Symposium

Case-mix measurement and assessing quality of hospital care

Introduction

Patients, insurers, and health care systems are increasingly counseled to consider both quality and cost in purchasing health care services, particularly hospital care, to "buy right." Although far more cost information is available today than 25 years ago, purchasers are no better off when trying to select hospitals of high quality. Nevertheless, quality issues are increasingly in the public mind because of fear that efforts to cut costs will reduce quality as well.

The heart of the problem is that research on quality of care has focused on the way care is rendered, the process of care, rather than on outcomes of care. Such process measures are critical to quality assurance activities meant to correct problems, but most consumers and payers are more interested in using quality assessment to select good hospitals and avoid the bad. For such comparison shopping, information on outcomes is more objective and more relevant. Outcome measures such as death and rehospitalization may also be less expensive than process measures because they are available from administrative data. Unfortunately, comparing outcomes intelligently requires that we be able to correct for differences in hospital case mix that might account for differences in outcomes and that we be confident that differences in outcomes that remain after correcting for case mix truly reflect differences in quality of care.

One example of using outcomes as a screen for quality of care is the release of hospital-specific Medicare mortality rates, which will occur about the time this supplement appears. Another is the recent decision of the Pennsylvania Health Care Cost Containment Council to require all hospitals to collect case-mix data on all discharges in order to determine quality of care. These public strategies are controversial, principally because of uncertainty as to whether the state of the art in case-mix adjustment really allows meaningful comparisons of outcomes such as death rates. These strategies imply that case-mix measurement is, at least potentially, a key to measuring quality of care. However, case-mix measurement of this kind is at the cutting edge of health services research and is quite controversial.

To provide readers with a sense of the diversity of scholarly views on the application of case-mix measurement to measuring quality of care, I asked six investigators to respond to questions related to three areas. In putting together the contributors' comments, I have preserved their arguments as they presented them, even at the expense of occasional repetition, in order to indicate the degree of consensus and disagreement that actually exists in this field. The symposium that follows may also sometimes appear disjointed because responses have been put together as if they resulted from a face-to-face encounter that never actually took place. This is, in fact, a kind of symposium by mail. The contributors to this symposium are Robert H. Brook, Lisa I. Iezzoni, Stephen F. Jencks, William A. Knaus, Henry Krakauer, Kathleen N. Lohr, and Mark A. Moskowitz.

Reprint requests: Stephen F. Jencks, Office of Research and Demonstrations, Health Care Financing Administration, Room 2-B-14 Oak Meadows Building, 6325 Security Boulevard, Baltimore, Maryland 21207.

Relation of case mix and quality measures

For what kinds of quality measurement is case-mix adjustment useful at the hospital level: Technical quality of care? Access to care? Patient satisfaction? Measures of outcomes that can be determined from administrative data? Is it true that, if we want to collect enough information about quality of care to allow purchasers to buy right, we will have to rely on case-mix measurement and comparison of outcomes rather than process-of-care measures? If not, what are the alternatives?

Lohr: The underlying premise appears to be that assessing quality of care and, when necessary, intervening to improve care, will increasingly be oriented to patient outcomes, but that such outcomes will have to be measured nonintrusively, that is, via insurance claims or other administrative data that do not require special data collection. Corollary assumptions are that:

- Process-oriented quality assessment is too time consuming and costly.
- Implementation of such approaches in the past has been, at best, disappointing.
- Processes and outcomes are the chief dimensions by which quality might be measured.

All these premises deserve examination, although they cannot be either confirmed or disproved here. Further, quality measurement or assessment is not synonymous with quality assurance, which encompasses activities to improve care when problems are detected. Moreover, quality measures are not synonymous with indicators, which are screens or early warning devices that may signal the need for further quality measurement.

Traditionally, process and outcome have been considered the only two dimensions for the direct assessment of quality of care, although structural characteristics of hospitals are often used as proxies for process and outcome. Within this framework, some take the view that patient outcomes are better,

being closer to the presumed end product of health care. That is a defensible view in classic quality assessment terms and has guided the field for more than two decades (Donabedian, 1966, 1980, 1982, and 1985).

Let me digress for a moment into quality assurance. In taking the classic view, one is also led to a method of quality assurance that depends on detecting problems in the end result of health care rather than trying to anticipate and correct problems at an earlier point in the delivery of health care. Designing delivery systems that incorporate quality controls throughout may be a more efficient approach to improving quality of care than initiating corrective action only when deficient outcomes are detected. Hence, in a health care world as complex as the one the Nation (or even Medicare) faces, placing central emphasis on outcomes as the *sine qua non* of quality assessment may not be the best tactic for quality assurance. We need, perhaps, some additional, or orthogonal, ways of thinking that incorporate concepts of clinical epidemiology, efficacy versus effectiveness, and medical technology assessment (Brook and Lohr, 1985).

Nonintrusive outcomes may be good screening tools for quality problems. (The jury is still out.) However, in this context they suffer from the same drawbacks as directly measured patient outcomes and have the added disadvantage of being only screens, not actual measures, of quality of care. No amount of case-mix or severity adjustment will compensate for those limitations, even though such adjustments (sometimes called patient classification systems) are imperative to allow the valid use of such indicators for quality assessment, consumer information, or quality assurance programs.

Iezzoni and Moskowitz: In assessing the utility of patient classification or case-mix measurement systems for evaluating quality of care, we must first emphasize the diversity of the systems and the implications that arise from the differences. Existing classification schemes can be arrayed along two dimensions.

The first dimension pertains to the data requirements, with systems sorted into those requiring only the Uniform Hospital Discharge Data Set (UHDDS), e.g., Computerized Disease Staging and Patient Management Categories, and those requiring primary data collection, (e.g., Acute Physiology and Chronic Health Evaluation II (APACHE), MedisGroups, and Computerized Severity Index). The latter systems provide more clinically meaningful comparisons of patient health status and risks for negative outcomes, but the nature of some of the data elements raises some questions about the comparability of case-mix measurements across hospital types. For example, teaching hospitals may be more likely than nonteaching hospitals to document abnormal physical examination findings. Likewise, systems such as MedisGroups, which incorporate findings from cardiac catheterization, angiography, endoscopy, and other expensive procedures, may measure differences in the frequency with which hospitals use such procedures in addition to, and perhaps sometimes instead of, actual differences in patient condition.

The second dimension involves timing of the patient assessment. At one end of the scale lie systems that can be applied only after discharge and encompass the entire hospital stay, such as Horn's uncomputerized Severity of Illness Index and systems using UHDDS (Horn, Horn, and Sharkey, 1984). These systems may permit identification of patients at risk for a poor outcome, but they do not distinguish risk factors present at admission from those that arise during hospitalization and may result from poor care. At the other end of the scale are systems that can be applied at several points during the hospital stay and are sensitive enough to identify day-to-day changes in clinical status (e.g., APACHE II, Computerized Severity Index, and MedisGroups). We will focus on this latter type of system.

Such case-classification systems are useful in stratifying patients' risks for negative outcomes. They were designed to do so, and the evidence from at least one system indicates success. Increasing APACHE II score is positively correlated with mortality rates among intensive care unit patients. However, their utility for quality assessment depends on how they are used. For example, looking at the worst APACHE II score over the entire hospital stay would reveal little about quality of care because the user would not be able to distinguish risk factors present on admission from problems that develop during, and perhaps as a consequence of, the provision of care.

The best way to identify quality shortfalls is to apply the classification system several times so as to measure change. For example, MedisGroups stipulates that patients who remain hospitalized a certain length of time receive a second review. An optimal scheme might involve looking at patient status at admission, at the worst point of the hospital stay, and at discharge. Multiple measures permit assessment of a range of quality concerns. For example:

- If clinical status is worse following performance of a procedure more often at one hospital than at another, a problem may exist with the technical execution of the procedure at that hospital.
- Patient satisfaction may be lower at a hospital where severity worsens following admission than at a hospital where clinical status improves steadily from its nadir on admission.
- Access to care may be poor at a hospital where patients are systematically more gravely ill on admission. Conversely, inappropriate admissions are a potential problem at hospitals where patients appear quite well on admission.
- Premature discharge and patient dissatisfaction may occur at hospitals where clinical status is poor on discharge.
- Clinical status at discharge (e.g., death) may be viewed as an outcome of the hospital stay.

The multiple-measures approach would produce a wealth of useful information, but it would serve only as a screen to suggest areas of potential quality

shortfalls. To ascertain whether care is poor, we still must look at the process of care. If one hospital provides extensive postdischarge planning to ensure that adequate health services and supports are available for patients in poor condition but another neglects postdischarge arrangements, then quality of care differs even though condition at discharge is the same. If one hospital admits patients referred for extensive diagnostic workups unavailable at other local facilities, while another inappropriately admits patients who are not sick, identical case-mix findings again have different implications.

From a consumer's viewpoint, information from such case-mix adjustments is also inadequate to support a buy-right strategy because of the limited perspective. Although we have focused on a single admission, consumers may be more interested in patient experience over a complete episode of illness. For example, suppose two clinically identical patients with congestive heart failure seek care at different hospitals over a 2-month period. Both patients have a waxing and waning course. At one hospital, the patient is admitted, quickly discharged with slight improvement, and readmitted; this happens five times, and he dies during the last hospital stay. At the second hospital, the patient is admitted and stays the entire 2 months before dying. The case-mix-adjusted death rate per admission is 20 percent at the first hospital and 100 percent at the second, but these differences are clearly deceptive; they would be even more deceptive if the latter hospital transferred the patient to a skilled nursing facility just before death.

As another example, suppose that the cardiology staffs at two hospitals have different approaches to the use of coronary artery bypass surgery in their angina pectoris patients. The first hospital tries to postpone surgery as long as possible using multiple medical regimens. Patients may be admitted several times, with alterations in the medical approach, before they go to surgery. Because of this, fewer patients have surgery. At the second hospital, patients are referred to surgery quickly, with minimal medical intervention. Thus, more patients have surgery. Both hospitals have an equal case-mix-adjusted mortality rate for the surgical admission, but clearly the implications for overall mortality rates are quite different. An information system must take account of these differences in practice in order to aid consumers in buying right.

Finally, we would point out that quality assessment and quality assurance are both processes with a single (albeit complicated) ultimate goal, that of improving quality of care. The information obtained for one process is certainly useful for the other. In a way, they are flip sides of the same coin, and we feel that excessive emphasis on their differences is somewhat misleading.

Lohr: I think your emphasis on measurements over time is important, but it is more difficult to operationalize than one might imagine, as we are finding from the validation work on our study of nonintrusive outcome measures. This is one area that calls for additional research.

Krakauer: The quality of medical care is judged by two criteria: effectiveness, which can be assessed objectively, and satisfaction, which is subjective. Four activities are required for the evaluation and optimization of the effectiveness of care:

- Monitoring the outcomes of care to assess the current status and time trends in the effectiveness of care.
- Analyzing variations in the utilization and outcomes of care among geographic areas, subpopulations, and providers of care to identify deficiencies in care.
- Assessing the effectiveness of identified interventions by measuring their impact on outcomes to ascertain the causes of the variations attributable to medical practices.
- Providing the resulting information to the medical community and providing incentives for its effective use.

Present work with Medicare data bases to support these four activities (Krakauer, 1987) has produced a body of information that bears directly on the importance of case-mix or risk adjustment in assessing the care delivered in the hospital and its aftermath.

This analysis focuses on death as an outcome. We analyze survival from a patient's first admission in a calendar year until 30 days after the end of the calendar year. Using this interval of observation is preferable to analyzing deaths that occur in the hospital because it removes bias that might result from changes in length of stay or differences in length of stay among hospitals (U.S. Department of Health and Human Services, to be published). Examining only one admission for each patient eliminates some of the potential problems that Moskowitz and Iezzoni identify.

Trends in hospital-associated death rates have been of great concern during the implementation of the Medicare prospective payment system. From 1984 to 1985, the hospital-associated death rate rose 9 percent. During that period, conditions such as malignancy, stroke, sepsis, pulmonary disease, renal disease, and heart attacks and failure, which are associated with greater than average likelihood of death, accounted for an increasing fraction of Medicare admissions. This increase accounts for about one-half of the increase in hospital-associated death rates.

These findings may reflect data problems, the most worrisome of which is the virtual absence of any systematic effort to ascertain the validity of the information entered on the bills and transmitted to the Health Care Financing Administration (HCFA). Peer review organizations do undertake to validate diagnosis-related group (DRG) assignments, but their sample is biased toward DRG's that are subject to coding uncertainty and conditions in which abuse of coding is likely. Patricia Booth of HCFA reports that validated DRG assignments currently differ from the original assignments in 10-13 percent of cases (Booth, 1987).

The data show changing patterns of coding over time, and additional diagnoses were coded more frequently in 1985 than in 1984. This was true for comorbid conditions associated with higher death rates, such as chronic renal and lung disease, malignancy, and diabetes, but it was also true for hypertension, which is actually associated with lower death rates once its effects on heart, brain, and kidneys are taken into account. In addition, the frequency of recorded comorbidities increased not only for admissions whose mortality rates rose, such as coronary artery bypass, but also for admissions whose mortality rates fell, such as major joint procedures. At least part of this increase in recorded comorbidities is almost certainly the result of greater diligence in recording diagnostic codes, because the increase is not clearly and consistently associated with a higher incidence of adverse outcomes. Consequently, assessing secular trends in the complexity of the illnesses of patients by counting additional diagnoses on the bills is hazardous.

There is less reason for concern about the principal diagnoses and procedures, although these too are subject to some manipulation, because the increases in the proportion of patients admitted for high-risk conditions are clearly accompanied by increases in postadmission mortality rates. In this instance, adjustment for case mix in the assessment of quality of care provides an important corrective to the initial impression of significant deterioration from 1984 to 1985. The stability of age-adjusted mortality rates in the Medicare population as a whole, independent of hospitalization, also suggests that there has not been a deterioration in quality of care (discussed later).

Knaus: First, we must define what is meant by case-mix adjustment. We believe that, if we ever want to be able to tell which components of the medical care process are important in influencing patient outcome, we must be able to describe the probability of an outcome objectively and reliably using information available at the time the decision to treat is made. Conceptually and scientifically, we do not think there is or ever will be any short cut to avoid this type of data collection. As more hospitals develop automated laboratory systems linked to hospital information networks, collecting data on the patient's condition at admission will become much less expensive and more reliable.

We also contend that the information used to adjust for patient risk must depend on patients, not treatments. Remember, we will be using the results of our predictions to judge the appropriateness of these treatment or process decisions. Therefore, our definition of case-mix adjustment is patient-specific information on type of disease, severity of disease, and physiologic reserve (age and chronic health), all determined at or shortly after the patient presents for the medical procedure or treatment we are attempting to evaluate.

With these patient characteristics in hand, we can prospectively stratify patients by their probability of having certain outcomes, such as death, postoperative complications, and infections. We can then assess which components of process are important in achieving an outcome. This precise description of patients will enable us eventually to separately identify the value of specific technical versus managerial components of the process of care. Such pretreatment risk stratification will also be useful to evaluate which patients have greatest access to care and how differing arrangements of care influence patient satisfaction. We acknowledge that this type of analysis is a significant departure from previous efforts at case-mix adjustment using discharge diagnoses. Unfortunately, one simply cannot use discharge diagnoses to judge the quality of the diagnostic or medical-surgical treatment process. Patient characteristics vary too much within diagnoses; adjustment is based on diagnostic information that may not have been available in making the treatment decision (Knaus et al., 1985a; Knaus and Zimmerman, 1985-86); and some additional diagnoses may not even have been present at admission.

This does not mean we will be able to eliminate disease designations as part of our case-mix adjustment. Proper disease designation is critical to patient description and risk stratification. Unfortunately, the classification system of the *9th Revision International Classification of Diseases*, ICD-9-CM (Public Health Service and Health Care Financing Administration, 1980), on which the DRG categories are based, is not very precise. In our own work on intensive care unit (ICU) patients, we have had to develop a more precise, less redundant taxonomy to specify disease. The observation from Dr. Krakauer's work that diagnostic codes explain less than one-half (4 of the 9 percent) of the increase in crude Medicare death rates highlights this shortcoming.

We cannot yet tell whether we will need disease- or diagnosis-specific criteria for case-mix adjustment. For some conditions, such as certain elective surgical procedures, minimal adjustment for chronic conditions may be sufficient. For many acute diseases, we may be able to use routine and uniform physiologic data whose relation to outcomes, although somewhat disease specific, may not vary greatly from one condition to another because, we believe, there is a uniform relationship between physiology and patient outcome for most acute life-threatening conditions (Wagner, Knaus, and Draper, 1986). We believe that this is why studies that have used APACHE II on non-ICU patients have found that it has good risk stratification properties (McClish et al., 1985; Coulton et al., 1985).

I would also like to comment on your suggestion that we use outcome as a measure of quality so purchasers can buy right. It brings up an important point. Who wants to be a patient in a hospital that has been identified as having a worse than average death rate? Hospitals so identified will be widely publicized and may suffer radical reductions in volume. Physicians whose sole hospital admitting privileges are in such hospitals may also suffer loss of

reputation and income, and they may have difficulty obtaining admitting privileges at other institutions.

One result will be tremendous pressure to keep the outcome results within acceptable limits by either improving quality in the hospital (which is fine) or trying to differentially influence the case mix (not fine). Unfortunately, the latter is likely to be much easier than the former. The pressures for admitting healthier patients unnecessarily and referring sicker than average patients to regional referral centers will be radically stronger than it is under the DRG system. Although there are some financial incentives for triage or transfer of critically ill high-cost patients under the current prospective payment system, the incentives are relatively small because of uncertainties over the cost of care for a patient and the fact that a filled bed brings in more revenue than an empty one.

Any movement toward direct measurement of quality through outcomes, however, could seriously threaten access to care for severely ill patients within any diagnostic category. Pressures for triage, recruitment of easy cases, and transfer of difficult ones might become intense. One implication, therefore, is that the case-mix system used to evaluate quality of care must adjust for risk of poor outcome within diagnostic classes with a precision close to that possible for clinicians. Adjustment must be made for patients at all levels of risk, from very high to very low, so that we can monitor small differences at either end of the spectrum. This could be done, as Lohr suggests, throughout the care process, rather than waiting for a final poor outcome to occur. For example, a trauma patient is admitted to a hospital and immediately goes to the operating room. The surgery is successful and lowers his previous high risk of death. Two days later, however, he develops a complication for which the clinical response is not appropriate, and he requires a second operation. Appropriate risk measurement would indicate where in the course of his care the poor process occurred. It could also serve as an early warning of developing problems to the clinicians treating the patient.

Brook: To assess the quality of care, one needs to decide whether to measure outcome of care, process of care, or patient satisfaction. If one chooses to measure patient satisfaction or outcome of care, then adjustment for differences in case mix must be made in order to make a definitive statement about the impact of a structural measure, such as the hospital, on quality of care. Usual case-mix adjustments for predicting outcomes adjust only for age, sex, and the kind of disease, and do not take account of severity of illness. At this time, we do not know how good these adjustments are for comparisons of hospitals on outcome measures. We do not know the sensitivity or specificity of adjusted outcomes in detecting poor hospital performance. This is true for any outcome measure one wishes to name, whether it be mortality, morbidity, or functional status. We also do not know how good age, sex, and condition adjustments are for contrasting hospitals that score differently on patient satisfaction measures. Finally, we do not even know what patient attributes (such as belief in efficacy of medical care) should be considered when comparing hospitals that score differently on patient satisfaction measures.

Lohr: I am not sure I agree that patient satisfaction measures need to be severity- or case-mix-adjusted. Given the state of the art of measuring patient satisfaction, however, it may be desirable and possible to take into account cultural and demographic characteristics, such as first language.

Jencks: I think that the panel's response can fairly be summed up by saying that case-mix measurement is potentially useful for quality measurement of many kinds at the hospital level but must be interpreted with great caution. Lohr points to the importance of distinguishing quality measurement from quality assurance and cautions that case-mix-adjusted outcomes are probably not adequate for quality assurance. Iezzoni and Moskowitz point out that multiple hospitalizations raise difficult problems in defining outcome, but Krakauer uses a method of analysis that allows for many (but not all) of these difficulties. Lohr and Brook suggest that case-mix adjustment has special meanings when we consider patient satisfaction as an outcome. Iezzoni, Moskowitz, and Knaus emphasize that case-mix adjustment must reflect the patient's diagnoses and severity of illness at admission; using worst severity during the illness, discharge severity, or discharge diagnoses confounds case-mix adjustment with the quality of care we seek to measure. Krakauer's data on changing diagnostic patterns over time raise concern in my mind as to whether coding practices are sufficiently stable to allow comparisons of patterns of secondary diagnoses, even if we are comfortable with using discharge diagnoses. Finally, Knaus points out the adverse effects that might result from using an inadequate method for inferring quality of care.

Measuring variation among providers

How well can severity and functional status measures distinguish among providers (e.g., hospitals, physicians, and capitated systems) as to their patients' risk for adverse outcomes? What do we know about how well these measures account for differences in outcomes across hospitals and physicians? How can we decide if severity and functional status measures account for enough of the variation in outcomes to allow us to conclude that the remaining difference reflects true differences in quality of care?

Knaus: The question of how much variation in outcome we must account for, or explain, before we are willing to attribute variations in outcome to hospitals' or physicians' treatment decisions is crucial. It has two separate answers, one scientific and one policy.

From a scientific perspective, we conceptualize patient outcome as a function of the patient's presenting characteristics, the subsequent diagnostic and treatment decisions made by the medical staff,

and the patient's response to the treatment. This response, in turn, will be influenced both by the patient's ability to respond and by the exact nature and timing of the treatment decisions. In a comprehensive analysis, we would collect sequential information on all these factors, and we should be able to explain virtually all of the variation in patient outcome. This type of analysis is complex, but it is definitely possible and will rapidly become more practical as our measurement and computing abilities expand. In our own area of research on intensive care patients, we expect to be able to do this within the next decade because we contend that there is a firm, predictable underlying relationship between acute physiologic abnormalities and outcome from a severe illness (Wagner, Knaus, and Draper, 1986).

It can be argued that response to therapy is part of the dependent, or outcome, variable and, for predictions at any one point in time, patient response obviously is the key outcome. We must acknowledge, however, that case-mix adjustment may have to be dynamic, with updates in the patient's condition as he or she receives treatment. This will probably be most necessary, as previously described, if we are using the case-mix adjustment as part of an ongoing quality assurance system. For quality assessment activities, initial patient characteristics should be sufficient.

In a number of fields, we can already adequately characterize patients prior to treatment and precisely predict their outcomes. When we compare these predicted outcomes with what actually occurred, substantial variations must be a function either of the quality and timing of the subsequent treatment decisions or of the patient's individual ability to respond to treatment.

Although patients vary in their individual abilities to respond, most of this variation is a function of aspects of the patients' initial condition, i.e., age or severity of disease (Knaus et al., 1985b), that are measurable at admission. Far more variation in factors influencing outcomes is found in the nature and timing of treatment decisions. Therefore, from a practical and policy perspective, we believe that it is legitimate initially to attribute unexplained differences between predicted and actual patient outcomes to the quality of care unless the institution can make a strong case for its population having a unique response to that care.

Even if there are no claims that the patients are unique, attributing unexplained variation to any single cause is often difficult, and pinpointing it to a single physician is often impossible. Some of the residual variation may result from random events, some from imperfect measurement of patient risk factors, and some from patient wishes. Ruling out these and competing causes can be done only slowly and through close cooperation between statistical and clinical investigators. In such endeavors, however, the existence of a bad process associated with a poorer than predicted outcome creates a compelling argument for poor quality of care.

Such linkages between process and outcome have been demonstrated in studies contrasting the quality of care of acute burn patients (Wolfe et al., 1983), trauma victims (Baxt and Moody, 1983), and neonatal intensive care units (Avery, Tooley, and Keeler, 1987), as well as our own work in adult medical-surgical ICU's (Knaus et al., 1986). In each case, substantial differences in predicted versus observed patient outcomes were attributed to specific changes in treatment decisions over time or to variations in the type of treatment decisions made among various centers.

In each of these studies, substantial prospective efforts had been undertaken to control for case mix and potential referral bias. None of the researchers can claim a direct cause and effect relationship, but each study had sufficient impact to prompt either further investigations or actual changes in institutional policy. Therefore, the real question ultimately may concern not the amount of the variation that can be accounted for, but rather the strength of the clinical study pointing out variations in outcome and the description of the variations in process that could have directly influenced those outcomes.

For example, in a risk-adjusted eight-center comparison of outcome from neonatal intensive care, one of the eight centers was found to have significantly fewer low-birth-weight survivors who developed chronic lung disease (Avery, Tooley, and Keeler, 1987). That center was also the only one of the eight that used a unique, coordinated, nonintrusive method for treating neonates with acute respiratory problems. Is it not reasonable to conclude that this approach may be influential in the quality of care?

In regard to your question about adjusting risk in capitated systems, the issue is one of measuring chronic disease or physiologic reserve rather than acute severity. This is difficult but, again, far from impossible. Screening new Medicare health maintenance organization patients by their baseline vital capacity, serum cholesterol, and smoking history (or even comparing their physical or physiologic profiles with Framingham data) would be an excellent way to establish baseline population risk and thereby to assess outcome and perhaps even adjust reimbursement levels for population-based risks.

Lohr: Virtually by definition, a psychometrically reliable and valid severity or functional status measure will distinguish among patients as to their probability of being in a given health state; it will (with some degradation in precision) thus provide relevant information on the probability that those patients will remain in that state or move to another. By extension, controlling for differences in severity of illness or functional status among patients cared for by different providers (or by networks of providers in a preferred provider organization) should help us to evaluate the significance of differences in outcomes. Nonetheless, combining these data for dissimilar patients, diagnoses, practice settings, and the like presents formidable, although probably not intractable, problems.

Using such measures to distinguish among providers as to their patients' underlying risk for adverse outcomes is still a developing technology. We still have no proof that true differences in outcomes remain once severity has been taken fully and accurately into account or that whatever differences do remain reflect parameters of the health care process that are amenable to intervention or change.

Krakauer: Case mix, as measured by age, sex, DRG mix, and comorbidities, accounts for 44 to 50 percent of the interhospital variation in mortality associated with hospitalization, depending on the number of variables used. Interestingly, nearly all (85-90 percent) of the explanatory power of the model is attributable to 10-12 DRG's or DRG complexes. In examining mortality rates at the hospital level rather than the patient level, comorbidities appear to be less important. Adding counts of comorbidities to age, sex, and DRG mix increases the predictive power of the model by only 0.5 percent. Thus, the group of selected comorbidities that has substantial and statistically significant power in predicting the likelihood that individual patients will die does not appreciably increase the explained variance at the hospital level. Its contribution is negligible for overall mortality rates and is not at all impressive for mortality rates associated with selected individual conditions. It is, of course, possible that use of a different or more comprehensive array of comorbidities would yield better results.

Another potentially severe limitation of the Medicare data bases results from the lack of adequate information on the physiologic state of the patient (severity of illness) at admission (Wagner, Knaus, and Draper, 1986; Brewster et al., 1985). These limitations hamper comparisons of alternative approaches to management, which are needed to develop objective measures of the effectiveness (Krakauer, 1986) and quality of care. Comparisons of the performance of hospitals are also uncertain because there may be significant differences in severity mix among hospitals (Horn et al., 1985).

Jencks: Lohr has already emphasized that the jury is still out on our ability to infer quality from nonintrusive outcomes. Knaus argues that the burden lies on the hospital to show that poor risk-adjusted outcomes do not indicate poor quality of care. Elsewhere, he has suggested that coordination is so critical to good care that poor coordination is evidence of poor quality, even if no specific clinical errors are apparent; his perspective challenges traditional measurement of the process of care. On the other hand, Krakauer argues that the case-mix adjustment that we can make with Medicare administrative data is not good enough to allow us to infer poor quality of care in hospitals where outcomes are worse than expected.

Further research

What kinds of further research, including comparisons among case-mix systems, are needed?

Brook: A two-pronged approach seems necessary. First, a major research endeavor should be undertaken and funded to examine the value of outcome measures adjusted for age, sex, and condition as a screening tool to identify hospitals with low quality of care. The predictive validity, both positive and negative, as well as the sensitivity and specificity of such a screening tool need to be determined. This research activity should be undertaken immediately, should be done at a disease and hospital level, and should include a well thought out plan to test the validity of risk-adjusted outcome as a screening tool. The best one can do is to determine the effect of a particular case-mix adjustment on how hospitals rank with regard to a particular risk-adjusted outcome. The baseline for such comparisons is the rank of outcomes risk adjusted using only age, sex, and principal diagnosis. In addition, we need to understand how hospital ranking on the basis of both implicit and explicit judgments of process of care compare with hospital rankings on outcomes that are risk adjusted in different ways. By examining such data, we can ascertain the adequacy of an adjusted outcome as a screening tool for identifying hospital quality-of-care problems.

Such research is likely to take 5 to 10 years. What should we do in the interim? Without good data to guide us, I can only suggest that outcome data adjusted for age, sex, and diagnosis should be made publicly available, but with a lot of red flags. In addition, because the data are likely to be controversial and some are likely to be inaccurate, hospitals should be given at least 6 months to respond to preliminary data. If they can show, by means of a process analysis or some other analysis, that the data really are not outcome outliers, information should be corrected before the data are publicly released.

In summary, a concerted research strategy is needed to address the characteristics of outcome measures that have been risk adjusted in different ways. The strategy used should be similar to that used to validate any new screening tool. However, if the screening tool must be used now, then it should be used with caution, and hospitals should have sufficient time before public release of results to allow correction of the errors that are sure to occur.

Iezzoni and Moskowitz: The concerns we have raised suggest research directions. More work is needed on defining meaningful episodes of care and categorizing types of admissions. To assist in describing episodes of care, linkage of hospital and outpatient data bases (e.g., Medicare Part A and B data bases) must become easier. Parallel endeavors should focus on linking data on posthospital outcomes, such as deaths and readmissions, at the patient level.

Investigators should also concentrate on a number of concerns about the classification methods themselves. First, we need to know more about the implications of differences in data documentation and procedure patterns for case-mix comparisons across

hospital types. Second, we need to study the costs of implementing case-classification measures that are useful for quality assessment, because widespread application of such systems at multiple points in time may itself prove prohibitively expensive. Third, we need to prove that we can identify quality shortfalls by measuring patient status at several points over the hospital stay. Finally, we need to devise ways to identify the components of a case-classification system that best predict shortfalls in quality of care. This may involve reweighting individual components of case-mix indexes to come up with new scoring schemes that indicate the presence of a risk that quality has been compromised.

Knaus: By this time, our prejudice must be clear. We need large-scale prospective-measurement reliability studies so that we can move away from use of discharge abstract data. We need to develop better, more reproducible measures of acute severity of illness for the majority of the hospital population, measures that match the precision now possible only in ICU's. These severity measures must be based on objective patient characteristics such as physiology, not therapy. We also need a more precise and reproducible measure of a patient's biological age or physiologic reserve so that we can abandon our reliance on chronologic age and crude definitions of chronic health status. This is especially true considering the increasing number of the elderly.

We believe that this type of fundamental research should be the priority, not comparisons among existing systems. Existing case-mix systems have largely been developed outside the scientific arena with the objective of explaining variations in hospital charges, not patient outcomes. They have had subjective criteria and no firm conceptual base. We do not believe that comparisons among them would create any new or important information.

Rather, we hope that the government will take a longer view, will recognize that scientifically valid case-mix adjustments will be necessary, and will sponsor fundamental work on them now. Proper case-mix adjustment is the only way to ensure that future access, diagnostic, and treatment decisions are based more on patients' ability to benefit than on their ability to pay.

Lohr: In the short run, of course, there may be few alternatives to relying on administrative data, nonintrusive outcomes, and quality assessment based on process measures for monitoring inpatient care. Nonetheless, I would like to reiterate that, because we still have little hard evidence regarding the process-outcome links, we must remain skeptical in this area. Furthermore, the issues warranting intensive research attention that were pertinent a few years ago—collection of data related to general levels of quality of care in the hospital setting, investigations of the impact of the prospective payment system (positive and negative), long-term care, posthospital care, home health care, ambulatory care, and episodes of care (especially in capitated settings)—remain salient (Lohr et al., 1985). The need for more conceptual and practical work on methods is similarly unmet.

In the longer run, however, we should seek progress in two areas. First, the entire health care delivery enterprise is undergoing rapid and revolutionary change (blurring of insurance with provider systems, vertical and horizontal integration of provider institutions, technological advance, and so forth). The conceptual and practical tools necessary for reliable and valid quality assurance have yet to be fully articulated, but they will surely be drawn from expanded views of quality that include not only traditional process and outcome measures but also measures of access, underservice, patient satisfaction with care, and provider satisfaction and morale. Second, the extraordinary growth in the capabilities and applications of computers (e.g., portable computers and "user-friendly" software) may open up remarkable avenues of collecting, processing, and reporting information to payers, providers, and patients that are only dimly perceived at the moment. In this arena, patient and provider satisfaction and process measures may emerge as key quality domains.

Irrespective of the uses to which case-mix adjusters (e.g., severity or health status measures) might be put in the quality arena, they are likely to have important reimbursement applications. Continued and expanded investigations concerning the clinical validity, ease of data collection, interpretability, and other characteristics of the major competing severity measures are needed so that the "black box" aspects of these systems are mitigated to some extent. Studies should be initiated to explore the ways in which health status measures (functional or physical health, mental health, and general health perceptions or ratings measures) could be used to characterize populations (e.g., those enrolled in capitated systems) as well as to monitor change over time.

With respect to quality of care per se, studies to improve quality assessment techniques and quality assurance programs are needed. Elements of this line of research might include expanding the definition of quality of care, as suggested earlier; revisiting the question of the current limitations and future opportunities for peer-review-based quality assurance in Medicare; and examining whether and how diagnosis-specific process criteria can be combined with computer-based quality assessment to monitor and improve care in the full range of settings that serve the elderly.

Krakauer: Case-mix adjustment is clearly essential in assessments of the quality of care, whether of secular trends, of the comparative utility of competing patient management techniques, or of the comparative performance of hospitals. The preliminary exploration of the Medicare data base that I have described indicates that it contains sufficient data, probably of adequate accuracy, at least for the purpose of monitoring. If data on the physiologic state of the patient sufficient to permit adjustment for the severity of illness could be added, the Medicare data base would serve as a secure foundation for a systematic process of assessing and improving quality of care.

Jencks: These comments suggest a core research agenda to which I believe all of the panel subscribes:

- We need research to test the sensitivity and specificity of case-mix-adjusted outcomes as screening tools for aberrant quality of care. Work under way at the Rand Corporation by Drs. Brook, Kahn, and Chassin (work with which Dr. Lohr is associated) is probing the relationship between case-mix-adjusted outcomes and quality of care at the patient level. However, Krakauer's work illustrates that patient-level results may not be safely applied to hospitals, and I am unaware of ongoing projects to explore the relation of risk-adjusted outcome to quality of care at the hospital level.
- We need to validate instruments for measuring the type and severity of illness, as well as the patient's physiologic reserve, at the time of admission rather than at discharge. Dr. Thomas MacKenzie and associates at Queens University (Ontario) are testing currently available alternative case-mix systems and will establish a data base for creating new combinations of data elements. However, this work will not provide new definitions of the nature of admissions, and the sensitivity of systems to a number of parameters will not be tested. In particular, I am not aware of work in progress that will substantially clarify the importance of measuring patient condition at admission rather than a few hours or days later. Further, this work will not develop the new instruments that Knaus considers so important.
- We need better ways of looking at patient-level data so that we can account for differences in patterns and purposes of care and address such issues as how to compare hospitals that treat similar patients with different patterns of admissions and services. Some work in the area of cancer care promises to help us better understand patterns of care for that disease, but extending those efforts to other diseases may take a long time.
- We need expanded definitions of outcomes and of quality of care that include not only death but also patient satisfaction, functional status, and the consequences of limited access to care. This may be the most challenging area in the use of case-mix-adjusted outcomes to identify problems with quality.

At a more general level, I hear both an important consensus and important differences of opinion in the panel regarding our ability to go from data on outcomes that have been adjusted for hospital case mix to inferences about quality of care in hospitals. The consensus is that such inferences are not safe today; more research and more data are necessary. Brook urges that outcomes be considered only screening tools for quality problems and that they be validated as we would validate any other screening tool. Lohr suggests that we are still far from being able to make such inferences. Iezzoni and Moskowitz point to a variety of clinical difficulties in interpreting outcome data. Krakauer emphasizes the need for adequate severity adjustment. Only Knaus suggests that we now have the ability "in a number of fields, to adequately characterize patients prior to treatment and precisely predict their outcomes," and he also counsels the need for further research.

I hear much less consensus on how much further research is needed and what level of inference may soon be safe. On the one hand, I believe that Krakauer and Knaus are saying that such inferences may be possible with moderate amounts of further research. On the other hand, Brook and Lohr argue that we are basically talking about screening tools, and I think this means that, for the foreseeable future, they would require validation of screening results by actual review of patient records and other firsthand evidence. Iezzoni and Moskowitz frame a broad research agenda that perhaps lies somewhere in between.

References

Avery, M. E., Tooley, W. H., and Keeler, J. B.: Is chronic disease in low birth weight infants preventable? A survey of eight centers. *Pediatrics* 79:26-30, 1987.

Baxt, W. G., and Moody, P.: The impact of rotorcraft aeromedical emergency care service on trauma mortality. *Journal of the American Medical Association* 249:3057-3061, 1983.

Booth, P.: Health Standards and Quality Bureau, Health Care Financing Administration. Personal communication. Baltimore, Md. 1987.

Brewster, A. C., Karlin, B. G., Hyde, L. A., et al.: MEDISGRPS: A clinically-based approach to classifying hospital patients at admission. *Inquiry* 22:377-387, 1985.

Brook, R. H., and Lohr, K. N.: Efficacy, effectiveness, variations, and quality: Boundary-crossing research. *Medical Care* 23:710-722, 1985.

Coulton, C. J., McClish, D., Doremas, H., et al.: Implications of DRG payments for medical intensive care. *Medical Care* 23:977-985, 1985.

Donabedian, A.: Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly* 44(Part 2):166-206, July 1966.

Donabedian, A.: *Explorations in Quality Assessment and Monitoring, Volume I, The Definition of Quality and Approaches to its Assessment.* Ann Arbor, Mich. Health Administration Press, 1980.

Donabedian, A.: *Explorations in Quality Assessment and Monitoring, Volume II, The Criteria and Standards of Quality.* Ann Arbor, Mich. Health Administration Press, 1982.

Donabedian, A.: *Explorations in Quality Assessment and Monitoring, Volume III, The Methods and Findings of Quality Assessment and Monitoring: An Illustrated Analysis.* Ann Arbor, Mich. Health Administration Press, 1985.

Horn, S. D., Bulkley, G., Sharkey, P. D., et al.: Interhospital differences in severity of illness. *New England Journal of Medicine* 313:20-24, 1985.

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

Knaus, W. A., and Zimmerman, J. E.: Prediction of outcome from intensive care. *Clinics in Anaesthesiology* 3(4):811-829, Oct. 1985.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E.: APACHE II: A severity of disease classification system. *Critical Care Medicine* 13(10):818-829, 1985a.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E.: Prognosis in acute organ system failure. *Annals of Surgery* 202(6):685-692, 1985b.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E.: Evaluation of outcome from intensive care in major medical centers. *Annals of Internal Medicine* 104:410-418, 1986.

Krakauer, H.: Assessment of alternative technologies for the treatment of end-stage renal disease. *Israel Journal of Medical Sciences* 22:245-259, 1986.

Krakauer, H.: Outcomes of In-Hospital Care of Medicare Patients in 1983-1985: The Medicare Experience. Draft Report. Office of Medical Review, Health Standards and Quality Bureau, Health Care Financing Administration. Baltimore, Md. 1987.

Lohr, K. N., Brook, R. H., Goldberg, G. A., et al.: *Impact of Medicare Prospective Payment on the Quality of Medical Care: A Research Agenda.* Rand Report No. R-3242-HCFA. Prepared for Health Care Financing Administration. Santa Monica, Calif. The Rand Corporation, Mar. 1985.

Health Care Financing Administration: DRG Refinement: Outliers, Severity of Illness, and Intensity of Care. *Report to Congress.* HCFA Pub. No. 03254. Office of Research and Demonstrations. Washington. U.S. Government Printing Office, Sept. 1987.

McClish, D. K., Russo, A., Franklin, C., et al.: Profile of medical ICU vs. ward patients in an acute care hospital. *Critical Care Medicine* 13:381-386, 1985.

Public Health Service and Health Care Financing Administration: *International Classification of Diseases, 9th Revision, Clinical Modification.* DHHS Pub. No. 80-1260. Public Health Service. Washington. U.S. Government Printing Office, Sept. 1980.

Wagner, D. P., Knaus, W. A., and Draper, E. A.: Physiologic abnormalities and outcome from acute disease. *Archives of Internal Medicine* 146:1389-1396, 1986.

Wolfe, R. A., Roi, L. D., Flora, J. D., et al.: Mortality differences and speed of wound closure among specialized burn care facilities. *Journal of the American Medical Association* 250:163-166, 1983.