that some of the deaths in those clusters (which total 28 percent of all deaths) may reflect certain specific diagnoses with higher-than-average risks.

CPHA grappled explicitly with the problem of distinguishing comorbidities present at hospital admission from complications of treatment, given the ambiguities that characterize the way that secondary diagnoses are currently recorded. It identified 70 diagnoses from the HCFA complications and comorbidity list (used for making DRG assignments) that its medical consultant determined were most likely to represent true complications. These were excluded from consideration in constructing the predictor variables for expected mortality. In this way, the impact of identifiable complications of treatment was separated from the estimation of patient condition at admission.

However, there were also a number of other diagnoses that sometimes are complications and sometimes comorbidities, depending on when they occur. Pneumonia and urinary tract infections are two common examples. In those instances where CPHA could not determine whether the diagnosis typically represented a complication or comorbidity, it decided to count them as comorbidities. This choice means that when these diagnoses actually represent complications, the assessment of the outcomes of the responsible hospitals credits them instead with serving more severely ill patients.

## Technical Adequacy

The RAMI approach applies a relatively simple methodology to those DRG clusters where there is little variation in mortality to explain or predict. Therefore, it is the more sophisticated analysis of the clusters with higher mortality rates that raises more questions. For example, in attempting to go beyond the presence or absence of comorbidities in general and characterize the nature, number, and intensity of those comorbidities, CPHA has developed specific indicators constructed from the raw data on recorded secondary diagnoses. More information is needed to determine the validity of these particular indicators; that is, how well they represent the characteristics of comorbidities they are

intended to capture (breadth, intensity) compared to the wide range of potential alternative constructs.

A second issue concerns the potential bias introduced into comparisons of outcomes across hospitals, or other patient subgroups, that could derive from the interrelationship among these multiple measures of comorbidity. CPHA constructed these indicators from the same set of data on secondary diagnoses, thus the values assigned to these variables for individual patients tend to be correlated to one another, although the magnitude of this association varies across the different DRG clusters.[4] The main effect of this correlation among independent variables, known technically as multicollinearity, is to make the estimates of the regression coefficients generated for each of those variables more uncertain.

Since it is these coefficients that are used to calculate the expected mortality of individual patients, this greater uncertainty in their value could affect the results of outcomes analyses employing RAMI. It could, that is, if the errors thereby introduced into the estimation of individual mortality probabilities are not randomly distributed across the patient subgroups being compared. For example, if the regression coefficients systematically underestimate the effect of a risk factor for one category of patients and overestimate its effect for another and these patient categories are disproportionately admitted to different hospitals, then the comparison of hospital outcomes could be biased in favor of certain hospitals and against others. However, until the implications of these correlations among the diagnostic predictor variables is explored in detail for a range of major patient subgroups, we will not know the extent to which this potential bias actually affects the results produced by this approach.

Among the other predictor variables, the use of race raises the same questions that we have discussed with respect to other approaches. The limitations of inpatient mortality as an outcome measure have also been described above.

In sum, RAMI's use of logistic regression techniques, as well as binomial and Poisson tests of significance, seems appropriate as applied to the identification of individual outlier hospitals. However, CPHA conducted

[4]For example, if a patient has no secondary diagnoses at all (variable 4), that means that he, by definition, will also have no cancer-related secondary diagnoses (variable 5), and none that has a risk of mortality that exceeds the average mortality rate of the DRG cluster (variable 9).

no tests comparable to the chi-square simulation employed in the Nonintrusive Outcomes Study to see whether the overall pattern of outliers deviated significantly from random variation. The substantial number of outliers reported at the hospital level (13 percent high and 8 percent low using a 99-percent confidence interval) suggests that the probability that all these outliers simply represent random variation is fairly low, even without a formal test of significance. Still, it would be useful to know for the analysis of DRG clusters, how many and which ones showed patterns of outliers that exceeded the numbers expected owing to chance variation.

## Data Quality

CPHA has developed some broad information on the effect of data imperfections for RAMI analyses through its site visits to selected hospitals. For several hospitals with relatively large discrepancies between their observed and expected mortality rates, a major contributing factor proved to be incomplete recording of secondary diagnoses. Apparently, these hospitals only entered those diagnoses that affected reimbursement under Medicare's prospective payment system. As a result, RAMI systematically underestimated the severity of illness characterizing patients at these hospitals.

Currently, there are no data on the extent to which this problem prevails among hospitals overall or on the magnitude of its effect on RAMI-adjusted outcomes. While the institution of outcome analyses based on RAMI could, over a period of time, create incentives for more complete coding of secondary diagnoses, this problem would cloud the interpretation of a RAMI analysis of Medicare files as they now exist.

Although CPHA has reported no similar findings regarding the impact of inaccurate principal diagnosis codes, the critical role these data play in RAMI's severity adjustment suggests that any problems with these codes in the Medicare Statistical System (the potential for which is discussed in chapter 2) could also seriously affect its results.

## Validation

CPHA tested the RAMI procedures on a sample of 300 CPHA hospitals, selected to match the characteristics of general acute care hospitals in the United States as a whole. It has also applied RAMI to a 17-member group of nonprofit hospitals and is currently conducting site visits at several of these hospitals. During these visits, hospital administrators and physicians have assessed the validity of the "problem areas" identified by RAMI for their institution, based on their own quality assurance

activities and case reviews. CPHA reports that these hospital officials tend to confirm the accuracy and utility of the RAMI analysis.

The statistical tests of RAMI conducted by CPHA focus on its performance as a whole and show, most notably, a high degree of correspondence between predicted and actual mortality for all patients and across hospitals and diagnostic clusters within hospitals. Strictly speaking, the magnitude of deviation or convergence between expected and observed mortality reflects the characteristics of the population of hospitals being examined; specifically, the extent to which their outcomes differ from one to the other with a given set of adjustments. That in itself does not, in our view, demonstrate the appropriateness of those adjustments; that is, the degree to which they control for differences in patient condition and not for differences in hospital performance.

Arguing in favor of the CPHA severity adjustment is that all the factors used are by their nature patient characteristics; CPHA specifically excluded from consideration institutional characteristics such as a hospital's size or teaching status. Nevertheless, as discussed previously with respect to race (see chapter 2), if hospital performance varies systematically with one of those patient characteristics, statistical adjustment for that patient characteristic removes the component of hospital performance associated with it from the analysis. At this stage there is only limited corroboratory evidence—primarily the reports from the hospital site visits—showing that variation in RAMI-adjusted mortality rates reflects real differences in the quality of hospital treatment.

## Overall Assessment

The Risk-Adjusted Mortality Index shows promise as an alternative or supplement to HCFA's current practices in evaluating individual hospital performance. Its adjustment for patient severity takes advantage of available information on the specific risk of mortality associated with individual principal diagnoses and a wide range of comorbidities. To date CPHA has used the mortality index primarily for analyses of provider performance. The fact that the index employs patient-level data means that it could be flexibly applied to analyses of patient subgroups and change over time as well.

There are, nevertheless, some points of potential controversy that characterize this approach. They include the use of race as a predictor variable, the counting of possible complications as comorbidities, the use, so far, of inpatient mortality as the outcome, and concerns that interrelationships among some of the predictor variables might lead to biased

comparisons of outcomes, for example, among hospitals. More extensive testing of this approach than has occurred to date, particularly across patient subgroups, could help to allay these concerns.

# Disease Staging Adapted to Mortality Analyses

## General Purpose

HCFA has contracted with SysteMetrics, Inc., to apply its Disease Staging methodology to address the question of why the overall mortality rate of hospitalized Medicare beneficiaries increased between 1984 and 1985. Disease Staging is a system for assessing the severity of a principal or secondary disease—but not overall patient condition—based on formal judgments by physician panels of the clinical implications of particular combinations of diagnoses for specified disease categories. SysteMetrics has refined this system over a number of years and has developed a computerized version that can automatically "stage" patients based on data from standard hospital abstract files.

For HCFA, Disease Staging provided a way to evaluate the hypothesis that the increased mortality from 1984 to 1985 reflected an increase in the aggregate severity of patients admitted to hospitals, rather than a decline in quality of care. HCFA's Office of Research and Demonstrations contracted for this analysis specifically for its reports to Congress on the impact of the PPS system on Medicare beneficiaries. To accomplish this, SysteMetrics adapted the computerized version of Disease Staging to Medicare mortality data. It also developed a separate variable representing "unrelated high-risk comorbidities,"[5] which in conjunction with the stage of the principal disease permitted SysteMetrics to characterize the overall severity of illness for individual patients.

## Substantive Focus

Although primarily concerned with explaining trends in overall Medicare mortality over the 2-year period, the approach that SysteMetrics

[5]Only comorbidities that do not influence the Disease Staging score assigned to the principal diagnosis (i.e., unrelated) are counted for this variable. Comorbidities are designated "high risk" if a particular diagnosis (or related set of diagnoses) had, as a principal disease, a crude Medicare mortality rate that exceeded 10 percent in 1984. See Jonathan E. Conklin and Robert L. Houchens, "PPS Impact on Mortality Rates: Adjustments for Case Severity," Final Report (Santa Barbara, Calif.: SysteMetrics, Inc., Oct. 6, 1987), p. 21.

adopted to adjust for patient severity also permitted analysis of some patient subgroups, particularly individual disease categories. All analyses focused on 30-day postadmission mortality.

Disease Staging is structured around 398 disease categories. SysteMetrics defined the patient subgroups first by disease category and then by patient age, sex, number of unrelated comorbidities, and stage of primary disease. The trend analysis examined total Medicare discharges in 1984 and 1985, while the patient subgroup analysis largely focused on 31 separate disease groups, including the 20 diseases accounting for 75 percent of all Medicare 30-day postadmission mortality, plus another 11 with significant variation in mortality rates across hospitals.

## Severity Adjustment

In this study, SysteMetrics developed several different, though related, procedures for severity adjustment of Medicare mortality data. For the main comparison of 1984 and 1985 overall mortality, SysteMetrics adjusted for severity by controlling for changes in the distribution of patients in risk groups defined by principal disease, stage of the principal disease, number of unrelated high-risk comorbidities, age, and sex. To do this, it applied the logic of indirect standardization on a very large scale.

SysteMetrics calculated separate mortality rates for 1984 Medicare discharges for each of 12,390 different strata; that is, combinations of values for those five variables.[6] It then computed an expected overall mortality rate for 1985 by assuming a constant mortality rate within each of those strata but adjusting for the number of 1985 discharges falling into those strata. A comparison of the observed 1985 mortality rate to this expected rate indicated the extent to which the observed increase in overall Medicare mortality reflected a shift in the distribution of patients into higher risk categories.

The analysis of individual disease categories followed the same basic logic, but with variations designed to reduce, where feasible, the number of different strata. SysteMetrics examined the potential for consolidating strata by testing to see if "adjacent" strata had mortality rates that were not significantly different from each other. It defined adjacent

[6]To determine the strata, multiply 398 (principal diseases) x 3 (disease stages) x 4 (0, 1, 2 or 3+ unrelated high-risk comorbidities) x 3 (age groups) x 2 (sex groups) = 28,656. The difference between this figure and 12,390 represents potential cells for which no actual cases were found in the 20-percent sample of 1984 Medicare discharges analyzed.

strata in terms of a specified sequence of consolidation: first, groups that were the same for all variables except sex; second, age; third, number of unrelated comorbidities; and finally, disease stage. The rationale for following this sequence was to avoid arbitrary combinations and to restrict the number of times that significance tests were employed.

This procedure reduced the 54 possible strata[7] to between 10 and 37 for each of the 31 "high-mortality" disease categories. SysteMetrics consolidated the remaining 359 "low-mortality diseases" in a similar fashion, using fewer stratification variables and categories within them, including the 309 least fatal diseases, which were simply divided among 16 body system codes. This second set of analyses enabled SysteMetrics to use logistic regression analysis to test for a significant change in mortality between 1984 and 1985 (controlling for patient severity) in each of the 31 "high-mortality" disease categories. It also aided in determining the role of disease stage, comorbidities, age, and sex in defining groups of patients at higher or lower risk of mortality within those disease categories.

The problem of distinguishing between comorbidities that existed at the time of admission and complications of hospital treatment applies to both the definition of stages in the Disease Staging system and the scoring for the "unrelated high-risk comorbidities" variable. SysteMetrics dealt with this problem for comorbidities by having a panel evaluate all diseases in the Disease Staging system and identify those that, when they appeared as a secondary diagnosis, were likely to indicate a condition present prior to admission 75 percent of the time. Only these diseases were counted under the comorbidity variable, provided that they also were unrelated to the principal diagnosis and qualified as "high risk."

Secondary diagnoses also play an important role in determining the stage assigned to the principal disease category. SysteMetrics made no adjustments to the Disease Staging algorithms to remove or reduce the effect of those secondary diagnoses likely to have occurred after admission. It has cited evidence to show that in practice the "peak" stage is "almost always" the same as the stage at hospital admission. However, that may be less true for patients who die in the hospital, particularly of acute conditions, since by definition such patients deteriorate over the

[7]To determine possible strata, multiply 3 (disease stages) x 3 (0, 1, or 2+ unrelated high-risk comorbidities) x 3 (age groups) x 2 (sex groups) = 54.

course of their hospital stay. In that case, Disease Staging might tend to overestimate the severity of patient condition at admission for those who die, compared to those who survive.

## Technical Adequacy

The measurement issues raised by this approach to mortality analysis largely focus on the merits of the Disease Staging system itself (in its computerized form) as an indicator of patient condition. Researchers at the University of Michigan conducted an independent comparative assessment of four patient severity systems, including Disease Staging.[8] This study explicitly considered several aspects of validity and generally rated the computerized version of Disease Staging lower than the other systems. However, because the evaluation tested Disease Staging without the separate variable for unrelated high-risk comorbidities, these overall ratings may not apply to the adaptation of Disease Staging to mortality analyses. Moreover, the evaluation found that the clinical logic underlying the Disease Staging system gave it an intuitive validity as a measure of the severity of an illness.

One advantage that this form of Disease Staging has over alternative approaches, particularly the Risk-Adjusted Mortality Index, is that it explicitly distinguishes between comorbidities that enter into the designation of stage (i.e., severity of the principal diagnoses) and those that are unrelated. Thus, each diagnostic code is considered only once, either in the staging process or the calculation of the number of unrelated high-risk comorbidities. This should largely eliminate the potential for systematic multicollinearity among the predictor variables, which we suspect could occur with the RAMI approach.

## Data Quality

Although SysteMetrics noted the importance of accurate dates of death for calculating 30-day postadmission mortality, it made only a small number of adjustments to correct illogical combinations of discharge and death dates. It performed no general analyses on the vulnerability of the approach to data imperfections. However, findings of low reliability for staging determinations in the University of Michigan study based on a

[8]J. William Thomas et al., An Evaluation of Alternative Severity of Illness Measures for Use by University Hospitals, 3 vols., Department of Health Services Management and Policy, School of Public Health (Ann Arbor: University of Michigan, Dec. 29, 1986). The other severity adjustment systems examined were APACHE II, Medisgrps, and Patient Management Groups. The authors evaluated both the "clinical" version of Disease Staging, which relies on medical record abstraction, and "coded" Disease Staging, which uses computerized discharge abstracts.

reabstraction of medical records indicate a substantial problem in this area.[9]

The Disease Staging software should automatically produce consistent results from equivalent combinations of diagnostic codes. Therefore, low reliability strongly suggests that there is sufficient variability in the way diagnostic information is entered on bills to affect the stages assigned by Disease Staging to cases in Medicare's Statistical System.[10] However, given the lack of comparable information on the other approaches we examine in this report, we cannot say whether this is more or less of a problem for Disease Staging than for other approaches.

## Validation

While the basic Disease Staging system has been used for some years and evaluated fairly comprehensively,[11] SysteMetrics has only recently developed the version created for this analysis of Medicare mortality trends. Therefore, no specific validation of this approach has yet occurred.

## Overall Assessment

Although designed for analysis of trends rather than assessments of hospital performance, Disease Staging Adapted to Mortality Analyses demonstrates many of the same attributes as CPHA's Risk-Adjusted Mortality Index. Both estimate the risk of individual patients dying based on specific combinations of principal and secondary diagnostic codes, derived from empirical outcomes in CPHA's system and from criteria developed by physician panels combined with empirical outcomes (for the "high-risk" comorbidities) in SysteMetrics' system.

Disease Staging has handled the conundrum of distinguishing comorbidities from complications slightly differently than RAMI, but both attempt to identify clear-cut complications of treatment in order to limit the influence of complications on severity adjustment. However, Disease Staging differentiates more clearly between comorbidities that increase the severity of the principal diagnosis and those that are unrelated. Disease Staging is sensitive to inconsistencies in diagnostic coding, but the

[9]Thomas et al., vol. 2, p. 64

[10]This undoubtedly reflects the sophistication of the system. The Michigan study notes that DRGs, probably because they are clinically less precise, achieve much higher levels of reliability using the same hospital abstract diagnostic codes. Thomas et al., vol. 2, p. 67.

[11]Most notably by Thomas et al.

relative vulnerability of this adaptation of Disease Staging to data problems compared to other approaches is not known.

As with RAMI, we believe that Disease Staging shows promise as an alternative or adjunct to HCFA's current practices in its intramural analyses of Medicare patient outcomes, primarily because of its more sophisticated adjustment for patient severity. However, our conclusion is subject to the results of future testing and validation.

# National Hospital Rate-Setting Study

## General Purpose

As part of a HCFA-sponsored, multiyear evaluation of the effects of "prospective reimbursement" programs in 15 states,[12] Abt Associates developed an approach for assessing changes in quality of care that focuses on "care-sensitive conditions." These represent a subset of all medical and surgical conditions that are perceived to be "sensitive to hospital administrative and policy action." In other words, hospital administrators looking for ways to economize in the face of changed economic incentives would be relatively more likely to influence the pattern of treatment for these conditions in ways that could affect the quality of care.

In addition, Abt's approach to mortality analysis is notable for its examination of Medicare data over a relatively long time period (from 1974 to 1983, covering a full decade prior to the introduction of PPS in Medicare), as well as for the wide range of factors that it attempts to take into account when assessing the effect of state cost-containment efforts on quality.[13]

To a large extent, the characteristics of the National Hospital Rate-Setting Study reflect the specific analytical purpose for which it was designed. The study focused exclusively on the effects, including quality

[12]These programs represented a wide spectrum of hospital cost-containment efforts, involving both voluntary and mandatory controls on the structure and amount of reimbursement for hospital services. They included some state-level precursors of PPS.

[13]Information on this approach was drawn from relevant excerpts of the draft report on quality of care for the National Hospital Rate-Setting Study. HCFA approved the final report for release in May 1988.

of care, of state-level prospective reimbursement (as defined by the diverse state regulatory programs that Abt investigated). In this report, however, we are not concerned with how well Abt addressed that specific question; rather, we want to evaluate the potential for applying this technique more broadly to monitoring quality of care in the Medicare program. In chapter 7, we consider its applicability for a longitudinal assessment of the effect of Medicare's prospective payment system on quality of care.

## Substantive Focus

The focus of the Abt study on the effects of prospective reimbursement narrowed the scope of the approach in terms of both medical conditions considered and provider subgroups analyzed. The selection of "care-sensitive" conditions was intended to maximize the likelihood of finding effects, if any, on quality of care associated with prospective reimbursement. The conditions picked by Abt's panel of physicians and hospital administrators represented about 15 percent of all Medicare patients. They included 59 "urgent care" diagnostic categories and 8 "elective" surgical procedures, all defined by ICDA-8 or ICD-9-CM codes.[14]

Abt analyzed these two groups as aggregates, and made separate analyses of four relatively common conditions drawn from the urgent care (acute myocardial infarction and congestive heart failure) and elective surgical groups (inguinal herniorrhaphy and transurethral prostatectomy). Abt also conducted a parallel analysis of a random sample of all Medicare discharges to put the results for the "care-sensitive" conditions into perspective.

Abt chose to measure the outcomes in terms of postadmission mortality rates over various time periods, usually 30 and 360 days.[15] It based all analyses on data for Medicare patients over 65-years-old, but analyzed patient subgroups as defined by diagnosis only. Abt made comparisons across years (1974-1983) among states with prospective reimbursement programs, between those states and all others without such programs, and across different types of prospective reimbursement systems (such as voluntary and mandatory programs).

[14]ICDA-8 stands for the Eighth Revision, International Classification of Diseases, Adapted for Use in the United States. This was the diagnostic classification system generally used by hospitals until supplanted by the ICD-9-CM system in 1979.

[15]Abt used other indicators to assess structure and process of care aspects of quality.

## Severity Adjustment

Severity adjustment took place at several levels. First, Abt adjusted outcomes on the basis of diagnosis or procedure, age, and sex. For each of the 456 possible combinations on these three dimensions, it calculated a standard mortality rate from a random sample of Medicare patients treated in hospitals outside the 15 states of interest. Abt used these rates to calculate an expected mortality rate for each hospital in the states with prospective reimbursement in each year, based on its distribution of patients among those 456 categories. It then computed a standardized mortality ratio (observed/expected mortality). This was the outcome measure used for most of the analyses. Comorbidities and severity of illness within diagnostic categories did not enter into these calculations, owing to problems in the available diagnostic data.

In addition, Abt made a series of statistical adjustments in the regression analyses used to assess the effect of prospective reimbursement. Many of these explanatory variables involved characteristics of the communities in which hospitals were located, while others concerned the characteristics of the particular hospital.[16] Abt justified the inclusion of all these variables on the basis of their frequent use by other researchers in modeling hospital behavior and their utility in reducing the amount of variation between prospective reimbursement and nonprospective reimbursement hospitals that remained unaccounted for in their analysis.

This procedure makes sense for Abt's analysis of the specific effects of prospective reimbursement on outcomes; however, the inclusion of statistical controls for such a broad range of variables would raise more serious questions were they intended to provide an overall adjustment of health care outcomes for patient severity. Conceptually, the potential problem is the same as that described in chapter 2 with reference to adjustment of outcomes by race. Part of the variation in outcomes associated with at least some of these variables could be related to systematic differences in the quality of patient care, rather than differences in patient condition. To the extent that this is true, controlling for these variables in the regression analysis will mask the differences in quality of care that they represent. The likelihood that this problem will arise increases in rough proportion to the number of variables included in the analysis.

[16] Community characteristics included, region, racial composition of area population, proportion receiving Aid to Families With Dependent Children, proportion with private health insurance, per capita income, median education level, number of nursing home beds per capita, proportion of physicians who were specialists, and the existence of state "certificate of need" procedures; hospital characteristics included teaching status, ownership type, and number of total admissions.

## Technical Adequacy

Of particular interest from a measurement perspective are the "care-sensitive conditions." The main support for their validity and utility derives from the expertise of the panel of physicians and hospital administrators who selected them originally. Further evidence corroborating the susceptibility of these conditions to medical practice changes inimical to good quality care, and their capacity to detect differences among providers, would be desirable before adopting these conditions in other quality of care studies.

Beyond the issue raised above concerning potential masking of quality differences, more general questions can be raised about the substantive analytical implications of the broad range of control variables entered into Abt's regression equations. For example, it is not clear what community characteristics such as median income or education levels actually represent when analyzed in terms of their statistical relationship to the standardized mortality ratio of individual hospitals. Abt did not pay much attention to this problem, in part because of its particular focus on the effects of prospective reimbursement. In fact, Abt constructed certain variables to incorporate otherwise unspecified differences among states that predated the introduction of prospective reimbursement. This whole analytical strategy assumes an exclusive interest in a specific intervention—in this case, the introduction of "prospective reimbursement"—which has been separately entered into the analysis.

Even with the inclusion of all these variables, the overall regression equations did not obtain a "good fit" of these data, averaging around 4 percent of the variance in standardized hospital mortality ratios explained. Nevertheless, the regression coefficients for the variables representing prospective reimbursement programs usually achieved statistical significance. These results supported the conclusion that prospective reimbursement affected somewhat the performance of hospitals relative to a crude national standard, but most of the fluctuation in hospital performance defined in those terms remains unexplained in the Abt analysis.

In addition, the Abt analysis made no adjustments for potential analytical problems raised by statistical relationships among its predictor variables (multicollinearity) and the lack of independence among observations of outcomes at the same hospitals from year to year (serial correlation). The expense of making these adjustments for a data set of this size was judged prohibitive.

## Data Quality

Abt was sensitive to the magnitude of potential inaccuracies, including systematic as well as random error, in the Medicare data. These concerns applied particularly to diagnostic data. For example, Abt rejected the use of secondary diagnoses as a patient standardization variable because of inconsistencies in those data. In addition, Abt computed expected outcomes for hospitals separately for the periods 1974-1978 and 1979-1983 to accommodate differences in the ICDA-8 and ICD-9-CM diagnostic classification systems, which affected assignment of patients to "urgent care" categories.

More generally, Abt argued that problems with the data would have limited impact on its particular analysis of prospective reimbursement effects, unless they were related to the introduction of prospective reimbursement. Conceding that evidence of such a relationship did exist; that is, a tendency for hospitals to give patients more "severe" diagnoses when such information began to influence their reimbursement, Abt cautioned that this would tend to conceal effects from prospective reimbursement on quality. However, Abt made no adjustments in its analysis to compensate for this problem, other than suggesting a relaxation of the conventional 5-percent threshold of statistical significance for detecting an effect.

## Validation

No validation of the analytical approach employed in the National Hospital Rate-Setting Study has yet occurred, either through case record reviews or other independent evidence. In part, this reflects the focus of the study on substantive issues—the impact of specific prospective payment programs—rather than methodology development.

## Overall Assessment

Two aspects of the approach developed for the National Hospital Rate-Setting Study are potentially useful, although the approach as a whole is not a promising alternative to HCFA's current practices for general Medicare outcome analyses. First, the concept of "care-sensitive" diagnoses offers a different strategy for reducing the effect of random variation in analyses of outcomes for hospitals and other specific providers. However, its validity has not yet been systematically tested. Second, the overall approach was designed to assess the impact of a specified programmatic change on health care outcomes. Abt was concerned in this instance with state rate-setting programs, but its approach should be applicable wherever the question focuses on the effect of a particular program or policy change. We evaluate the potential utility of this approach for making such an assessment regarding PPS in chapter 7.

# Non-HCFA Approaches for Analyzing Outcome Data

After extensive search through literature and discussion with experts, we found only two alternatives to the approaches that HCFA has developed that (1) are applicable to existing Medicare data sets, (2) have actually been applied for quality monitoring purposes, and (3) represent a substantial departure from the approaches assessed in chapters 2 and 4. We describe in detail the two approaches that met our criteria, after a brief discussion of why much of the work we looked at did not prove relevant. The bibliography lists other studies that we considered and rejected as potential alternative approaches.

As do the approaches developed through HCFA's extramural research, these two non-HCFA approaches have some potential advantages over HCFA's current intramural analyses. At this point, however, neither is sufficiently well tested to clearly establish its relative merit overall, and both have only been applied to a subset of surgical procedures. Our analyses of the two main approaches described in this chapter are summarized in chapter 6, table 6.1.

## Approaches Not Meeting Our Criteria

Although the literature on quality of care is extensive, empirical efforts to actually assess quality of care by analyzing hospital outcomes are much fewer. In April 1987, Arlene Fink et al. of the Rand Corporation reported finding only 18 separate studies that had collected and analyzed data on the relationship of hospital care to patient outcomes.[1] The large majority of these involved the collection of patient data that are not available in Medicare's administrative files. For example, the Stanford Institutional Differences Study, which compared surgical outcomes across 1,224 hospitals in 1972, used laboratory results from CPHA patient abstracts as one of its risk prediction factors.

We examined a number of studies that focused on the relationship of outcomes to volume of care; that is, whether hospitals or physicians tend to produce better outcomes when they perform a given surgical procedure relatively frequently. (See Bibliography, "Volume and Outcome Literature.") In principle, the techniques developed for addressing this specific issue could be adapted to a variety of comparisons among Medicare patient and provider subgroups.[2] However, some of studies we

[1] Fink et al., "Uses and Misuses of Hospital Outcome Data" (Presentation to the American Medical Review Research Center Executive Training Program on Peer Review Outcome Data, Washington, D.C., Apr. 21-22, 1987).

[2] The problem of distinguishing real differences in outcome from random fluctuations, which is so critical to analyses of individual providers and small patient subgroups, is much less salient for the large data sets generally used in research studies of this kind.

considered require clinical data that are not available in Medicare's data sets.[3] Others analyzed administrative data using procedures to adjust for patient severity and to compare outcomes similar to those employed in the HCFA intramural and extramural approaches described in chapters 2 and 4. These procedures range from relatively simple indirect standardization based on age, sex, and presence or absence of multiple diagnoses in some of the original work, to more recent applications of logistic regression using Disease Staging to control for severity of illness.[4] We found no examples in this literature of approaches for adjusting and assessing outcomes that would substantially improve or expand upon those HCFA and its contractors have already employed.

In contrast, the two approaches discussed in detail below are quite different from those assessed in the previous two chapters.

# Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery

## General Purpose

Mark Blumberg of the Kaiser Foundation Health Plan has developed a procedure for computing expected mortality rates that builds on a statistical analysis of actual outcomes for a large population of patients.[5]

What distinguishes Blumberg's approach is its focus on nonelective surgery, the statistical technique it uses to assign cases to risk groups (recursive partitioning), and the range of tests applied to study results that check for bias and other potential problems. Blumberg initially tested this approach on all Maryland hospital discharges from April

[3]See Ann Barry Flood et al., "Does Practice Make Perfect? Part I: The Relation Between Hospital Volume and Outcomes for Selected Diagnostic Categories" and "Part II: The Relation Between Volume and Outcomes and Other Hospital Characteristics," Medical Care, vol. 22 no. 2 (Feb. 1984), pp. 98-125. This work builds on the analysis and data of the Stanford Institutional Differences Study.

[4]Harold S. Luft et al., "Should Operations Be Regionalized? The Empirical Relation between Surgical Volume and Mortality," New England Journal of Medicine, vol. 301, no. 25 (Dec. 20, 1979), pp. 1364-69; and Joyce V. Kelly and Fred. J. Hellinger, "Physician and Hospital Factors Associated with Mortality of Surgical Patients," Medical Care, vol. 24, no. 9 (Sept. 1986), pp. 785-800.

[5]Mark S. Blumberg, "Maryland Mortality for Non-Elective Surgery: A Prototype RAMO System," unpublished paper, May 6, 1987

1984 through March 1985. In this analysis, he both assessed the performance of individual providers and compared outcomes among a variety of patient subgroups.

## Substantive Focus

Blumberg designed his approach specifically for surgical conditions. Within those, he concentrated on approximately 250 nonelective surgical procedures with overall mortality rates of 6 percent or more. These procedures accounted for 5 percent of all surgery cases in the state that year (1.7 percent of all hospitalizations), but over 44 percent of all surgical deaths and 9 percent of all hospital deaths. Forty-six percent of the cases (and 57 percent of the deaths) involved Medicare patients.

Blumberg's initial analysis centered on inpatient mortality because of its availability in the Maryland hospital discharge data set. However, he noted that his approach would lend itself to a variety of alternative outcomes, including deaths 30 days after surgery and postoperative complications, providing that good data on those outcomes could be obtained.

Blumberg excluded from his analysis all cases with a principal or secondary diagnosis of metastatic cancer. He argues that patients with advanced cancer should not be evaluated in terms of individual hospital episodes. Rather their care should be assessed using data from tumor registries, which collect detailed diagnostic and treatment information on patients over an extended period. Surgery for cancers that have not metastasized are included in his system.

Since the expected and observed mortality rates are calculated from patient-level data, in principle any groups that can be defined using information in the data file can be analyzed for differences in outcomes. Thus, Blumberg's approach can be applied to analyses of a wide range of patient and provider subgroups. He has compared the performance of individual physicians and hospitals for specified classes of surgical procedures as well as overall. In other analyses Blumberg has examined patterns of outcomes related to whether a case was admitted during the week of a national holiday.

## Severity Adjustment

Blumberg's approach draws on many of the same data elements employed by other systems to adjust for severity: patient age and sex (but not race), principal and secondary diagnoses, and surgical procedures. In addition, Blumberg used type of admission (e.g., elective, urgent, and emergent), source of admission (e.g., transfer from a nursing

home), and dates of admission, procedure, and discharge. He combined these data elements in various ways to produce 47 different predictor variables of several types. One set rank-ordered diagnoses and procedures by observed death rates in the state as a whole and assigned each to one of six ordered categories (fractiles) reflecting that ranking. Other variables noted the presence of specific diagnoses, procedures, or particular combinations of diagnoses singled out by Blumberg on the basis of their overall observed mortality rate and clinical logic.

In his analysis, Blumberg considered secondary diagnoses in the construction of these predictor variables only if they involved chronic conditions (specific diagnoses within diabetes, chronic renal disease, hypertension, chronic heart disease, malignancy, and obesity) or trauma. He did not include diagnoses that could represent either complications of treatment or comorbidities, such as pneumonia. Thus, Blumberg took an approach similar to that of Disease Staging but carried it somewhat farther, which was to err on the side of missing potential comorbidities so as not to adjust inappropriately for possible complications of treatment.

Unlike many of the approaches described previously, Blumberg chose not to develop separate models to derive expected mortality rates for different conditions or procedures. Instead, he sought to generate one model for all nonelective surgery. To do this, he applied a statistical technique known as recursive partitioning. This technique sorted all the cases involving the selected nonelective surgical procedures into groups defined by the particular combinations of predictor variables that differentiated most effectively among cases on the dimension of inpatient death. On the basis of this analysis, Blumberg formed 121 mutually exclusive "equal risk of outcome groups." Every case was assigned to one of these groups and assumed the group's observed mortality rate as its expected mortality rate.

A key advantage of recursive partitioning for this analysis is its ability to uncover significant interactions among predictor variables. In other words, certain combinations of values for different factors may have a stronger impact on the probability of dying than any of those factors viewed alone. For example, if men with hypertension were more likely to die from surgery and women with hypertension less likely, each variable—sex and hypertension—alone would, on average, have a weak relationship to mortality, but the combination would have a strong association. Interaction effects can only be picked up by regression analyses, both linear and logistic, if they are individually specified in the analysis.

Recursive partitioning, by contrast, actively searches among all potential combinations of the specified predictor variables and selects those that have strong associations with the outcome of interest.

For any groups of patients being analyzed (e.g., those treated at different hospitals), Blumberg computed the number of observed and expected deaths. He assessed the statistical significance of the difference between observed and expected deaths by a chi-square test wherever the number of expected deaths was five or more, and by a 95-percent confidence interval for a Poisson distribution where fewer than five expected deaths were involved.

## Technical Adequacy

Blumberg's use of diagnostic and procedure data raises concerns about measurement and analytical techniques that parallel those discussed above with reference to CPHA's Risk-Adjusted Mortality Index. First, Blumberg selected highly complex predictor variables on the basis of their success in predicting mortality, as well as their consistency with clinical logic. No studies have systematically tested the validity and reliability of these variables compared to possible alternatives.

Second, the implications of correlations among multiple variables derived from the same data elements need to be explored.

Recursive partitioning, like logistic regression, is affected by correlations among the predictor variables being tested. To the extent that two predictor variables are correlated, the technique tends to choose one, potentially by a very close margin, and discounts the importance of the other. As with the CPHA logistic regressions, we believe that this might bias comparisons of outcomes across patient subgroups, if discrete categories of patients were differentially affected by an exclusive reliance on one of these two variables. However, the risk that this would occur is reduced by the sensitivity of recursive partitioning to differences among subgroups (i.e., interactive effects). If the differential effect of the rejected correlated variable is strong enough, the technique should select that variable in subsequent partitions that specifically involve the subgroups of cases affected by it.

Moreover, Blumberg found little evidence of bias in the tests that he applied to his model. For example, he compared the expected-to-observed mortality ratios for all categories of the predictor variables to ensure that the variables included in the recursive partitioning procedure were all adequately adjusted for. He also compared observed and

expected outcomes for detailed procedure and diagnosis codes not entered into the model and found few instances where they diverged significantly. Blumberg did report a slight tendency for his model to underestimate risks for certain high-risk cases.[6]

## Data Quality

Blumberg underscored the general vulnerability of his approach to systematic errors in the data to which it is applied. As one possible check for this, he suggested monitoring trends in the expected outcomes of individual hospitals, with nonrandom shifts possibly indicating changes in a hospital's coding procedures. More generally, Blumberg emphasized the importance of examining data that are used for key predictor variables. He did this to a limited extent in his analysis of the Maryland data, which resulted in his dropping 11 hospitals (accounting for 18 percent of all surgical cases) from his analysis because he found inconsistencies in the hospitals' coding of cases as elective or nonelective.

## Validation

In addition to the tests for bias noted above, Blumberg has collected some limited, independent evidence concerning the overall effectiveness of his approach in identifying genuine quality problems. Most of this evidence derives from contacts with hospitals that did well or poorly in his analysis of Maryland surgery cases. Blumberg found that the quality assurance staff in several of the hospitals with adverse results had independently reached similar conclusions in terms of departments or types of patients demonstrating poor outcomes. Similarly, a hospital that did especially well in one category of surgery turned out to specialize in that area and to have a particularly high volume of those cases.

## Overall Assessment

Although Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery only relates to surgical conditions, it applies to almost half the cases, for the general population in Maryland, where surgery results in death. The approach employs a sophisticated technique for patient severity adjustment that is particularly sensitive to interactions among predictor variables. However, as we noted, its complex predictor variables have not yet been systematically tested for validity and reliability. Moreover, correlations among variables created from common data elements might affect the statistical procedure (recursive partitioning)

[6]These were sets of cases with expected death rates much higher than the mean for similar cases (i.e., procedures within the same body system).

employed to estimate expected mortality rates. Blumberg has gone farther than most to test for bias in the results of his analyses, but further testing of the overall validity of the approach is needed.

## Computerized Identification of Surgical Complications

### General Purpose

Leslie Roos and colleagues have proposed a system for monitoring outcomes from administrative data sets that differs substantially from the types of analyses that HCFA currently performs.[7] This approach uses diagnostic codes to assess outcomes directly. Specifically, it identifies cases readmitted to a hospital owing to complications of surgery (adverse outcomes such as infection directly attributable to the surgery), based on the diagnoses recorded for the second hospital admission and the time elapsed since the operation. Two physicians identified the diagnostic codes indicative of complications by reviewing abstracts of medical histories (but not medical records) for a large group of rehospitalized patients in Manitoba, Canada. Roos et al. developed and tested computer programs that permit automatic screening of large data sets for cases that fit the specified diagnostic patterns for surgical complications.

### Substantive Focus

This approach requires the development of separate diagnostic indicators of complications for different surgical procedures. So far, Roos et al. have done this for three relatively common procedures: hysterectomies, cholecystectomies, and prostatectomies. Certain types of patients have been excluded, such as those operated on for malignancy and some emergency or repeated procedures. Only those complications serious enough to lead to rehospitalization are counted in this approach.

[7] Leslie L. Roos, Jr., et al., "Using Computers to Identify Complications After Surgery," American Journal of Public Health, vol. 75, no. 11 (Nov. 1985), pp. 1288-95; and Leslie L. Roos, Jr., et al., "Centralization, Certification, and Monitoring: Readmissions and Complications After Surgery," Medical Care, vol. 24, no. 11 (Nov. 1986), pp. 1044-66. John Wennberg has led a team conducting related research analyzing data from both Manitoba and Maine on prostatectomies and examining mortality as well as one major complication (reoperation). See John E. Wennberg, et al., "Use of Claims Data Systems to Evaluate Health Care Outcomes: Mortality and Reoperation Following Prostatectomy," Journal of the American Medical Association, vol. 257, no. 7 (Feb. 20, 1987), pp. 933-36.

Roos et al. have applied their approach to evaluate the performance of individual hospitals in Manitoba, as well as to assess the general impact on surgical outcomes of such factors as hospital volume for a given procedure and the extent of physician experience with that procedure.

## Severity Adjustment

The diagnostic indicators of complications do not in themselves incorporate an adjustment for patient severity. In part, this reflects the intrinsic nature of this outcome; unlike death, complications are by definition conditions that, in principle, should not occur. Nevertheless, it is not reasonable to expect to reduce the occurrence of complications to zero, particularly for patients whose poor health status (e.g., multiple chronic comorbidities) makes them especially vulnerable. Precisely because patients vary in their vulnerability to complications, we believe that some adjustment for patient severity is called for when comparing outcomes among groups of providers or patients.

In fact, Roos et al. do make a separate adjustment for several patient characteristics in their analyses of factors affecting surgical performance. These characteristics include age, sex, and residence (urban vs. rural), three medical history variables (any prior hospitalization in the last 2 years, number of prior ambulatory physician visits with chronic disease diagnoses, and whether any prior ambulatory physician visits had diagnoses of heart disease), and finally the number of diagnoses recorded for the hospital admission when the surgery was performed.

The emphasis on medical history provides yet another strategy for dealing with the general problem of clearly identifying preexisting comorbidities that could influence patient condition at admission. However, only the prior hospitalization factor would be feasible using existing Medicare data bases, since the diagnoses associated with ambulatory physician visits are currently not recorded in the Medicare Statistical System.

Unlike several other approaches that adjust uniformly for patient characteristics across conditions, Roos et al. made adjustments only for those patient variables that attained statistical significance in the separate logistic regression analyses performed for the three surgical procedures. Therefore, the risk factors adjusted for differed substantially among the three procedures, with only prior hospitalization and ambulatory visits for chronic diagnoses included in as many as two of the three analyses.

## Technical Adequacy

A key measurement issue raised by this approach is the reliability and validity of its identification of complications using hospital admission diagnoses. The computer programs developed by Roos and his colleagues were designed to replicate the decisions made by the two physicians who established the diagnostic indicators. The physicians in turn identified complications based not on actual medical record reviews, but on medical history summaries containing the dates and diagnoses of all hospitalizations and physician visits in a 2-year period. (For a few cases, these physicians reviewed some additional information abstracted from medical records.)

Cases were counted only if both physicians agreed that they represented complications (either independently or after discussion). Until these judgments receive additional testing, we will not know how closely the decision rules on complications adopted by these two physicians correspond to those that other practicing clinicians might set, either overall or among those with the highest technical expertise.

Further, this procedure relies heavily on the accuracy and completeness of the recorded diagnoses. For all three procedures, the primary diagnosis alone was the basis for identifying over 90 percent of the complications. Roos et al. cite previous studies that found few errors in the data files they used, managed by the Manitoba Health Services Commission, with respect to correct abstraction of diagnoses from patients' medical records. However, the medical records themselves may contain diagnoses that are not complete and accurate, or not fully consistent across comparable cases.

Roos et al. provide some indication that such factors could affect their results. For example, they express concern about the potential for "opportunistic coding" should their system be used to monitor hospital performance. They also characterize "post operative wound infection," one of the more frequently cited diagnoses indicative of complications, as "notoriously subjective." Both these comments suggest that physicians have some discretion in deciding whether or not to enter in a patient's medical record the diagnoses used to identify surgical complications.

In our view, this element of discretion is likely to lead to systematic errors in recorded diagnoses. An unknown, but potentially substantial, proportion of the diagnoses for readmission are made by the physicians

responsible for the original surgery.[8] These physicians, especially those whose complication rates are relatively high, would have an incentive to record diagnoses that were less indicative of complications, even without the institution of a formal monitoring mechanism. In the absence of a detailed comparison of cases selected and not selected by the Roos approach, preferably involving detailed reviews of the basis for the recorded diagnoses, we cannot determine the proportion of complications that this approach misses owing to the unreliability of the basic diagnostic data used.[9] Such diagnostic errors would also affect other systems that employ diagnostic data for severity adjustment, but the effect would probably be greater where surgical complications are the specific outcome of interest.

## Data Quality

Roos and his colleagues note that the Manitoba data they analyzed are likely to be of higher quality than Medicare's data. Therefore, they suggest that checks on data quality precede application of their approach to Medicare administrative files. They do not specifically analyze how the results of their approach might be distorted as a result of these anticipated limitations in the quality of the data recorded in Medicare's computerized claims files.

## Validation

As noted above in the discussion of diagnostic coding, this approach has not yet been validated using medical record reviews or other independent indicators of surgical complications.

## Overall Assessment

The Computerized Identification of Surgical Complications provides a specific technique for using claims data on three surgical procedures to analyze morbidity rather than mortality, thereby demonstrating an overall approach that could be replicated for other types of surgery. This approach draws on the clinical judgment of physicians to identify patterns of diagnoses that are indicative of surgical complications in a large administrative data file of Manitoba hospital patients. Roos et al.

[8]Roos et al. provide data on cholecystectomies that show that about 69 percent of cases with complications are readmitted to the same hospital where the surgery was originally performed.

[9]Evidence of underreporting of certain types of complications in the medical record is presented in R. Michael Massanari et al., "Reliability of Reporting Nosocomial Infections in the Discharge Abstract and Implications for Receipt of Revenues under Prospective Payment," American Journal of Public Health, vol. 77, no. 5 (May 1987), pp. 561-64. However, these data refer to complications detected during the initial hospital episode, not to subsequent readmissions that were caused by surgical complications.

have also shown how these diagnostic indicators can be analyzed in conjunction with other data on patients to adjust for differences in severity, although several of the specific variables they use are not available in Medicare administrative data sets.

The available evidence on the validity of this approach is still somewhat limited. The critical decision rules for identifying surgical complications reflect the collective judgment of just two physicians, and this has not yet been independently verified. Moreover, potential problems with the reliability of the diagnostic data used in making these determinations could result in an unknown proportion of cases with complications being missed.

Chapter 6

# Comparison of HCFA and Alternative Approaches

In this chapter we compare the relative strengths and weaknesses of the nine evaluated approaches to analyzing Medicare outcome data according to the criteria laid out in chapter 1. We focus on whether benefits could be gained by expanding HCFA's current practices to include aspects drawn from the six extramural and independent approaches. We present our conclusions and recommendations at the end of the chapter, together with relevant comments from the Department of Health and Human Services and our response to those comments.

Because none of the approaches has yet been fully validated, we cannot establish which works best in practice to identify quality of care problems. Our assessment of relative strengths and weaknesses, therefore, relies primarily on a logical analysis of how completely and carefully different approaches deal with specific problems faced by outcome analyses using administrative data. These problems would include, for example, the difficulty of distinguishing secondary diagnoses indicative of preexisting comorbidities as opposed to complications of treatment for purposes of severity adjustment. (See appendix I for a full listing of the specific analytical issues we addressed in our assessment.) Clearly, until researchers have validated these approaches using independently derived evidence such as medical record reviews, we cannot determine whether the logically more complete and careful strategies for dealing with these problems actually work as intended. However, our analysis should help to assign priorities among competing approaches for future validation efforts. Moreover, it can serve to identify specific techniques that might usefully be borrowed from one analytical approach and adapted to others.

Table 6.1 summarizes under general purpose the basic issues addressed as each approach has been applied to date. Diagnoses included describes the patient population covered by the approach as defined by principal diagnosis. Severity adjustment gives an overall rating based on the degree to which diagnostic data were used to adjust individual patient risks and care was taken to distinguish complications from comorbidities. Quality of measurement rates the validity, reliability, and sensitivity of measures used in the analysis. The appropriateness of application rates the extent to which application of the analytical technique used accords with its assumptions and limitations. Data quality summarizes the probable impact of missing or inaccurate data elements on results. The extent of validation summarizes the available evidence on the effectiveness of the approach in identifying quality of care problems.

**Table 6.1: Comparison of Approaches for Analyzing Medicare Patient Outcomes**

| | | Substantive focus | | |
|---|---|---|---|---|
| **Approach** | **General purpose** | **Outcome type** | **Diagnoses included** | **Discrete populations analyzed** |
| **HCFA-intramural** | | | | |
| 1986 hospital mortality analyses | Provider performance | Inpatient mortality | All | 9 diagnosis-related groups |
| 1987 hospital mortality analyses | Provider performance | Mortality within 30 days of admission | All | 16 diagnostic clusters |
| HSQB monitoring systems | Trends over time, patient subgroups | Mortality and readmission over multiple time periods following admission, morbidity and disability based on postdischarge costs | All for trend analyses, 9 specific medical and surgical conditions for patient subgroup analyses | 7 "major conditions" for trend analyses; patient subgroups defined by 9 "tracer conditions," further divided by race, sex, age, and presence of 9 comorbidities |
| **HCFA-extramural** | | | | |
| Nonintrusive Outcomes Study | Provider performance | Inpatient mortality, postdischarge mortality | 48 medical and surgical conditions | Individual hospitals for the 48 conditions |
| Risk-Adjusted Mortality Index | Provider performance | Inpatient mortality | All except neonatal conditions | Individual hospitals for 310 diagnostic groups |
| Disease Staging Adapted to Mortality Analyses | Trends over time | Mortality within 30 days of admission | All | 31 specific disease categories plus age, sex, comorbidities, and disease stage |
| National Hospital Rate-Setting Study | Trends over time, patient subgroups | Mortality within 30 days and 1 year of admission | 59 "urgent care" conditions and 8 elective surgical procedures | 15 state programs regulating hospital revenues |
| **Non-HCFA** | | | | |
| Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery | Patient subgroups, provider performance | Inpatient mortality | 250 higher risk surgical procedures | Hospital, physician, diagnosis, and others |
| Computerized Identification of Surgical Complications | Provider performance, patient subgroups | Surgical complications resulting in hospital readmissions | Hysterectomy, cholecystectomy, prostatectomy | Hospital, physician specialty, urban or rural location, and others |

| Severity adjustment | Technical adequacy: Quality of measurement | Technical adequacy: Analytical technique | Technical adequacy: Appropriateness of application | Data quality | Extent of validation |
|---|---|---|---|---|---|
| Rudimentary | Medium | Multiple regression, t-test for significance | Significance test not appropriate | Unknown | Fragmentary evidence based on PRO reviews employing varying approaches |
| Moderate | Medium | Logistic regression, formula for range of "expected" mortality incorporates both sampling variance and overall interhospital variance | Appropriate | Unknown | None based on independent data sources; consistency using alternative statistical procedures tested |
| Moderate | High for mortality, low for morbidity | Life table analyses, Cox proportional modeling | Appropriate | Unknown | Limited case record reviews underway |
| Rudimentary | High | Indirect standardization, binomial significance tests | Appropriate | Currently unknown, but under study | Extensive case record reviews underway for two conditions |
| Sophisticated | Medium | Indirect standardization, logistic regression, binomial and Poisson significance tests | Appropriate | Results influenced by incomplete secondary diagnosis coding | Limited number of hospital site visits; consistency of results using data from different years tested |
| Sophisticated | Medium | Indirect standardization, logistic regression, chi-square significance tests | Appropriate | Results influenced by random variation in coding of diagnoses in claims files | None |
| Moderate | Medium | Multiple regression | Appropriate | Results influenced by changes in diagnostic coding over time, but only principal diagnosis used | None |
| Sophisticated | Medium | Indirect standardization, recursive partitioning, chi-square and Poisson significance tests | Appropriate | Results influenced by accuracy of coding for elective and nonelective surgery | Tests for potential bias, limited number of hospital site visits |
| Moderate | Medium | Logistic regression, chi-square significance test | Appropriate | Results influenced by thoroughness and accuracy of coding in medical record of surgical complications | None |

# Comparative Assessment

## General Purpose

HCFA's intramural analyses of outcomes collectively address all three of the general purposes identified in chapter 1: assessing the performance of individual providers, monitoring changes in outcomes over time, and comparing the outcomes of different patient groups. The main advantage that other approaches, particularly Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery, have demonstrated is greater flexibility to accomplish several of these purposes with the same data set. This requires the ability to compute both the observed and expected frequency of the designated outcome for all individual patients. These results can then be aggregated along any dimension by which those patients can be categorized, including their demographic characteristics, their surgical or medical condition, and the hospital or physician that treated them. In principle, similar flexibility would be available over a wider range of medical conditions using HCFA's extramurally developed Risk-Adjusted Mortality Index and Disease Staging Adapted to Mortality Analyses or the patient-level data employed in HCFA's 1987 hospital mortality analyses.

## Substantive Focus

### Types of Outcomes

The nine approaches differ on two major points concerning the types of outcomes monitored. The first is consideration of any outcomes other than mortality. Almost all the approaches we examined focus predominantly or exclusively on mortality outcomes, primarily because of the limited data available in administrative data files. Nevertheless, HCFA currently analyzes readmissions and costs of subsequent Medicare services as proxies for morbidity and disability in HSQB's monitoring systems. Several of HCFA's extramural approaches also examine readmissions to some degree. The one approach that would expand on HCFA's current practice is the system for Computerized Identification of Surgical Complications. Although limited so far to three procedures, similar criteria could be developed to identify complications for other types of surgery.

The second main issue dividing approaches concerns the way mortality is defined. Some focus on deaths that occur in the hospital. Others evaluate hospital outcomes in terms of deaths within a certain period of time following hospital admission, whether or not the patient was discharged during that time.[1] Each has characteristic strengths and weaknesses. Inpatient mortality is more clearly related to hospital treatment, assuming appropriate adjustments for patient severity, but variations in average lengths of stay among hospitals can make comparisons across providers deceptive. Table 6.1 lists the type of mortality measure each approach has adopted. HCFA has employed both in the past, but has recently tended to use mortality within fixed time periods. In principle, one could use either definition of mortality with any of these approaches.

## Range of Medical Conditions

HCFA's intramural analyses of provider performance and trends over time have covered the full range of medical and surgical conditions from the start. Its analysis of patient subgroups, by contrast, has largely focused on nine tracer conditions representing about a third of all Medicare discharges, although the 1987 hospital mortality analyses introduced a new set of 16 diagnostic clusters encompassing 70 percent of Medicare patients. Two of HCFA's extramural approaches include all Medicare hospital patients, while the rest are more limited in their scope. The two non-HCFA approaches have a narrower focus in terms of conditions analyzed; both are limited to a subset of surgical procedures, although the cases reviewed by the Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery included nearly half of all inpatient deaths related to surgery

## Range of Patient and Provider Subgroups

HCFA has so far given relatively little attention to analyzing patient and provider subgroups. The 1986 and 1987 analyses of hospital mortality have considered only interhospital comparisons, with separate analyses of specific clusters of diagnoses. The HSQB monitoring systems have examined the relative impact of a few selected factors: race, sex, age, and nine specific comorbidities.

The four approaches that HCFA has developed extramurally do not expand this dimension much, largely owing to their focus on comparisons across providers or aggregate trends over time. Still, the Risk-

[1]The Rand Nonintrusive Outcomes Study looks as well at "total deaths," defined as all inpatient mortality plus any deaths within 30 days of discharge (not admission).

Adjusted Mortality Index has examined more extensively than any of HCFA's intramural analyses the distribution of potential problems within hospitals by diagnostic groups, while Disease Staging Adapted to Mortality Analyses provided a similar breakdown of trends over time for individual disease categories. The two non-HCFA approaches have performed a substantially wider range of subgroup analyses.

## Severity Adjustment

In order to make credible inferences about variation in quality of care from analyses of health care outcomes, some means must be found to adjust for differences among individuals in their intrinsic vulnerability to adverse outcomes. This variability in patient severity is most likely influenced by numerous factors, including the specific medical condition afflicting the patient, the intensity of that disease, the presence of other medical conditions (comorbidities), and basic demographic factors such as age and sex that could affect the patient's overall physiological and psychological resources. Ideally, an approach to analyzing outcomes would take the effects of all relevant severity factors into account, so that any remaining differences in outcomes would directly reflect differences in the quality of care provided.

All the approaches we examined made some effort to adjust for patient severity, although the level of sophistication varied greatly. Among HCFA's intramural outcome analyses, adjustments made in the initial examination of hospital mortality in 1986 had a number of shortcomings (see chapter 2 for details), which the 1987 analyses largely overcame. Several of the approaches that HCFA has developed extramurally represent additional improvements, while the two non-HCFA approaches offer variations with distinctive advantages and disadvantages. Nevertheless, even the most sophisticated approaches have recognizable limitations. These reflect the current state of medical knowledge upon which to base predictions of patient outcomes and the limited amount of clinical information available in existing administrative files.

Several characteristics distinguish the more sophisticated adjustments for patient severity from approaches that we have rated as rudimentary on this dimension in table 6.1. First is the range of factors for which the approaches make adjustments. The rudimentary approaches either limit themselves to controls for age and sex within clusters of related diagnoses or else use flawed measures in attempting a more complete adjustment for patient condition. By contrast, the most sophisticated systems adjust for demographic variables in conjunction with detailed diagnostic

information on both principal diagnoses and comorbidities. Approaches rated moderate fall somewhere in-between.

A second characteristic of sophisticated approaches is their effort to compensate for known problems with the relevant data available in Medicare's administrative files. For example, these approaches do not rely on Medicare's system of DRGs to adjust for differences in case-mix, owing to the clinical heterogeneity of patients within DRGs. Similarly, in considering secondary diagnoses, the more sophisticated approaches attempt to distinguish between comorbidities present at the time of hospital admission and complications of treatment. Without such an effort, the results of poor quality care may be inappropriately treated as evidence of a more severe illness and the effects of genuine comorbidities neglected in adjusting outcomes for patient condition.

Another factor that could affect the appropriateness of severity adjustments is the inclusion of factors that potentially mask real differences in quality of care. For example, if racial minorities systematically received poorer care across the groups being compared (e.g., hospitals), then adjusting for race before comparing outcomes would conceal that aspect of relative performance. Since we do not know the degree to which such factors reflect differences in quality as opposed to genuine differences in patient condition, adjustment for race did not affect our ranking of approaches in table 6.1. However, we note that only four of the nine approaches included race as a predictor variable: HCFA's 1986 hospital mortality analyses, the HSQB monitoring systems, the Risk-Adjusted Mortality Index, and the National Hospital Rate-Setting Study.

We rated three approaches as "sophisticated" in table 6.1: the Risk-Adjusted Mortality Index, Disease Staging Adapted to Mortality Analyses, and Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery. Although each is different in several notable ways, each represents an improvement over HCFA's current practice in severity adjustment. In particular, each uses information about the probability of death associated with specific principal and secondary diagnoses in estimating expected mortality, in contrast to the clusters of diagnoses that HCFA has relied on to adjust for differences in case-mix and the effect of comorbidities.

Comparing the three sophisticated systems, we found differing strategies and compromises for dealing with the same inherent constraints. For example, the Risk-Adjusted Mortality Index counts ambiguous secondary diagnoses, those that could have developed during the hospital stay or been present at the time of admission, as comorbidities present

at admission. The Disease Staging system, by contrast, only includes in its unrelated high-risk comorbidity variable those secondary diagnoses that were likely to have been present at admission at least 75 percent of the time. However, the Disease Staging system also uses secondary diagnoses to assign stages to the principal condition, and these may include complications as well as preexisting comorbidities. Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery is the most restrictive, considering as comorbidities only a limited number of specific diagnoses that clearly represent chronic illnesses or other conditions, such as trauma, that are highly unlikely to occur after admission.

## Technical Adequacy

### Quality of Measurement

Relatively little attention has been paid to issues of measurement in the development of any of these systems. We found few examples of formal tests of validity, reliability, and sensitivity. Therefore, the systems that rate higher on this dimension do so because they have limited their analysis and severity adjustment to factors that intrinsically raise few measurement issues. The Nonintrusive Outcomes Study, for example, only controls for age and sex. The HSQB monitoring systems also employ age and sex as predictive factors, plus race and the presence of specific comorbidities defined by ICD-9-CM codes. By tracking mortality in terms of a survival analysis following admission over several years, HSQB incorporates unusually complete information on this outcome in its analysis. However, the measures derived from Medicare costs used by HSQB as proxies for morbidity outcomes bear only limited logical relation to the concept of morbidity and are not justified on any basis other than data availability.

The systems that include more complex indicators in their severity adjustments have yet to systematically evaluate those constructs. For example, the tests performed on the Risk-Adjusted Mortality Index and Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery have examined the results generated by their "models" of expected mortality, rather than the component parts of those models. Even for the Disease Staging system, which has been developed, refined, and used over a substantially longer period of time, no formal evaluation of the version that includes the unrelated high-risk comorbidity factor has yet been undertaken, particularly for the purpose of calculating expected mortality rates from large administrative data sets. Regardless of the success of

these approaches overall in predicting deaths across hospitals or other groups, unless the factors used to adjust outcomes can be supported individually as plausible indicators of patient condition at admission, the legitimacy of those adjustments as representative of differences in patient severity will remain in doubt.

With the limited exception of Computerized Identification of Surgical Complications, we found that neither HCFA nor others have succeeded in devising credible indicators of morbidity (as opposed to mortality) outcomes using the types of administrative data currently available. Nevertheless, there is a need to go beyond mortality analyses if variations in patient outcomes and quality of care are to be assessed more fully. HCFA has recognized this need and included as a priority in its latest solicitation for grants and cooperative agreements research leading to "reliable and valid measures of patient outcome that are more sensitive to variations in the quality of care than currently used measures [mortality or rehospitalization rates]."[2]

## Type of Analytical Technique

As table 6.1 indicates, the nine approaches employ diverse analytical techniques to make statistical adjustments for differences in patient severity and to analyze remaining differences in outcomes. They range from complex applications of regression or recursive partitioning procedures to build elaborate models of expected mortality to quite simple examples of indirect standardization across a limited number of patient subgroups. Several of the approaches employ a combination of advanced and basic techniques, matching the complexity of the analysis to the level of variability in expected deaths found in different medical conditions.

## Appropriateness of the Analytical Technique's Application

By and large the approaches we examined respect the assumptions and limitations of the analytical techniques they employ. The main exception was the original 1986 HCFA analyses of hospital mortality, particularly its application of statistical significance tests based on normal rather than binomial or Poisson distributions.[3]

[2]Health Care Financing Administration, "Medicare and Medicaid Programs; Health Care Financing Research and Demonstration; Availability of Funds for Cooperative Agreements and Grants," 52 Fed. Reg. 34307 (1987).

[3]As explained in chapter 2, significance tests based on binomial or Poisson distributions are more appropriate for analyses of mortality, especially where the number of cases at risk is small, because they reflect the fact that deaths only occur as whole numbers, not fractions, and cannot go below zero.

We have raised some potential concerns regarding several of these approaches, however. One has to do with possible biases in comparisons of outcomes across patient subgroups, including hospitals, which could derive from correlations among multiple risk factors constructed from common data elements on diagnoses. This concern particularly applies to the Risk-Adjusted Mortality Index. Disease Staging Adapted to Mortality Analyses minimizes the potential for such bias by considering each diagnostic code only once, either in the designation of the stage of the principal disease or in the assessment of unrelated comorbidities. Only the Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery has undergone specific testing for bias. Until such tests are performed across a range of major patient subgroups, we cannot determine whether biases of this sort exist and, if they do, how they might affect the results.

The regression analyses of the National Hospital Rate-Setting Study also had to contend with problems of correlations among its independent variables (in this case caused by the number of factors entered into the analysis, rather than structural relationships among factors derived in part from the same data elements) as well as serial correlations in observations from year to year. However, the limited analytical objective of the study—assessing the impact of specified program changes on outcomes—makes these problems less acute. Similarly, the use of age as a predictor variable over a several year period in the HSQB monitoring systems could in some instances be problematic, given the assumption in the Cox proportional hazards model that such variables will have a consistent effect over the period under study.[1] For the most part, however, the deviation from constancy over the 2- to 3-year period that HCFA has examined to date is not very large.

## Data Quality

Little is known about the vulnerability of most of these approaches to imperfections in the data; however, the fragmentary evidence currently available suggests that their results are affected by data problems in Medicare data sets. For example, a number of the PROs charged with evaluating the hospitals identified as outliers in HCFA's 1986 hospital mortality analyses reported finding major discrepancies between their own data sets and HCFA's on such basic information as the number of

[1]As explained in chapter 2, when dealing with a relatively elderly population, the risk of dying associated with higher ages (80+) increases more rapidly from year to year over the study period than for lower ages (65-70), so that the relative risk associated with age is not constant.

discharges and inpatient deaths experienced by individual hospitals. According to the PRO data, many of these hospitals were not outliers.

Additional evidence of data problems potentially affecting results were reported for CPHA's Risk-Adjusted Mortality Index, SysteMetrics' Disease Staging Adapted to Mortality Analyses, and Blumberg's Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery. CPHA has learned from its site visits that substantial variations among hospitals in the extent to which they record secondary diagnoses has a decided impact on the results of their analysis. Specifically, the Risk-Adjusted Mortality Index seriously underestimates expected mortality for those hospitals recording few secondary diagnoses, making their outcomes appear worse than they would with complete data. The low rating of Disease Staging on reliability indicates that it is vulnerable to inconsistencies in diagnostic coding on claims files. Finally, Blumberg reported a substantial proportion of hospitals with inconsistent coding of elective and nonelective surgery, a key variable for his system.

Aside from the distinction between elective and nonelective surgery, which only Blumberg emphasizes, the problems noted above are likely to affect many of the nine approaches. All the approaches with sophisticated adjustments for patient severity rely heavily on detailed diagnostic coding, for example. They might all prove as unreliable as Disease Staging if put to the same test as that applied in the University of Michigan evaluation.

With only piecemeal indications of the nature of these problems, we have very little information upon which to base an estimate of the probable overall impact of missing or inaccurate data on the results that any of these approaches would generate using existing Medicare data files. In large part, this reflects the lack of information on the percentage of cases in Medicare files with missing or inaccurate entries for specific data elements. Without such information, the developers of these approaches would have to undertake large-scale medical record reviews of their own to address this issue.

So far, only the Rand Nonintrusive Outcomes Study has adopted this approach. If, however, specific information on missing and inaccurate data were available, it could aid in assessing the effect of these inadequacies in Medicare's data on each approach. Researchers could conduct a series of simulations designed to test the sensitivity of their findings to specific types of data discrepancies. Until this is done, we will not know how much, and in what way, deficiencies in Medicare administrative

data affect the capacity of these approaches to analyze Medicare outcomes.

## Validation

None of these approaches has yet undergone extensive validation using independently derived information. Such information should usually include, but not be limited to, medical record reviews for a representative sample of patients. Because the persuasiveness of different data sources varies depending on the quality issues in question, no single source of independent information is likely to be definitive. For example, medical records can document that hospitals carried out certain procedures on a patient, but often provide little insight on how well they were performed. Researchers could assess validity in terms of the number of different data sources drawn upon and the quality of corroboratory evidence obtained from each.

For some approaches, work is underway that could begin to provide this type of validating information in the relatively near future. This work includes the Nonintrusive Outcomes Study and, on a more limited scale, a pilot project conducted by eight PROs, which will collect clinical data for patients with five of the conditions examined in HSQB's monitoring systems. CPHA and Blumberg have followed a more informal approach, which draws on the reactions of hospital administrators and physicians in selected hospitals to outcome analyses of their own institutions.

Some tentative inferences about limitations in HCFA's 1986 hospital mortality analyses may also be made from the PRO review of those outlier hospitals. Validation of the remaining approaches has so far involved only tests of internal consistency and logic, based on the same administrative data as the original analysis of patient outcomes.

# Conclusions, Recommendations, and Comments From HHS

HCFA has contributed to the development of most of the relevant available approaches for analyzing Medicare outcomes using existing administrative data. Of the nine distinct approaches that we examined, three, including the two hospital mortality analyses, were developed and used by HCFA's Health Standards and Quality Bureau in its intramural monitoring of the Medicare program. Four of the nine emerged from HCFA's extramural research program. We identified only two independent approaches beyond HCFA's intramural and extramural efforts that were applicable to outcome analyses using existing Medicare data files.

Conclusions about the relative strengths and weaknesses of these nine approaches are constrained by the fact that none has been extensively validated by systematically testing its overall effectiveness in identifying cases or patterns of quality problems. We therefore lack an empirical basis for rating or ranking these approaches in their entirety. However, we can compare the relative strengths and limitations of the approaches on the specific dimensions by which they were assessed.

Looking first at HCFA's intramural efforts, we found numerous improvements in the 1987 analyses of hospital mortality compared to those conducted in 1986. Most notable was the greater sophistication in adjustment for patient severity—shifting from hospital-level to patient-level data, using ICD-9-CM codes instead of DRGs, and incorporating data on comorbidities. We also found the use of statistical techniques and significance tests more appropriate in 1987.

Nonetheless, our comparison of HCFA's intramural analyses with the other six approaches we examined suggests a number of additional improvements that could be made in HCFA's analyses of Medicare outcomes based on existing administrative data. In the discussion that follows, we make five specific recommendations to the Secretary of Health and Human Services (HHS). As part of its review of a draft of this report, HHS provided us with specific comments on each of these recommendations. We have reproduced these comments in full below, immediately following the recommendation to which they apply. Where appropriate, we respond to the Department's comment, before moving on to the next recommendation. The full text of the Department's general and technical comments on our draft report, together with our response to the technical comments, appear in appendix II.

## Improving Adjustments for Patient Severity

Several of the approaches we examined, for example, the Risk-Adjusted Mortality Index and Disease Staging Adapted to Mortality Analyses, provide a more precise adjustment for patient severity than HCFA's intramural analyses have yet obtained.[7] This advantage derives from their use of data on specific principal and secondary diagnoses to adjust for the different mortality risks associated with particular diagnoses. HCFA might well obtain similar advantages by adding to its 1987 hospital analyses one or more predictor variables that capture the differences among discrete diagnoses in mortality risk.

[7]The Risk-Adjusted Monitoring of Outcomes for Nonelective Surgery provides a more precise adjustment for a subset of the Medicare population.

## Recommendation to the Secretary of HHS

We recommend that the Secretary of HHS direct the Administrator of HCFA to strengthen HCFA's analyses of mortality data by testing and incorporating more sophisticated adjustments for patient severity, especially adjustments that exploit more fully the available diagnostic information.

## HHS Comment

"We concur with this recommendation in principle. However, the methodologies mentioned, RAMI and Disease Staging, being based solely on billing data remain subject to substantial criticism. The problem of adequately accounting for interhospital differences in the severity of illnesses of patients will not be satisfactorily resolved until recourse is made to data in the medical record. This latter course has been chosen by HCFA. A severity adjustment tool applicable to four high-risk conditions is being developed by HCFA for use by hospitals in responding to the planned 1988 release of information on outcomes of hospitalization. A longer-range effort, currently being undertaken by HCFA, is the development of a uniform clinical data set which would automatically provide the needed data on the condition of the patient at the time of admission."

## Our Response

We believe that the eventual inclusion of additional clinical information in administrative records should not preclude making better adjustments for patient severity in analyses of currently available data. HCFA's immediate plans to develop a severity adjustment tool for voluntary use by hospitals will apply to just four conditions. Hospitals may refer, if they choose, to the results of medical record reviews that they conduct using this instrument when they comment on HCFA's 1988 analysis of their mortality rates. However, the observed and predicted mortality rates for those hospitals reported by HCFA will not be affected by that analysis. It is uncertain when HCFA's longer range efforts will produce an operational data base with uniform clinical indicators on all patients. In the meantime, HCFA plans to continue conducting outcome analyses using the existing administrative data files. Our recommendation would enable HCFA to make incremental improvements now in these ongoing analyses of Medicare outcomes by using diagnostic information from billing data to improve adjustments for patient severity.

# Limiting the Impact of Random Variation

In analyzing Medicare patient outcome data and applying them in quality monitoring and assurance, HCFA has emphasized assessments of individual hospital performance. One basic problem that all such analyses confront, particularly those that focus on relatively infrequent outcomes such as mortality, is the difficulty of interpreting results based on small

numbers of cases, either for small hospitals with few admissions or for analyses of particular conditions with few cases in a given institution. Because death is typically a low-probability event for most hospital patients, just one or two deaths may cause the observed rate to increase sharply and exceed expectations. Thus the random variation in actual deaths for the total aggregate of patients may push those smaller hospitals that happen to have one or a few deaths in a given year above the level of their "expected deaths" and into the outlier category. In evaluating the results of HCFA's 1986 hospital mortality analyses, a number of PROs observed that some of their outlier hospitals were only outliers for that particular year, and not in preceding or subsequent years.

The solution to the problem of random variation is to analyze more cases, either by consolidating medical or surgical conditions or drawing on a longer time period. HCFA's 1987 hospital mortality analyses incorporated the former strategy through its analyses of 17 broad diagnostic categories. However, these analyses were restricted to patients discharged in 1986, even though comparable data were available for 1984 and 1985.

## Recommendation to the Secretary of HHS

We recommend that the Secretary of HHS direct the Administrator of HCFA to employ data for several years when analyzing outcomes such as mortality rates for small groups of cases across individual hospitals. Hospitals that demonstrate a consistent pattern of observed outcomes that deviate significantly from the expected should be considered the prime candidates for intensified review, as should hospitals whose deviation beyond the range of expected mortality in a single year is based on a number of cases large enough to reduce the effect of random variation.

## HHS Comment

"We concur with and have already implemented this recommendation. Multi-year analyses were performed in preparation for the 1987 release. In fact, such multi-year analyses (1984-1986, and outcomes of first admissions in 1986 in addition to last admissions) were carried out and their results are briefly described in the last paragraph of the Technical Appendix of the release. This effort will be continued."

## Our Response

The Department indicates that it had already implemented this recommendation in the hospital mortality analyses released in December 1987. However, those analyses fall short of what we recommended on two

counts. First, only information on the performance of individual hospitals in 1986 was reported to either the hospitals or the public. The 1984 and 1985 data files were used to assess the overall level of agreement in the designation of outlier hospitals from year to year, with no information presented on the performance of specific hospitals in those years. Consequently, neither consumers nor hospital officials could evaluate the consistency of the information HCFA provided on 1986 mortality rates for individual hospitals with the experiences of those hospitals in previous years.

Second, HCFA's analyses were conducted separately for patients discharged in 1984, 1985, and 1986. Checking the consistency of results from year to year does help to clarify the effect of random variation. But to narrow the range of predicted mortality for smaller hospitals, analysts would have to pool and analyze together the data from several years. That, combined with information on the consistency of a hospital's performance over several years, would produce the most useful information from the available data.

## Expanding Analyses of Patient Subgroups

HCFA's use of outcome analyses in Medicare program operations has concentrated primarily on comparisons of mortality rates among individual hospitals. However, any of the existing approaches that build on patient-level data could be applied to analyses of a wide range of patient and provider subgroups. Therefore, the limited scope of HCFA's efforts to date to analyze patient subgroups does not derive from limitations in available analytical approaches or data. Moreover, the fact that interventions to correct quality problems typically involve either physicians or hospitals does not mean that the initial screening of cases needs to focus in the first instance on individual providers.

For purposes of targeting PRO reviews, the key issue is the number and proportion of cases with genuine problems correctly selected for review, compared to those that should be included but are not and those chosen that prove upon review not to have quality problems. Once a subset of cases has been selected for review, the PROs can perform their usual examination of the medical record and take action against the providers responsible for those cases with confirmed quality problems.

HCFA's own limited comparisons of outcomes across patient subgroups in the HSQB monitoring systems demonstrate that significant differences do exist among groups that vary in sex, age, and presence of comorbidities. These particular variables are typically related to differences in patient

severity; however, physicians may not always be sensitive to the potential impact of such factors on the probability of adverse outcomes. Therefore, analyses that show that 80-year-old men are nearly three times as likely to die from a transurethral prostatectomy as 65-year-old men could help to identify cases where performance of that procedure may have been questionable.

Similar systematic differences in outcomes could potentially be associated with factors not related to patient severity. More extensive comparisons of outcomes across a range of patient subgroups would be needed to determine whether, and where, such differences occurred within the Medicare population.

If HCFA were to expand its analyses based on patient subgroups, preferably with more sophisticated adjustment for patient severity, it could then test whether outcome analyses focused on patient subgroups would usefully supplement, or partially substitute for, hospital-based analyses as a way of targeting PRO quality reviews. Information from subgroup analyses could be used in conjunction with information on the outcomes experienced by individual hospitals to select cases for review; the two strategies are by no means mutually exclusive.

## Recommendation to the Secretary of HHS

We recommend that the Secretary of HHS direct the Administrator of HCFA to expand HCFA's analysis of comparative outcomes among patient subgroups, such as those defined by diagnostic and demographic characteristics. If substantial differences in outcomes among such groups are found after adjusting for differences in patient severity, HCFA should experiment with strategies for targeting quality of care reviews based on these analyses.

## HHS Comment

"We agree that mortality rates alone do not adequately measure how effectively a hospital cares for its patients. Accordingly, we have started an ongoing effort involving all interested parties in improving the accuracy of outcome predictions by refining the model and methodology. We hope to employ other measures which would contribute to assessment of improvements in effectiveness of medical services. These measures would make more important the use of complete files and we are working to define, measure and implement procedures to validate and ensure file accuracy and completeness."

## Our Response

We fully agree with HCFA on the desirability of both developing outcome measures other than mortality and refining the models and methods. However, this recommendation addresses a different point, the potential advantages of targeting quality of care reviews using outcome analyses focused on patient subgroups rather than hospitals. We urge the agency to refine and expand the analyses it has already performed that examine the relative outcomes of patient groups defined by such factors as age, race, diagnosis and presence of comorbidities (i.e., the HSQB monitoring systems described at the end of chapter 2). To the extent that such analyses identify patient subgroups with relatively poor outcomes even after taking differences in patient severity into account, they could prove more effective than analyses of hospital outcomes (or broad diagnostic clusters within hospitals) in targeting cases for quality of care reviews.

## Validation and Periodic Assessments of Alternative Approaches

All nine analytical approaches would benefit from additional testing of their overall validity and effectiveness in targeting cases for review. We have noted above that many individual measures or indicators had not been formally examined for validity, reliability, and sensitivity. We also noted the lack of information on vulnerability to data imperfections.

But beyond this, information is needed on the success of these systems in accurately locating patterns of quality problems. The primary practical objective for any of these approaches is to maximize both the number of cases uncovered where poor-quality care led to poor outcomes (true positives) and the number of cases without quality problems that are appropriately screened out (true negatives). This assessment basically calls for a comparison of the results produced by these analytical approaches to findings derived from independent sources of information. Several of the approaches we examined have begun collecting some of this validating information, primarily through medical record reviews and hospital site visits.

One approach that has received almost no validation through independent data sources is HCFA's 1987 hospital mortality analyses. This reflects two factors: (1) HCFA's decision to develop a new approach to analyzing mortality outcomes, and (2) HCFA's adherence to its self-imposed deadline to release these analyses by December 1987. As the discussion in chapter 2 demonstrates, HCFA had good reason to abandon the heavily criticized approach used in its 1986 analyses. It also chose

not to apply any other existing approaches, including the Risk-Adjusted Mortality Index and Disease Staging Adapted to Mortality Analyses.[6]

The decision to develop an essentially new analytical approach within a constrained time period meant that whatever HCFA devised could receive only limited testing and validation before its release. By the time HCFA had made the final decisions on the analytical procedures it would follow, it had only a brief period in which to produce the preliminary results it needed to send to the hospitals for comment. It had no time to go beyond statistical tests to see how the system actually worked in identifying patterns of quality problems.

The lack of validation for HCFA's 1987 hospital analyses is of particular concern because HCFA moved so quickly beyond the research and pilot-testing applications that generally have characterized the other analytical approaches and publicly issued "information" on all Medicare hospitals. HCFA was careful to use the term "information" rather than "rating" or "performance indicator." It stressed the limited adjustment it made for patient severity of illness, and it encouraged hospitals to comment on the information pertaining to their respective institutions. Nevertheless, HCFA's decision to publish these analyses lent the results a credibility that they may not have deserved. In effect, HCFA determined that, although these analyses were imperfect, they provided some information that was better than none. The main risk inherent in proceeding with an untested approach is that HCFA's method for computing "expected" mortality may inadvertently make some hospitals providing quality care look worse than they should, or suggest all is well in hospitals with substantial quality problems. Without validating evidence, we do not know how often this occurs.

In practice, the rigor and extent of validation required before applying analyses of outcomes depends in part on the proposed uses. Internal analyses providing background to HCFA policymakers need not await full-scale validation, so long as they understand the uncertainties affecting the approaches. By contrast, analyses intended to influence specific decisions by Medicare program officials or beneficiaries should be systematically validated.

[6] Both these approaches were still undergoing refinement in the first part of 1987, but the basic system was in place and some validating evidence (the site visits for RAMI and past experience with earlier versions for Disease Staging) was available or being generated.

Our examination of the relative strengths and limitations of nine specific approaches to analyzing outcome data necessarily reflects what we found at a particular point in time. The continuing genesis of new approaches, and the development of additional information about old ones, means that such findings are likely to change over time. Thus, regular reassessments of available approaches for analyzing Medicare outcomes are needed.

### Recommendation to the Secretary of HHS

We recommend that the Secretary of HHS direct the Administrator of HCFA to assess periodically the relative strengths and limitations of available approaches for analyzing Medicare patient outcome data in terms of their substantive focus, technical adequacy, and degree of validation (i.e., their overall effectiveness in identifying patterns of patient care with quality problems). These assessments should guide the selection of analytical approaches used in future HCFA reviews of Medicare patient outcomes. HCFA should ensure that analyses of Medicare patient outcomes from administrative files employ approaches that have been validated to some degree through independent data sources, and any results publicly released should describe the extent of that validation.

### HHS Comment

"We concur with this recommendation. The proper validation of statistical analyses of variations in outcomes is a difficult matter and was discussed above. All efforts are under continuous review and reevaluation. The usefulness of the techniques used in the 1987 analyses and possible modifications, improvement, or full substitution with an alternative are scheduled to be fully evaluated prior to initiation of 1988 analyses."

### Our Response

We have noted that HHS concurs with our recommendation.

## Assessing the Effects of Inaccurate Data

In evaluating the capacity of these approaches to draw practical inferences about outcomes from existing Medicare data sets, analysts should examine how missing or inaccurate data would affect each approach. They could readily calculate the range of likely effects, but only if they have information on the nature and magnitude of data problems in Medicare's key administrative files. At this point, HCFA does not routinely and systematically collect such information.

The Medicare Statistical System was designed to administer payments to hospitals and physicians, and its data quality procedures have focused primarily on problems that could affect payment amounts. Now that HCFA is using these data for the quite different purpose of monitoring health care outcomes, it needs new systems for measuring the magnitude of errors in the individual data elements.

Moreover, the increasing use of these data for monitoring purposes justifies an effort to correct deficiencies in the data. The threshold for initiating corrective action should vary, depending on the data elements. For example, diagnostic coding, though critical, is likely to be especially prone to error and difficult to correct since it involves (1) questions of medical judgment in initially specifying the diagnoses, (2) the training and judgment of persons abstracting information from the medical record, and (3) additional errors in data entry, editing, and processing. Inconsistencies in recorded dates, by contrast, may be detectable by computer edits and more easily corrected. For each data element, a specified threshold for corrective action could be set based on an assessment of the costs and benefits associated with a given level of completeness and accuracy.

## Recommendation to the Secretary of HHS

We recommend that the Secretary of HHS direct the Administrator of HCFA to evaluate periodically through medical record reviews of a nationally representative sample of Medicare patients the percentage of cases with missing and inaccurate data in the Medicare Statistical System for each of the individual data elements used by HCFA to analyze Medicare outcomes. The results of such assessments should be publicly reported, and corrective action taken for those data elements crucial for reliable outcome analyses. Meanwhile, all analyses of Medicare mortality rates and other outcomes should include an explanation that their findings could be in error by an unknown amount due to potential data inaccuracies.

## HHS Comment

"We concur with this recommendation and have already undertaken studies, mentioned in the report, and other tests to evaluate the appropriate approaches to this problem."

## Our Response

We welcome the Department's concurrence with this recommendation, but wish that its response had been more specific. In our view, none of

the studies described in our report, nor any others that we have identified, represent the kind of systematic effort called for in this recommendation. The only exception would be the Institute of Medicine studies, whose findings from the mid-1970s are now quite out-of-date. In the past, HCFA has typically relied on small-scale, intramural analyses to address relatively narrow issues of data accuracy. Such studies can provide indications of where problems might exist. However, without independent review of medical records for a substantial and representative sample of patients, reliable estimates of the magnitude of any data inaccuracies cannot be generated.

# Assessing the Effect of Prospective Payment on Patient Outcomes

The Senate Special Committee on Aging requested that we examine the potential to use Medicare's administrative data to assess changes in patient outcomes associated with the introduction of the prospective payment system in 1983. Specifically, the Committee asked us to address the following question: Is it feasible for HCFA to use existing patient outcome data to compare the quality of care received by Medicare beneficiaries before the introduction of the prospective payment system (PPS) in 1983 with that provided to beneficiaries under PPS? In this chapter we outline the potential implications of prospective payment for quality of care, discuss the constraints that any study of PPS effects needs to overcome, describe research currently underway to assess the effect of PPS on Medicare beneficiaries, and examine what would be required to go beyond these efforts to obtain more definitive information on the effect of PPS.

## Potential Effects of Prospective Payment on Quality of Care

The Medicare prospective payment system, authorized in 1983, was designed to control inpatient hospital reimbursements ($37 billion in 1983), the largest component of Medicare spending. Prior to the institution of PPS and the related cost controls mandated by the Tax Equity and Fiscal Responsibility Act a year earlier, Medicare typically reimbursed hospitals on an item-by-item basis. Therefore, every extra day in the hospital, or additional service provided, increased the hospital's payment for that case. PPS was structured to eliminate the financial incentive to provide more and more services. Now, with the exception of relatively few "outlier" cases, PPS pays hospitals a fixed amount for each patient, depending on the diagnosis-related group (DRG) to which the patient is assigned, regardless of how long the patient stays in the hospital and the level of services received.

The obvious concern for quality of care raised by this shift in payment systems is that it rewards hospitals for any cutback in services provided, whether or not those services are medically appropriate. Hospitals are still obliged to provide care which meets "professionally recognized standards," and the Peer Review Organizations (PROs) are charged with ensuring that they do. However, the effectiveness of these controls, together with the professional commitment of hospital staff to provide quality care, in counteracting the financial incentives of PPS to restrain services across the board has not been fully assessed.

To the extent that hospitals respond to the financial incentives of Medicare's PPS and seek to maximize the ratio of their payment (by DRG) to

the actual costs of treating patients, their treatment patterns are likely to change in three major ways:

1. Because DRGs do not adjust for severity of illness (except for a payment differential for some diagnoses based on age above 69 or the presence of certain comorbidities), there may be an incentive for hospitals to favor admission of patients with less severe forms of a condition. If less severely ill patients require on average fewer services and shorter hospital stays, they are more likely than other patients to cost the hospital less than the payment it receives from HCFA for cases within that DRG. This could lead both to unnecessary hospitalization (with increased risk of infection or other hospital-induced illness) for less severely ill patients and reduced access to care for more severely ill patients.

2. In an effort to decrease costs, there is an incentive to shorten the time a patient stays in the hospital, which could result in premature discharges that may in turn lead to adverse posthospital health outcomes.

3. Services provided during hospitalization could also be reduced, ranging from decreased intensity of nursing services to the provision of fewer diagnostic tests.

## Constraints Affecting Studies of PPS Effects

The attribution of causation or effect is an intrinsically difficult analytical task, even under the best of conditions. The basic problem is one of ruling out all other possible influences on the outcome of interest, so that the specific effect of one particular set of factors can be determined. In the case of PPS, there are two major circumstances that compound this difficulty, especially when the question is posed years after the change has taken place—the protracted implementation of PPS and the likelihood of systematic changes in diagnostic coding coinciding with the introduction of PPS. These circumstances heavily constrain the available options for analyzing changes in Medicare patient outcomes associated with the shift to a prospective payment system.

In an earlier report, we examined in considerable detail the constraints affecting an assessment of PPS effects and proposed a set of analytical strategies specifically designed to deal with those constraints.[1] Although that report focused on posthospital care rather than acute care, the

[1] U. S. General Accounting Office, Post-Hospital Care: Efforts to Evaluate Medicare Prospective Payment Are Insufficient, GAO/PEMD-86-10 (Washington, D.C.: June 2, 1986).

study designs it outlined would also apply to analyses of changes in inpatient outcomes associated with PPS.

## The Protracted Implementation of PPS

One way of distinguishing between the effects of a programmatic decision like PPS and all other factors influencing Medicare patient outcomes is to focus on those changes in outcomes that occur right at the point when the "intervention" takes effect. If it can be assumed that the influence of the other factors remains roughly constant, while the influence of the programmatic change will be fairly immediate, then discrete shifts in outcomes that occur simultaneously or immediately after the program change occurs can be plausibly linked to that change. Unfortunately, PPS took effect in a highly complex and extended fashion, which precludes a simple comparison of outcomes before and after PPS.

First, PPS itself was preceded by the complicated cost-containment program mandated by the 1982 Tax Equity and Fiscal Responsibility Act, which for many, but not all, hospitals created comparable incentives to cut back on patient services. Second, when PPS itself was introduced in October 1983, the system only took effect for individual hospitals as each began its next fiscal year. So for 11 months following October, some hospitals were being paid prospectively while others continued under cost-based reimbursement. Third, the amounts that hospitals received for each DRG were at first largely based on their own historic costs, with the transition to national DRG rates phased in over 4 years. Given this history, analysts need to consider a very broad period, essentially spanning at least 4 to 5 years, in assessing possible effects of PPS on outcomes. Other factors influencing patient outcomes, including changes in medical technology and treatment patterns and independent efforts to constrain costs mounted by other third-party payers, cannot be assumed to remain constant over such a protracted period.

Sometimes analysts can take advantage of a staged implementation of a program change to help to distinguish its effects from those of other factors influencing the outcomes of interest. For example, they can compare the behavior of hospitals already under PPS with those still operating under cost-based reimbursement. However, such comparisons work best when the effect of the policy change on individual cases is relatively immediate and discrete.

An analysis of the effects of PPS is considerably more complex because, for a number of reasons, hospitals were likely to vary in the timing of PPS-induced changes. For example, hospitals differed in the extent to

which the Tax Equity and Fiscal Responsibility Act had already focused their attention on reducing the costs of entire hospital stays, as distinct from individual services. Similarly, PPS exerted much stronger financial pressure on some hospitals to cut costs than others; those doing as well or better under PPS than before probably were slower to change established practice patterns. Some hospitals, particularly those owned or operated by multihospital corporations, may have made changes in policy and practice in anticipation of PPS before it actually took effect for them. Hospitals also vary in the degree of influence their administrators have over the behavior of attending physicians, which could affect the speed with which changes occurred in patterns of admission, discharge, and the use of ancillary services.

## Confounding Due to Systematic Changes in Diagnostic Data

In making causal attributions, the most difficult alternative explanations to rule out are those that change at the same time as those we are attempting to evaluate. In the case of PPS, there is considerable evidence to support the suspicion that the coding of diagnoses, a key source of information for patient severity adjustments, systematically changed as PPS took effect. This simultaneous shift was a logical consequence of PPS, because hospital payment was now affected by both the accuracy and number of diagnoses recorded. One Rand study estimated a 6.2-percent increase in the "case-mix index" of Medicare during PPS' first year specifically resulting from changes in hospital coding and documentation.[2]

Changes in diagnostic coding could result from several contradictory tendencies. One is an increase in the overall accuracy of coding, because HCFA, through the PROs, began verifying the diagnostic codes used in DRG assignments once PPS was introduced. A related trend would be increased thoroughness in recording secondary diagnoses, particularly those on the DRG complications and comorbidity list, which again could affect DRG assignment and therefore the amount of hospital payment.

For more complicated cases with several comorbidities, the appropriate designation of principal and secondary diagnoses may not always be straightforward. Before PPS, hospitals had no particular incentive to pick one diagnosis over another as the principal diagnosis. However, as their payment under PPS now depends on their choice of principal diagnosis, hospitals stand to gain financially by systematically favoring those diagnoses that result in higher paying DRG assignments. So the

[2]Grace M. Carter and Paul B. Ginsburg, The Medicare Case Mix Index Increase: Medical Practice Changes, Aging, and DRG Creep (Santa Monica, Calif.: Rand Corporation, June 1985), p. 27.

introduction of PPS created a complex set of incentives for both increased and decreased accuracy in the designation of patient diagnoses.

There is little available information as to the timing and magnitude of actual changes in diagnostic coding brought about by these new incentives. The last systematic investigation of coding accuracy (the Institute of Medicine studies described in chapter 2) examined data that predated the implementation of PPS by nearly a decade. However, recent evidence from the second year of PPS' implementation indicates both a substantial error rate in DRG assignments—21 percent of all Medicare cases—and a clear predominance of errors (62 percent) that increase hospital revenues.[3]

Any investigation of trends in outcomes over an extended period of time has to contend with changes not only in the accuracy of diagnostic codes linked to the transition to PPS, but also in the way these data have been recorded over time. In 1979, the generally accepted system for coding diagnoses shifted from ICDA-8 to ICD-9-CM. Differences between the two systems in how diagnoses are classified and coded complicate longitudinal analyses of hospital outcomes. Moreover, prior to 1981, hospitals provided narrative descriptions of diagnoses on their Medicare billing forms, from which HCFA derived specific diagnostic codes for a 20-percent sample of Medicare hospital patients. Beginning in 1981, providers had the option of reporting full diagnostic codes in lieu of narrative summaries. Concurrent with the initiation of PPS in 1983, hospitals were required to report full ICD-9-CM codes on all inpatient hospital bills. Therefore, a significant first step in attempting to compare outcomes before and after PPS would be to establish the comparability of diagnostic data over time.

The problems with diagnostic coding might not pose serious analytical difficulties if analysts could simply compare outcomes before and after PPS directly and ignore diagnoses. However, the only outcome for which data are readily available from Medicare data sets is mortality, and here, some adjustment for patient severity is essential for making valid comparisons. All the approaches examined in the previous chapters relied on diagnostic data to make such adjustments. Given the extended period over which PPS effects could arise, we cannot safely assume that the characteristics of patients in hospitals, in terms of their diagnoses

[3]David C. Hsia et al., "Accuracy of Diagnostic Coding for Medicare Patients Under the Prospective-Payment System," New England Journal of Medicine, vol. 318, no. 6 (Feb. 11, 1988), pp. 353-54.

and severity of illness, have not changed. In fact, available evidence suggests that both have shifted substantially in this period.[4]

In short, the diagnostic data required to assess any changes in outcomes associated with the implementation of PPS have most likely been substantially affected as well by PPS. Therefore, to evaluate changes in outcomes attributable to PPS, analysts need to correct any biases in the data created by these simultaneous changes. This would require much more detailed information about the nature and magnitude of changes in data recording that occurred during this time period than is currently available.

## Current HCFA Assessments of PPS Effects

HCFA has funded two major research projects intended to examine the effects of PPS on hospital care. One of these, a study by Abt Associates of PPS-induced changes in hospital service utilization, looks for indications of the types of changes in treatment patterns that PPS would logically tend to promote. The results of these analyses are expected to be released in early 1988. A second study, conducted by the Rand Corporation, addresses the issue of patient outcomes and quality of care more directly, but does so largely through medical record reviews supplemented by analyses of Medicare claims data.[5] The final report for this study is due in late 1988. The first study does not specifically relate to outcomes; the second relies primarily on data sources other than administrative files. Nevertheless, both will provide information relevant to an assessment of the effects of PPS on Medicare beneficiaries.

## Abt PPS Evaluation Analyses

As part of a $5 million contract to investigate a wide range of PPS effects, Abt Associates has undertaken a series of quantitative analyses for HCFA that use Medicare administrative data to look for evidence of PPS-induced effects on the use of hospital and nonhospital services. They involve the construction of several different data files from Medicare claims data and related sources, focusing on hospitals, patients, and beneficiaries for the years 1981-1985. Using a variety of measures and grouping strategies, Abt researchers are examining the extent to which the changes in hospital treatment patterns that PPS would logically

[4]Jonathan E. Conklin and Robert L. Houchens, "PPS Impact on Mortality Rates: Adjustments for Case Severity," Final Report, HCFA contract no. 500-85-0015 (Santa Barbara, Calif.: Oct. 6, 1987), p. 39.

[5]For a third perspective on PPS-induced effects on hospital outcomes viewed in terms of "outputs"; i.e., discharged to self-care, to nursing home or home health care, or dead, see Michael J. Long et al., "The Effect of PPS on Hospital Product and Productivity," Medical Care, vol. 25, no. 6 (June 1987), pp. 528-38.

promote have in fact come about. These include shifts in types of patients admitted, the lengths of hospital stays and days in intensive care units, use of ancillary services, rates and patterns of transfers and readmissions, and substitution of outpatient for inpatient care.

Abt is examining trends through a comparison of descriptive statistics for relevant groups and through multivariate modeling to adjust statistically for the effect of non-PPS factors on utilization. Much of the analysis concentrates on 10 specific hospital markets, so that detailed information about the characteristics of those markets can also be entered into the analysis. The results of these analyses are scheduled to appear in HCFA's 1986 Annual Report to Congress on the effects of PPS.

The Abt study does not attempt to address the issue of quality of care directly; that is, by identifying whether a given service was or was not medically necessary or appropriate or by assessing outcomes. However, several of its analyses investigate the degree to which utilization changes have focused on vulnerable patient groups, such as the disabled, those over 84-years-old, those receiving long-term care, and those with specific tracer conditions (e.g., pneumonia, stroke, total hip replacement, hip pinning, inguinal hernia repair, congestive heart failure, and transurethral prostatectomy).

Abt explicitly considered the constraints on studies of PPS effects described above and proposed a range of approaches to ameliorate their impact. A number of the analyses use quarterly data, which enables Abt to designate fairly precisely when both the Tax Equity and Fiscal Responsibility Act and the prospective payment system came into effect for specific hospitals. This allows Abt to build the staggered start of PPS across hospitals into the design of an interrupted time-series analysis, which helps to distinguish the effect of PPS from other factors influencing utilization over the same time period. Abt also assesses PPS effects through comparisons with hospitals in "waiver states," where PPS was not implemented (because the state had its own cost-containment program in place) and through a separate variable designed to measure the magnitude of PPS' financial consequences for individual hospitals. Abt recognizes that although these analytical strategies are helpful, they do not fully resolve the difficulties inherent in attributing observed changes in utilization specifically to PPS.

Abt also acknowledges the problems raised by probable distortions in available diagnostic data. The Abt study plan basically accepts the

absence of reliable data on diagnoses as a major limitation that precludes analyses of changes in patient severity. This in turn constrains the conclusions that the study can reach about the implications of observed changes in hospital utilization. However, Abt conducted a small-scale pilot study based on medical record reviews that found some increase in the severity of cases admitted to hospitals for three conditions after PPS took effect.

## Rand PPS Quality Impact Study

HCFA has also contracted with the Rand Corporation to conduct a 3-year, $4.5 million study that focuses specifically on the question of potential quality of care effects from PPS. The study is assessing the quality of care received by 17,000 Medicare patients through detailed reviews of their medical records. Half the cases are drawn from 1981-1982 discharges and half from 1985-1986. Rand is sampling in five states (California, Texas, Pennsylvania, Florida, and Indiana) the patient records of those hospitalized for one of six specific conditions: myocardial infarction, congestive heart failure, pneumonia, hip fracture, cerebrovascular accident (e.g., stroke), and depression. For each of these conditions, Rand has drafted highly detailed protocols that assess both the process and outcome of care in terms of the specific clinical circumstances experienced by individual patients.

Rand developed its protocols through an elaborate process of literature reviews, extended discussions in expert consensus panels, and extensive pilot testing. The protocols provide detailed but standardized information on both the severity of the case and the extent to which management of the case conformed to quality standards. They assess the process of care in terms of six stages of hospital treatment from diagnosis to discharge, including the treatment of any comorbidities. The outcomes monitored during the hospital stay are mortality, major and minor complications, and medical stability as well as functional status at discharge. In addition, Rand follows the Medicare claims data for all cases for 6 months after the hospital stay to determine subsequent mortality, readmissions, and movement into nursing homes. Rand will combine individual items from these protocols to produce composite process of care and outcome quality scores.

By reviewing medical records, Rand avoids most of the problems associated with diagnostic coding in computerized patient abstracts. The original list of hospital patients from which cases are sampled depends on the recorded diagnoses, but the Rand abstracters determine through a preliminary review of the record whether that case actually meets the

Rand criteria for inclusion. Rand has structured its data collection to permit testing for changes in the accuracy and thoroughness of recording between 1981 and 1986 that could affect the study's results.

A clear strength of the study is the highly detailed, condition-specific assessment of quality based on medical record reviews. The findings will be limited, however, to six medical conditions. Rand selected six conditions that are relatively common—and therefore directly represent a substantial number of Medicare patients. Three other prerequisites imposed by the study's design were that the conditions have well-defined criteria for diagnosis, that diagnostic and treatment practices did not change appreciably between 1981 and 1986, and that needed information on patient severity, process of care, and outcome be consistently recorded in medical records. Only one of the six conditions involves surgical treatment, even though the study plan originally called for an equal number of medical and surgical conditions.

In contrast to the sophistication with which the quality of care for individual cases is assessed, the Rand study deals with the problem of PPS's protracted implementation through a fairly rudimentary before-and-after PPS comparison. Focusing on conditions that did not change appreciably in diagnostic or treatment practice during that period helps somewhat in screening out other potential influences on quality, as does the fact that the same hospitals are sampled for the 1981-1982 and 1984-1985 cases. Nevertheless, the design of the Rand study will inherently provide a stronger basis for assessing changes in patient severity and quality of care over time for those six conditions than for attributing those changes, if any, to PPS.

## Potential for Further Analyses

The work that HCFA has underway represents a substantial commitment of resources to evaluating the effects of PPS on Medicare patients. The two studies described above deal in different ways with the intrinsic problems facing any study of PPS effects, and each has corresponding strengths and weaknesses. Are there alternative approaches available, using existing data, that could provide more direct information on outcomes than the Abt PPS Evaluation Analyses? Could such approaches profitably expand on the information already generated by the Rand PPS Quality Impact Study?

To some extent it is premature to pose these questions. Both the Abt and Rand studies are in progress. When their results are available we will be in a better position to judge what was learned from them and how their

analyses might be usefully extended. Nevertheless, the designs of these studies make clear that the Abt research will not be able to link changes in utilization patterns to outcomes, while the Rand findings can only cautiously be projected to other types of medical problems.

With respect to the question of available alternative approaches, of those examined and compared in chapters 2 to 6, one would appear particularly relevant to a before-and-after PPS analysis. This is the approach, also developed by Abt Associates, for the National Hospital Rate-Setting Study discussed in chapter 4.

In part because it was designed to answer a similar set of questions about state prospective reimbursement systems, the approach of the National Hospital Rate-Setting Study is potentially well-suited to addressing the issue of PPS effects on Medicare outcomes. Many of the state prospective reimbursement programs evolved over several years, thus the study is structured to assess the effect of a specified intervention over an extended period of time. Its adjustments for patient severity, although limited and perhaps distorted by an excess of predictor variables, have the advantage of depending somewhat less than other systems on diagnostic information. Secondary diagnoses are not used and primary diagnoses serve mainly to identify "urgent care" and "elective surgery" cases.

Moreover, the data sets created for the National Hospital Rate-Setting Study, extending from 1974 to 1983 and including all states (either individually as prospective reimbursement states or in the residual nonprospective reimbursement group), could be consolidated to form an extensive set of national baseline data against which to evaluate post-PPS changes. In other words, given relatively little effort to consolidate and extend the National Hospital Rate-Setting Study data sets for another 3 or 4 years, an assessment of national PPS effects on patient mortality and readmissions could be made quite analogous to those performed for the state programs.[6]

Extension of the National Hospital Rate-Setting Study to the Medicare PPS program would be feasible and relatively inexpensive; however, the overall validity of this approach requires a judgment based on more

[6]However, the analysis would focus more exclusively on changes over time. Because PPS was a national program, comparisons across states would be limited to contrasts with a few states where PPS was "waived." Moreover, comparison with these states would not provide a particularly good basis for examining PPS effects—these states had the option not to adopt PPS precisely because they had their own programs designed to achieve similar results.

information than is currently available. As noted in chapter 4, the designation of "urgent care" conditions, while intuitively appealing, has yet to receive empirical testing. Moreover, the problem of systematic errors in diagnostic coding would continue to bias the results of this approach, if perhaps somewhat less strongly than approaches that attempt to draw more out of diagnostic information. In short, substantial questions remain on whether this approach would produce a valid comparison of outcomes between the period immediately before and after PPS.

Ultimately, given the paucity of other relevant information in existing administrative data sets and the critical need to adjust for variations in patient severity in comparing outcomes, there is no way to finesse the problem of systematic error in patient diagnostic codes. Credible analyses of outcomes using administrative data depend on obtaining accurate information about principal diagnoses and any comorbidities. Such information does not now exist for the period immediately preceding and following the introduction of PPS. Conceivably, such information could be reconstructed, but it would require an extensive reabstraction of medical records on a wide scale, both across time and diagnoses. Moreover, any patient subgroups of potential interest would have to be well represented in the sample of reabstracted cases. The effort that this would require might more usefully be placed in other areas, such as assessing the accuracy of data currently recorded for Medicare cases as recommended in chapter 6.

The frustrations currently faced in attempts to analyze the effects of PPS underline the importance of systematically collecting baseline data prior to any major program changes. By planning an evaluation of the programmatic change in advance, and designing a data collection strategy to generate consistent, accurate, and relevant data prior to the implementation of the change, much stronger inferences about causation can be made. Similarly, efforts to maintain the accuracy and consistency of routine program monitoring data, and to implement and document uniformly any alterations in those data over time, will pay dividends in future uses of those data for assessing the effects of program changes.

To summarize, analyses of the effects of PPS on Medicare patient outcomes could be performed using existing administrative data. However, their results would remain open to challenge owing to at least two factors: the lack of comparable diagnostic data from the periods immediately before and after PPS upon which to base adjustments for any changes in patient severity, and the intrinsic difficulty of isolating the

effects of PPS from other factors influencing Medicare patient outcomes during its protracted implementation. Much of what can be learned about PPS effects on Medicare beneficiaries could emerge from the ongoing HCFA studies. Once those studies have been completed, the potential benefits of further analyses using administrative data can be more fully assessed.

# Criteria for Evaluating Approaches to Analyzing Medicare Outcome Data

Name of Approach:

Names of Principal Developers:

Source Documents:

a.

b.

A. General Purpose of the Approach

1. What is the basic unit of analysis: patient, provider, or other?

2. For or across what groups have outcome measurements and comparisons been made: hospitals, physicians, types of patients, Medicare beneficiaries, etc.?

3. Has the approach been used for longitudinal analyses, cross-sectional, or both?

B. Substantive Focus of Approach

1. What specific health care outcomes are monitored?

a.

b.

c.

2. For what range of medical conditions has this approach been applied?

1

3. How are the medical conditions operationalized (e.g., DRGs, consolidated DRGs (describe), "cleaned" DRGs (describe), clusters of ICD-9-CM codes, etc.)?

4. Are there any additional conditions that the proponents of the approach claim it could be extended to without additional development? If so, list.

5. On what population has the approach been tested (e.g., types of patients and providers, geographic location, time period, etc.)?

6. If a sample was employed, what were the characteristics and resulting limitations of the sampling procedure (e.g. size of sample, numbers of sites and strata, degree of randomization, etc.)?

7. What patient subgroups have been separately analyzed using this approach? Any problems with subgroup definition or identification?

8. What additional subgroups could be analyzed without further development of the approach?

C. Severity Adjustment

1. Are the outcomes adjusted for individual patient risk or severity of illness at time of hospital admission? If so, how? Does the adjustment focus on the acuteness of the principal diagnosis, the presence of unrelated comorbidities, or both? (List all independent variables entered into any model used for risk adjustment.)

2. Does the adjustment involve direct or indirect standardization techniques?

2

3. Do the authors specify criteria for the inclusion or omission of factors used in their procedures for the adjustment of risk? If so, what are the criteria (e.g., empirical analyses, a priori decision rules, clinical judgment, theory)? Were any factors likely to be relevant (based on their prominence in the quality of care literature as a whole) neglected or rejected inappropriately or inexplicably?

4. Does the adjustment distinguish between complications or comorbidities present at admission and those that develop during the hospital stay? If so, how?

5. Does the adjustment account for variations in admission stringency within similar diagnoses across providers? If so, how?

6a. What evidence is presented that any independent variables used for severity adjustment represent a valid indicator of individual patient risk prior to admission?

6b. If a model is used for this purpose, has it been tested for bias? If so, what tests were used and what did they show?

6c. Are any control variables used of questionable validity, such as hospital characteristics, geographic location, or race?

D. Technical Adequacy: Measurement Issues

1. Outcome Validity: Excluding severity issues, what evidence is presented that the outcomes tracked represent valid indicators of quality of care for the particular application in question (note evidence for each variable)? What are the major limitations of this evidence?

2. Reliability: What evidence is presented that the outcome measures and indicators used as independent variables are reliable for the particular application in question (note evidence for each variable)? What are the major limitations of this evidence?

3

3. Sensitivity: What evidence is presented that the outcome measures and indicators used as independent variables are sensitive measures for the particular application in question (note evidence for each variable)? What are the major limitations of this evidence?

E. Technical Adequacy: Analysis Issues

1. What specific analytical techniques are applied in this approach (e.g., linear regression, logistic regression, recursive partitioning, log linear modeling, discriminant analysis, etc.)?

2. What assumptions does this analytical technique make about either the nature of the data elements individually or in relation to each other (e.g., linearity, normally distributed, independent)? How do these assumptions affect the application of this approach?

3. Does the analytical technique make any inappropriate assumptions about the types of variables used, such as treating an ordinal scale as an interval or ratio scale?

4. Does the analytical technique employ significance tests? If so, are they used appropriately (e.g., using Poisson-based tests for small samples of discrete events)? How vulnerable is the approach to Type 1 (false positive) and Type 2 (false negative) error?

5. Are proper adjustments made for any known limitations in the reliability of the measures used?

F. Data Quality Issues

1. What specific data elements does the approach require?

2. What information is available on the effect of missing, inconsistent or inaccurate data on the results produced?

3. Does the approach incorporate any techniques to detect and compensate for missing, inconsistent or inaccurate data?

G. Extent of Validation

1. What independently derived evidence (e.g. from different data sources, such as medical record reviews) has been developed which tests the effectiveness of the approach in identifying cases, or aggregates of cases such as hospitals, with genuine quality of care problems?

H. Summary

1. What known limitations (if any) would need to be overcome to apply this approach for monitoring quality of care in the national Medicare population?

2. In what areas is information lacking on which to base an evaluation of this approach in terms of its potential use to monitor quality of care in the Medicare population (identify areas in terms of the above questions)?

3. If applied to existing Medicare data sets, what specific questions about quality monitoring or assurance could the approach address (e.g., target poor providers for intensive review, identify vulnerable patient populations, etc.)?

4. Considering the characteristics of the measures and analytical techniques employed in this approach, to what extent should any such inferences about quality of care received by Medicare beneficiaries be qualified or constrained?

5

# Comments From the Department of Health and Human Services

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

**DEPARTMENT OF HEALTH & HUMAN SERVICES** Office of Inspector General

Washington, D.C. 20201

MAR 21 1988

Mr. Lawrence H. Thompson
Assistant Comptroller General
U.S. General Accounting Office
Washington, D.C. 20548

Dear Mr. Thompson:

Enclosed are the Department's comments on your draft report, "Medicare: Improved Patient Outcome Analyses Could Enhance Quality Assessment." The enclosed comments represent the tentative position of the Department and are subject to reevaluation when the final version of this report is received.

The Department appreciates the opportunity to comment on this draft report before its publication.

Sincerely yours,

Richard P. Kusserow
Inspector General

Enclosure

Comments of the Department of Health and Human Services
on the General Accounting Office Draft Report,
"Improved Patient Outcome Analyses could Enhance Quality Assessment"

Overview

At the request of the ranking minority member of the Senate Special Committee on Aging, GAO examined the Health Care Financing Administration's (HCFA's) approach to analyzing the quality of care received by Medicare patients. The primary question was whether HCFA could obtain more or better information to guide Medicare quality assurance activities, using the administrative data on individual patients that it already collects. The study focused on five objectives:

-- to describe the analytical approaches HCFA currently employs to analyze existing Medicare administrative data on mortality and morbidity as an indicator of the quality of hospital care;

-- to examine the uses that HCFA has made of these outcome analyses to guide quality assurance in the Medicare program;

-- to identify other approaches for conducting outcome analyses which could be applied to Medicare administrative data;

-- to assess the relative strengths and limitations of HCFA's and other approaches in terms of their substantive focus and technical quality; and

-- to determine the feasibility of analyzing administrative data to assess changes in Medicare outcomes associated with the introduction of the Prospective Payment System in 1983.

GAO reports that a comparison of the 1986 and 1987 hospital mortality analyses shows that HCFA has strengthened the technical quality of its intramural analyses of Medicare outcomes based on administrative data. HCFA's application of these analyses has so far been limited, and not notably effective in identifying quality problems. In comparing HCFA's intramural analyses with six relevant other approaches, GAO found that additional improvements could be made in the key area of patient severity adjustment. Further, future analyses of Medicare outcomes would be more credible and useful if the analytical approaches selected were more fully validated, and the data which they analyze systematically checked for accuracy and completeness.

While we believe the report to be quite thorough and scholarly in its consideration of very complex issues and in its description of a number of useful and practical recommendations, we have developed technical comments which we believe deserve consideration in finalizing the report. These comments follow our response to GAO's recommendations.

GAO Recommendation

That the Secretary of HHS direct the Administrator of HCFA to strengthen HCFA's analyses of mortality data by testing and incorporating more sophisticated adjustments for patient severity, especially adjustments which exploit more fully the available diagnostic information.

Department Comment

We concur with this recommendation in principle. However, the methodologies mentioned, RAMI and Disease Staging, being based solely on billing data remain subject to substantial criticism. The problem of adequately accounting for interhospital differences in the severity of illnesses of patients will not be satisfactorily resolved until recourse is made to data in the medical record. This latter course has been chosen by HCFA. A severity adjustment tool applicable to four high-risk conditions is being developed by HCFA for use by hospitals in responding to the planned 1988 release of information on outcomes of hospitalization. A longer-range effort, currently being undertaken by HCFA, is the development of a uniform clinical data set which would automatically provide the needed data on the condition of the patient at the time of admission.

GAO Recommendation

That the Secretary of HHS direct the Administrator of HCFA to employ data for several years when analyzing outcomes such as mortality rates for small groups of cases across individual hospitals. Hospitals which demonstrate a consistent pattern of observed outcomes deviating significantly from the expected should be considered the prime candidates for intensified review, in addition to hospitals whose deviation beyond the range of expected mortality in a single year is based on a large enough number of cases to reduce the effect of random variation.

Department Comment

We concur with and have already implemented this recommendation. Multi-year analyses were performed in preparation for the 1987 release. In fact, such multi-year analyses (1984-1986, and outcomes of first admissions in 1986 in addition to last admissions) were carried out and their results are briefly described in the last paragraph of the Technical Appendix of the release. This effort will be continued.

GAO Recommendation

That the Secretary of HHS direct the Administrator of HCFA to expand HCFA's analysis of comparative outcomes among patient subgroups, such as those defined by diagnostic and demographic characteristics. If substantial differences in outcomes among such groups are found after adjusting for differences in patient severity, HCFA should experiment with strategies for targeting quality of care reviews based on these analyses.

Page 3

Department Comment

We agree that mortality rates alone do not adequately measure how effectively a hospital cares for its patients. Accordingly, we have started an ongoing effort involving all interested parties in improving the accuracy of outcome predictions by refining the model and methodology. We hope to employ other measures which would contribute to assessment of improvements in effectiveness of medical services. These measures would make more important the use of complete files and we are working to define, measure and implement procedures to validate and ensure file accuracy and completeness.

GAO Recommendation

That the Secretary of HHS direct the Administrator of HCFA to assess periodically the relative strengths and limitations of available approaches for analyzing Medicare outcome data in terms of substantive focus, technical adequacy, and degree of validation (i.e., their overall effectiveness in identifying patterns of patient care with quality problems). These assessments should guide the selection of analytic approaches used in future HCFA reviews of Medicare outcomes. HCFA should ensure that analyses of Medicare outcomes from administrative files employ approaches which have been validated to some degree through independent data sources, and any results publicly released should describe the extent of that validation.

Department Comment

We concur with this recommendation. The proper validation of statistical analyses of variations in outcomes is a difficult matter and was discussed above. All efforts are under continuous review and reevaluation. The usefulness of the techniques used in the 1987 analyses and possible modifications, improvement, or full substitution with an alternative are scheduled to be fully evaluated prior to initiation of 1988 analyses.

GAO Recommendation

That the Secretary of HHS direct the Administrator of HCFA to evaluate periodically through medical record reviews of a nationally representative sample of Medicare patients the percentage of cases with missing and inaccurate data in the Medicare Statistical System for each of the individual data elements used by HCFA to analyze Medicare outcomes. The results of such assessments should be publicly reported, and corrective action taken for those data elements crucial for reliable outcome analyses. Meanwhile, all analyses of Medicare mortality rates and other outcomes should include an explanation that their findings could be in error by an unknown amount due to potential data inaccuracies.

Department Comment

We concur with this recommendation and have already undertaken studies, mentioned in the report, and other tests to evaluate the appropriate approaches to this problem.

See comment 1.

See comment 2.

Now on p. 27.

See comment 3.

See comment 4.

Technical Comments

We believe that the report is so narrowly technical that it misses or, at least, takes for granted the most significant aspects of HCFA's recent efforts in this area. With the initiation of analyses of hospitalization-associated mortality rates and of the monitoring of outcomes, HCFA, prior to the issuance of the congressional mandates to undertake research on the evaluation of outcomes, shifted the focus of attention from process-based to outcomes-based assessment of the quality of medical care. In addition, the adoption of longitudinal follow-up techniques for the assessment of outcomes and the identification of mortality, morbidity, disability and cost as the outcomes to be measured has resulted in a sound and practical approach for the determination of the effectiveness of medical practices. We believe these points are lost in the mass of technical detail in the report. Furthermore, we believe that, as the 1987 analysis represents an analytical advance over the 1986 analysis, it is largely beside the point to critique the methodology used for the earlier study.

The GAO report's narrowly critical perspective is illustrated by a comment on page 2-17 concerning the validity of in- patient mortality: "Critics of this measure point to its sensitivity to variations in average lengths of stay. . . HCFA acknowledged this problem . . . ." Not stated is the fact that this problem was recognized by HCFA staff in the course of the 1986 analyses and that HCFA staff were the most vocal and vehement critics in the health services research community. Indeed, it was only the persistence of HCFA staff in discussions with its statistical and health services research consultants that prevented inpatient mortality from being used again in the 1987 analyses.

On another point of detail, in describing the monitoring studies, GAO appears to distinguish readmissions from morbidity. In fact, morbidity is defined in these studies as the deterioration of health to the point requiring medical intervention, whether in the inpatient or the outpatient setting. Therefore, the analysis of morbidity subsequent to an intervention such as a hospitalization has two components; i.e., readmissions and ambulatory care.

The methodology adopted for the analysis of variations in mortality rates associated with hospitalization was not solely the choice of HCFA staff. The techniques of Blumberg, CPHA and of the Rand NOS project were presented to statisticians who functioned as contract consultants and to external statistical advisers as options for active consideration. Similarly, options for the classification of patients were presented, together with supporting statistical analyses of predictive power and stability, to a group of clinical advisers. The methodology and the classification scheme adopted were recommended by these external advisers.

The 1986 analyses were also reviewed by external advisers and were modified prior to the issuance of the Peer Review Organizations (PROs) Request for Proposal as a result of their recommendations. (The introduction of the adjustment for state-average length of stay was, however, purely a HCFA initiative because, at the time they were convened, neither the advisers, nor HCFA staff for that matter, were aware of the extent to which inpatient mortality was a biased measure.)

The GAO report considers at length whether the 1986 release of "outlier hospitals" was of any practical use. The answer to this question consists of two parts: (1) whether the data were intrinsically useful, and (2) whether the PROs were able to make effective use of the data. We think the data were quite useful and we would cite interesting data from a New York State Department of Health report dated December 1987. On pages v and vi of the Executive Summary, it is stated that, in a ". . . study to test and improve upon the HCFA model . . .," when "non-targeted" cases were reviewed, ". . . 1 percent had care that caused or contributed to patient deaths . . ." and that ". . . 2.6 percent of the cases in outlier hospitals (identified in the HCFA release) were found to have problems that caused or contributed to patient deaths." Although these statistics are not tested for statistical significance, they tend to show that there are discernible differences in outcome in outlier and non-outlier facilities. In addition, New York State did find that by ". . . targeting rare-death DRGs. . .," they found 5.3 percent of cases in outlier hospitals had problems contributing to patient deaths. Regarding the utility of the information for PROs, the answer varies with the PRO. Some found it useful, some did not. There is, however, no formal study which describes the relative usefulness.

See comment 5.

See comment 6.

GAO is emphatic on the need for validation but not completely clear on how it is to be carried out, although it strongly implies that the method of choice is medical record review, as was carried out by the New York State Department of Health. We recognize that such an approach to validation is supported by the research community and has been the mainstay of review of quality of care. However, we are continuing to study this issue as well as other options as a means to validate the data.

See comment 7.

The GAO criticism of the limited explanatory power (R-square) of the models for the individual diagnostic categories in the 1986 and 1987 analyses, compared to the model for overall mortality, is based on a fundamental misunderstanding. GAO states clearly the objectives of the modeling; i.e., to account as completely as possible in the models for factors other than the quality of care so as to be able to examine specifically the variations in mortality rates attributable to variations in the quality of care. The explanatory power of a model under these circumstances will depend on how much variation is sought to be

See comment 8.

explained. Thus, the model for overall mortality must deal with variations in mortality rates due to variations in diagnostic case-mix and does so by specifying covariates which classify patients into diagnostic categories which differ among themselves in risk of death. Hence, much variation needs to be explained in overall mortality, and is explained by the diagnostic categories. However, this variation does not exist within diagnostic categories and, hence, the model explains much less of the variation. In both instances, it is the residual, unexplained variation that is of interest because it represents the variations in the quality of care, or would if severity of illness were adequately accounted for in the model. There is, therefore, nothing "wrong" with the low R-square of the models for the individual diagnostic categories. That they are considerably lower than that for the model for overall mortality is fully expected and quite appropriate.

See comment 9.

Finally, the GAO report has overlooked an additional benefit of these outcome studies; namely, to assess the impact of PPS on quality of care. These studies provide information on trends over time and the analyses performed to date cover the period of time over which the impact of PPS can be assessed. These studies characterize the trends in population-based and post-hospitalization mortality rates, in readmission rates and volumes of ambulatory (morbidity) and supportive care (disability). In addition, the detailed data they provide permit objective assessment of a number of the points upon which GAO speculates; e.g., the possibility of coding creep.

The following are GAO's comments on the Department of Health and Human Services' letter dated March 21, 1988.

## GAO Comments

Overall, HHS found the report "thorough and scholarly" and it generally concurred with our recommendations. We have addressed in chapter 6 the specific comments submitted by the Department for each of these recommendations. Here we respond to the additional "Technical Comments" from the Department provided on our draft report.

1. The Department apparently believes that the scope of the report is overly narrow or technical and ignores some of HCFA's most significant initiatives in the area of quality assurance. However, the scope of this study was defined by our request from the Senate Special Committee on Aging, and the issues addressed in the initiatives mentioned by the agency—the relative merits of outcome-based as opposed to process-based assessments of quality of care and the potential usefulness of longitudinal outcome analyses for determining the effectiveness of different medical practices—lie outside that scope.

While we appreciate and welcome the agency's comments for putting the report's findings into a broader perspective, we believe that these laudable efforts should not obscure the relevance of improving analyses of HCFA administrative data as they currently exist. The agency plans to continue conducting such analyses—most notably the annual assessment of hospital mortality rates—while the long-term development of better measures and improved data sets goes on. The findings and recommendations of our report bear most directly on what could be done in the relatively near term to enhance those interim analyses. In addition, the steps we recommend should strengthen the basis for making more fundamental improvements, by creating established procedures for systematically and periodically checking the accuracy of whatever data are analyzed and for assessing the validity of the analytical approaches used to examine those data.

2. We examined HCFA's first set of hospital mortality analyses in some detail for two reasons. First, this enabled us to determine whether the more recent analyses were in fact "an analytical advance," and to pinpoint where and how improvements had been made and where they were less evident. Second, the general approach of analyzing hospital-level data with multiple linear regression continues to have adherents

within and outside of HCFA. Therefore, strengths and limitations associated with this approach are still relevant, even if the particular analyses conducted by HCFA in early 1986 are unlikely to be replicated.

3. The Department appears dissatisfied with our description of the way HCFA defined morbidity in its monitoring studies. In fact, the report states quite explicitly that HCFA analyzed morbidity in terms of hospital readmissions and the costs of ambulatory care. (See p. 41.) However, the report goes on to raise some questions about the validity of both these indicators as measures of the core concept of morbidity.

Readmissions are an imperfect gauge of morbidity because factors other than health status or severity of illness enter into decisions to admit patients to the hospital. These factors include variations in medical practice among physicians, as well as differences among patients in their desire for hospital-based treatment and access to nonhospital alternatives (such as home health care by professional caregivers or family members).

Similarly, the costs of ambulatory care imperfectly reflect morbidity because of variations in the types and amounts of treatment that patients desire, as well as in the charges made by different physicians for equivalent services. Some morbidity may not be detected by either rehospitalization or ambulatory care charges, depending on patient proclivity to seek care and the availability of alternative forms of treatment other than acute inpatient and ambulatory care.

4. At several points in the report, we note HCFA's consultation with external statistical and clinical experts. (See p. 19 and 24.) Our main criticism of HCFA's implementation of the 1987 hospital analyses did not concern a lack of input from outside experts, but rather, the lack of validation based on independent data sources, such as hospital site visits or medical record reviews. Experts could help HCFA to choose wisely among alternative analytical strategies; however, such consultations cannot substitute for validation of the analytical approach adopted. Only a critical examination of the assessments generated by the chosen approach can provide a firm indication of how well it works in practice to identify potential quality of care problems.

5. In our view, the data cited from the New York State Department of Health do not support the utility of HCFA's 1986 hospital mortality analyses. First, they do not refer to the outlier hospitals identified by HCFA,

but instead, to hospitals designated as outliers in the Health Department's own regression analysis, which differed substantially from HCFA's approach. The New York analysis used New York State, rather than national data, combined the acute myocardial infarction DRGs, eliminated specialty hospitals, dropped the average length of stay variable, and added a separate case-mix variable.

Second, only 5 of the 10 high mortality outliers identified in the New York analysis coincided with high outlier hospitals in HCFA's analysis. Moreover, the "non-targeted cases," which HCFA compares to these outlier hospitals, were identified in a completely separate analysis by the New York Department of Health which sought to identify characteristics of individual patients (as opposed to hospitals) that might serve to indicate quality of care problems. Because these "non-targeted cases" were defined by their lack of any of the 11 patient characteristics being tested as indicators of poor quality care, they are unlikely to reflect an average level of quality problems. Likewise, the 2.6 percent of cases with confirmed problems from the outlier hospitals is not indicative of overall quality problems at those hospitals because they do not represent their full patient populations—only patients in DRGs for which that hospital had a mortality rate higher than the state average were sampled for this review.

6. While we did not ask PROs how useful they found the 1986 hospital mortality analyses overall, chapter 3 documents in considerable detail the outcome of the single instance where HCFA required virtually all the PROs to apply these analyses: that is, the evaluation of the outlier hospitals for possible inclusion in PRO quality of care objectives. The results of our survey are consistent with HCFA's view that the exercise was more useful for some PROs than for others. The 13 hospital outliers that PRO medical review subsequently confirmed as having quality problems, together with the six judged to have possible problems, were distributed among 10 states, leaving 41 states with no hospitals in either category. At most, two hospitals from the outlier lists were confirmed to have quality problems in any one state, although one state reported four hospitals likely to have quality problems.

7. We are uncertain whether HHS is referring here to the validation of analytical approaches for examining outcomes or validation of the data being analyzed. With respect to validating approaches to outcomes analyses, our view, stated on p. 17, is that there is no one "method of choice." Rather, the appropriate source and method will vary, depending on the relevant quality issue. Medical records are likely to be the

best source for many purposes, particularly when the process of care is at issue, but there are a number of areas where they contain relatively little information.

For example, medical records typically describe the procedures a patient has undergone, but convey little insight on how well those procedures were performed. Therefore, exclusive reliance on medical record review may provide less than optimal validation of approaches for assessing patient outcomes.

Checking the accuracy of data elements recorded in administrative data files is another matter. Here the issue is basically the correspondence of the computerized data set to information found in medical records, such as diagnoses and procedures. For this purpose, medical record reviews of representative samples of cases are the only logical source.

8. We have revised our discussion in chapter 2 to clarify the problem we see with regression analyses that have very limited explanatory power. We agree with the Department that these regression equations are intended to account only for that portion of variation in mortality rates that reflects differences in patient severity. We also agree that these equations inevitably represent an imperfect adjustment for patient severity, so that the unexplained variance in fact reflects, in addition to random variation, both differences in quality of care and differences in patient severity that have not been accounted for in the model.

We have no way of knowing how much of the unexplained variance represents quality differences and how much represents imperfect severity adjustment. However, as the total proportion of variance accounted for by these models gets very small—the nine DRG specific models range between 0.3 and 6.8 percent of variance explained—the rationale for using these equations to compute expected mortality becomes increasingly weak.

If on the one hand, the model in fact adjusts well for severity, it necessarily follows that severity has little to do with variations in outcomes among hospitals for the condition or procedure being analyzed—in which case there seems little point in making a severity adjustment at all. A simple comparison of observed mortality rates would be more direct and serve equally well.

If on the other hand, patient severity does play a major role in determining variations in hospital outcomes, adjustments based on a regression

equation with a low R-squared would only encompass a small part of the relevant differences in patient condition. In other words, the adjustment is either unnecessary or inadequate, though without independent validating evidence for the approach, we cannot say which.

9. Chapter 7 specifically addresses the question of what would be required to use administrative data to assess the impact of PPS on quality of care. There we lay out in detail the problems involved in obtaining comparable baseline data from the pre-PPS period, as well as the difficulty of isolating the effects of PPS from other factors that influenced Medicare patient outcomes over the years in which PPS gradually took effect. We also note the ways in which studies sponsored by HCFA to explicitly examine PPS effects have dealt with these issues.

However, in this comment the Department seems to ignore the HCFA studies designed to focus on PPS impacts. Instead, it apparently refers to the HSQB monitoring systems described in chapter 2, and perhaps the SysteMetrics application of Disease Staging. Neither of these analyses addresses the particular analytical problems raised by suspect diagnostic data from the pre-PPS era or the protracted implementation of PPS. In fact, HSQB's limited analysis of pre-PPS outcomes specifically excluded any adjustment for diagnoses because of the absence of adequate data. In this comment, the Department disregards these concerns and implicitly asserts that PPS effects, as well as related issues such as the extent of systematic change in diagnostic coding associated with the transition to PPS (i.e., "coding creep"), can be assessed without data from the period which preceded the implementation of prospective payment. For all the reasons described at length in chapter 7, we strongly disagree.

# Bibliography

## Approaches Examined—HCFA and HCFA-Sponsored


Conklin, Jonathan E., and Robert L. Houchens. "PPS Impact on Mortality Rates: Adjustments for Case Severity." Final Report. Santa Barbara, Calif.: SysteMetrics, Inc., Oct. 6, 1987.

DesHarnais, Susan, et al. "The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance." Typescript. Commission on Professional and Hospital Activities. Ann Arbor: University of Michigan, Sept. 28, 1987.

Gaumer, Gary. "Medicare Patient Outcomes and Hospital Organizational Mission." For-Profit Enterprise in Health Care, Bradford H. Gray, ed. Washington, D.C.: National Academy Press, 1986, pp. 354-74.

Health Care Financing Administration. Medicare Hospital Mortality Information: 1986, HCFA Pub. No. 01-002, 7 vols. Washington, D.C.: U.S. Government Printing Office, 1987.

———. "Medicare Program; Selected Performance Information on Hospitals Providing Care to Medicare Beneficiaries," 52 Fed. Reg. 158 (Aug. 17, 1987), 30741-45.

———. Memorandum releasing the list of hospital mortality outliers. Washington, D.C.: Mar 10, 1986.

Krakauer, Henry. "Outcomes of In-hospital Care in 1983-1985: The Medicare Experience." Typescript, Office of Medical Review, Health Standards and Quality Bureau, Health Care Financing Administration. Baltimore: no date.


## Approaches Examined—Non-HCFA


Blumberg, Mark S. "Maryland Mortality for Non-elective Surgery: A Prototype RAMO System." Typescript, May 6, 1987.

———. "Risk-adjusted Post-operative Mortality: Maryland Hospital Discharges April 1984-March 1985." Data for a presentation to the Maryland Health Services Cost Review Commission, Sept. 18, 1986.

Roos, Leslie L., Jr., et al. "Using Computers to Identify Complications After Surgery." American Journal of Public Health, 75:11 (Nov. 1985), 1288-95.

———. "Centralization, Certification, and Monitoring: Readmissions and Complications After Surgery." Medical Care, 24:11 (Nov. 1986), 1044-66.

Wennberg, John E., et al. "Use of Claims Data Systems to Evaluate Health Care Outcomes: Mortality and Reoperation Following Prostatectomy." Journal of the American Medical Association, 257:7 (Feb. 20, 1987), 933-36.

## Approaches Not Meeting Our Criteria

Administrator of Veterans' Affairs. "A Report on the Quality of Surgical Care in the Veterans Administration: The Phase I Report to the Congress of the United States under the Provisions of Public Law 99-166, Section 204." Typescript, Apr. 1, 1987.

Anderson, Gerald F., and Earl P. Steinberg. "Hospital Readmissions in the Medicare Population." New England Journal of Medicine, 311:21 (Nov. 22, 1984), 1349-53.

———. "Predicting Hospital Readmissions in the Medicare Population." Inquiry, 22 (Fall 1985), 251-58.

Bloom, Bernard S., and Osler L. Peterson. "End Results, Cost and Productivity of Coronary-Care Units." New England Journal of Medicine, 288:2 (Jan. 11, 1973), 72-78.

Bunker, J. P., et al., eds. The National Halothane Study. Bethesda, Md.: The National Institute of General Medical Sciences, 1969.

Dubois, Robert W., et al. "Adjusted Hospital Death Rates: A Potential Screen for Quality of Medical Care." American Journal of Public Health, 77:9 (Sept. 1987), 1162-66. (This study meets our four criteria listed in chapter 1, but was excluded because it employed an approach quite similar to the 1986 HCFA hospital mortality analyses.)

———. "Hospital Inpatient Mortality: Is It a Predictor of Quality?" New England Journal of Medicine, 317:26 (Dec 24, 1987), 1674-80.

Flood, Ann Barry, et al. "The Relationship Between Intensity and Duration of Medical Services and Outcomes for Hospitalized Patients." Medical Care, 17:11 (Nov. 1979), 1088-1102.

Francis, Anita M., et al. "Care of Patients with Colorectal Cancer: A Comparison of a Health Maintenance Organization and Fee-for-Service Practices." Medical Care, 22:5 (May 1984), 418-29.

Garber, Alan M., et al., "Case Mix, Costs, and Outcomes: Differences between Faculty and Community Services in a University Hospital." New England Journal of Medicine, 310:19 (May 10, 1984), 1231-37.

Gilpin, Elizabeth, et al. "Risk Prediction after Myocardial Infarction: Comparison of Three Multivariate Methodologies." Cardiology, 70 (1983), 73-84.

Hebel, J. Richard, et al. 'Assessment of Hospital Performance by Use of Death Rates: A Recent Case History." Journal of the American Medical Association, 248:23 (Dec. 17, 1982), 3131-36.

Kennedy, J. Ward, et al. "Multivariate Discriminant Analysis of the Clinical and Angiographic Predictors of Operative Mortality from the Collaborative Study in Coronary Artery Surgery (CASS)." Journal of Thoracic and Cardiovascular Surgery, 80:6 (Dec. 1980), 876-87.

Knaus, William A., et al. "Evaluating Outcomes from Intensive Care: A Preliminary Multihospital Comparison." Critical Care Medicine, 10:8 (Aug. 1982), 491-96.

Lamont, Campbell T., et al. "The Outcome of Hospitalization for Acute Illness in the Elderly." Journal of the American Geriatrics Society, 31:5 (May 1983), 282-88.

Lokkeberg, A. Russell, and Richard M. Grimes. "Assessing the Influence of Non-treatment Variables in a Study of Outcome from Severe Head Injuries." Journal of Neurosurgery, 61 (1984), 254-62.

Lubitz, James, et al. "Outcomes of Surgery Among the Medicare Aged: Mortality After Surgery." Health Care Financing Review, 6:4 (Summer 1985), 103-15.

Merrick, Nancy J., et al. "Use of Carotid Endarterectomy in Five California Veterans Administration Medical Centers." Journal of the American Medical Association, 256:18 (Nov. 14, 1986), 2531-35.

Miller, D. Craig, et al. "Discriminant Analysis of the Changing Risks of Coronary Artery Operations: 1971-1979." Journal of Thoracic and Cardiovascular Surgery, 85:2 (Feb. 1983), 197-213.

Moses, Lincoln E., and Frederick Mosteller. "Institutional Differences in Postoperative Death Rates: Commentary on Some of the Findings of the National Halothane Study." Journal of the American Medical Association, 203:7 (Feb. 12, 1968), 150-52.

Mossey, Jana M., and Leslie L. Roos, Jr. "Using Insurance Claims to Measure Health Status: The Illness Scale." Journal of Chronic Disease, 40:Supplement (1987), 41S-50S.

Moylan, Joseph A., et al. "Evaluation of the Quality of Hospital Care for Major Trauma." Journal of Trauma, 16:7 (July 1976), 517-23.

Pardaens, J., et al. "Multivariate Survival Analysis for the Assessment of Prognostic Factors and Risk Categories After Recovery From Acute Myocardial Infarction: The Belgian Situation." American Journal of Epidemiology, 122:5 (1985), 805-19.

Pollack, Murray M., et al. "Accurate Prediction of the Outcome of Pediatric Intensive Care." New England Journal of Medicine, 316:3 (Jan. 15, 1987), 134-39.

Riley, Gerald, and James Lubitz. "Outcomes of Surgery in the Medicare Aged Population: Rehospitalization After Surgery." Health Care Financing Review, 8:1 (Fall 1986), 103-15.

Roemer, Milton I., et al. "A Proposed Hospital Quality Index: Hospital Death Rates Adjusted for Case Severity." Health Services Research (Summer 1968), 96-118.

Seymour, David Gwyn, and Robert Pringle. "A New Method of Auditing Surgical Mortality Rates: Application to a Group of Elderly General Surgical Patients." British Medical Journal, 284 (May 22, 1982), 1539-42.

Shortell, Stephen M., and James P. LoGerfo. "Hospital Medical Staff Organization and Quality of Care: Results for Myocardial Infarction and Appendectomy." Medical Care, 19:10 (Oct. 1981), 1041-53.

Stanford Center for Health Care Research. "Study of Institutional Differences in Postoperative Mortality." Report prepared for the National

Center for Health Services Research, National Technical Information Service, no. PB 250 940. Dec. 15, 1974.

———. "Comparison of Hospitals with Regard to Outcomes of Surgery." Health Services Research (Summer 1976), 112-27.

## Hospital Volume and Outcome Literature

Flood, Ann Barry, et al. "Does Practice Make Perfect? Part I: The Relation Between Hospital Volume and Outcomes for Selected Diagnostic Categories"; "Part II: The Relation Between Volume and Outcomes and Other Hospital Characteristics." Medical Care, 22:2 (Feb. 1984), 98-125.

Goldberg, Jack, et al. "Mortality from Traumatic Injuries: A Case-Control Study Using Data from the National Hospital Discharge Survey." Medical Care, 21:7 (July 1983), 692-704.

Hughes, Robert G., et al. "Effects of Surgeon Volume and Hospital Volume on Quality of Care in Hospitals." Medical Care, 25:6 (June 1987), 489-503.

Kelly, Joyce V., and Fred. J. Hellinger. "Physician and Hospital Factors Associated with Mortality of Surgical Patients." Medical Care, 24:9 (Sept. 1986), 785-800.

Luft, Harold S. "The Relation Between Surgical Volume and Mortality: An Exploration of Causal Factors and Alternative Models." Medical Care, 18:9 (Sept. 1980), 940-59.

———, et al. "Should Operations Be Regionalized? The Empirical Relation between Surgical Volume and Mortality." New England Journal of Medicine, 301:25 (Dec. 20, 1979), 1364-69.

Riley, Gerald, and James Lubitz. "Outcomes of Surgery Among the Medicare Aged: Surgical Volume and Mortality." Health Care Financing Review, 7:1 (Fall 1985), 37-47.

Rosenfeld, Kenneth, et al. "Changes in Patient Characteristics and Surgical Outcomes for Coronary Artery Bypass Surgery 1972-82." American Journal of Public Health, 77:4 (Apr. 1987), 498-500.

Sloan, Frank A., et al. "In-hospital Mortality of Surgical Patients: Is There an Empiric Basis for Standard Setting?" Surgery, 99:4 (Apr. 1986), 446-53.

## Related Literature

Avery, Allyson Davies, et al., Quality of Medical Care Assessment Using Outcome Measures: Eight Disease Specific Applications. Santa Monica, Calif.: Rand Corp., Aug. 1976.

Blumberg, Mark S. "Comments on HCFA Hospital Death Rate Statistical Outliers." Health Services Research, 21:6 (Feb. 1987), 715-39.

———. "Risk Adjusting Health Care Outcomes: A Methodologic Review." Medical Care Review, 43:2 (Fall 1986), 351-93.

Bransome, E. D., Jr. "Assessment of the Impact of Medicare PPS on the Care of Persons with Diabetes Mellitus." Diabetes Care, 9:4 (July-Aug. 1986), 415-19.

Brook, Robert H., et al. Quality of Medical Care Assessment Using Outcomes Measures: An Overview of the Method. Santa Monica, Calif.: Rand Corp., Aug. 1976.

———, and Kathleen N. Lohr. "Efficacy, Effectiveness, Variations, and Quality." Medical Care, 23:5 (May 1985), 710-22.

Brown, Byron William, Jr. "Statistical Problems in Comparing Outcomes of Low Incidence." Health Care Delivery in Anesthesia, Robert A. Hirsh et al., eds. Philadelphia: George F. Stickley Co., 1980.

Bunker, John P., and Jinnet Fowles. "Medical Audit by Claims Data?" American Journal of Public Health, 75:11 (Nov. 1985), 1261-62.

Calore, Kathleen A., and Lisa Iezzoni. "Disease Staging and PMCs: Can They Improve DRGs?" Medical Care, 25:8 (Aug. 1987), 724-35.

Cislowski, Joseph A. "The Peer Review Organization (PRO) Program." Washington, D.C.: Congressional Research Service, Mar. 25, 1987.

Coffey, Rosanna M., and Marsha G. Goldfarb. "DRGs and Disease Staging for Reimbursing Medicare Patients." Medical Care, 24:9 (Sept. 1986), 814-29.

Conklin, Jonathan E. DRG Refinement: A Study of Alternative Groupings within Six Sets of Adjacent DRGs, Final Report. Santa Barbara, Calif.: SysteMetrics, Inc., Mar. 1985.

———, and Robert L. Houchens. DRG Refinement: Using Measures of Disease Severity, Final Report. Santa Barbara, Calif.: SysteMetrics, Inc., Aug. 1987.

———, et al. "Disease Staging: Implications for Hospital Reimbursement and Management." Health Care Financing Review, Annual Supplement (Nov. 1984), 13-22.

———, et al. DRG Refinement: A Feasibility Assessment Using Stage of Disease, Age, and Unrelated Comorbidity, Final Report. Santa Barbara, Calif.: SysteMetrics, Inc., Nov. 1984.

Gertman, Paul, and Steven Lowenstein. "A Research Paradigm for Severity of Illness: Issues for the Diagnosis-related Group System." Health Care Financing Review, Annual Supplement (Nov. 1984), 80-82.

Gonnella, Joseph S., et al. "Staging of Disease: A Case-Mix Measurement." Journal of the American Medical Association, 251:5 (Feb. 3, 1984), 637-44.

Goss, Mary E. W., and Joseph I. Reed. "Evaluating the Quality of Hospital Care Through Severity Adjusted Death Rates: Some Pitfalls." Medical Care, 12:3 (Mar. 1974), 202-13.

Hirsh, Robert A., et al., eds. Health Care Delivery in Anesthesia. Philadelphia: George F. Stickley Co., 1980.

Horn, Susan D., and Dale N. Schumacher. "Comparing Classification Methods: Measurement of Variations in Charges, LOS and Mortality." Medical Care, 20:5 (May 1982), 489-500.

Hornbrook, Mark C. "Hospital Case-mix: Its definition, Measurement, and Use. Part I: The Conceptual Framework"; "Part II: Review of Alternative Measures." Medical Care Review, 39 (1982), 1-43; 73-123.

Jencks, Stephen F., and Allen Dobson. "Refining Case-mix Adjustment: The Research Evidence." New England Journal of Medicine, 317:11 (Sept. 10, 1987), 679-86.

Long, Michael J., et al. "The Effect of PPS on Hospital Product and Productivity." Medical Care, 25:6 (June 1987), 528-38.

Luft, Harold S., and Sandra S. Hunt. "Evaluating Individual Hospital Quality Through Outcome Statistics." Journal of the American Medical Association, 255:20 (May 23/30, 1986), 2780-84.

McAuliffe, William E. "Studies of Process-Outcome Correlations in Medical Care Evaluations: A Critique." Medical Care, 16:11 (Nov. 1978), 907-30.

Moses, Lincoln E. "The Evaluation of Hospital Death Rates." Journal of the American Medical Association, 255:20 (May 23/30, 1986), 2801.

Rosen, Harry M., and Barbara A. Green. "The HCFA Excess Mortality Lists: A Methodological Critique." Hospital and Health Services Administration (Feb. 1987), 119-27.

Schroeder, Steven A. "Outcome Assessment 70 Years Later: Are We Ready?" New England Journal of Medicine, 316:3 (Jan. 15, 1987), 160-62.

Schumacher, Dale N., et al. "Severity of Illness Index and the Adverse Patient Occurrence Index." Medical Care, 25:8 (Aug. 1987), 695-703.

Scott, W. Richard, and Ann Barry Flood. "Costs and Quality of Hospital Care: A Review of the Literature." Medical Care Review, 41:4 (Summer 1984), 213-61.

Thomas, J. William, et al. An Evaluation of Alternative Severity of Illness Measures for Use by University Hospitals. Department of Health Services Management and Policy, School of Public Health. Ann Arbor: University of Michigan, Dec. 29, 1986.

Young, Wanda W. "Incorporating Severity of Illness and Co-morbidity in Case-mix Measurement." Health Care Financing Review, Annual Supplement (Nov. 1984), 23-31.

——, et al. "The Measurement of Hospital Case Mix." Medical Care, 20:5 (May 1982), 501-12.

## Data Quality

California Medical Review, Inc. Premature Discharge Study, Final Report. Typescript. San Francisco: Dec. 10, 1986.

Carter, Grace M., and Paul B. Ginsburg. The Medicare Case Mix Index Increase: Medical Practice Changes, Aging, and DRG Creep, Santa Monica, Calif.: Rand Corp., June 1985.

Corn, R. F. "Quality Control of Hospital Discharge Data." Medical Care, 18:4 (Apr. 1980), 416-26.

Cretin, Shan, and Linda G. Worthman. Alternative Systems for Case Mix Classification in Health Care Financing. Santa Monica, Calif.: Rand Corp., Sept. 1986.

Demlo, Linda K., and Paul M. Campbell. "Improving Hospital Discharge Data: Lessons From the National Hospital Discharge Survey." Medical Care, 19:1 (Oct. 1981), 1030-40.

———, et al. "Reliability of Information Abstracted from Patients' Medical Records." Medical Care, 16:12 (Dec. 1978), 995-1005.

Doremus, Harvey D., and Elana M. Michenzi. "Data Quality: An Illustration of Its Potential Impact Upon a Diagnosis-Related Group's Case Mix Index and Reimbursement." Medical Care, 21:10 (Oct. 1983), 1001-11.

Health Care Financing Administration. Medicare Statistical Files Manual. Baltimore: Apr. 1987.

Institute of Medicine. Reliability of Medicare Hospital Discharge Records, Final Report. Washington, D.C.: National Academy of Sciences, Nov. 1977.

Johnson, Allan N., and Gary L. Appel. "DRGs and Hospital Case Records: Implications for Medicare Case Mix Accuracy." Inquiry, 21 (Summer 1984), 128-34.

Kosecoff, Jacqueline, et al. "The Appropriateness of Using a Medical Procedure: Is Information in the Medical Record Valid?" Medical Care, 25:3 (Mar. 1987), 196-201.

Lloyd, Susan S., and J. Peter Rissing. "Physician and Coding Errors in Patient Records." Journal of the American Medical Association, 254:10 (Sept. 13, 1985), 1330-36.

Massanari, R. Michael, et al. "Reliability of Reporting Nosocomial Infections in the Discharge Abstract and Implications for Receipt of Revenues

under Prospective Payment." American Journal of Public Health, 77:5 (May 1987), 561-64.

Mullin, Robert L. "Diagnosis-Related Groups and Severity: ICD-9-CM, the Real Problem." Journal of the American Medical Association, 254:9 (Sept. 6, 1985), 1208-10.

Simborg, Donald W. "DRG Creep: A New Hospital-Acquired Disease." New England Journal of Medicine, 304:26 (June 25, 1981), 1602-04.

Worthman, Linda G., and Shan Cretin. Review of the Literature on Diagnosis Related Groups. Santa Monica, Calif.: Rand Corp., Oct. 1986.

*U.S. G.P.O. 1988-201-749:80158

Requests for copies of GAO reports should be sent to:

U.S. General Accounting Office
Post Office Box 6015
Gaithersburg, Maryland 20877

Telephone 202-275-6241

The first five copies of each report are free. Additional copies are $2.00 each.

There is a 25% discount on orders for 100 or more copies mailed to a single address.

Orders must be prepaid by cash or by check or money order made out to the Superintendent of Documents.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100