

The Quality of Quality Measurement in U.S. Nursing Homes

Vincent Mor, PhD,¹ Katherine Berg, PT, PhD,^{1,2} Joseph Angelelli, PhD,¹ David Gifford, MD, MPH,^{1,3} John Morris, PhD,⁴ and Terry Moore, RN, MPH⁵

Purpose: This article examines various technical challenges inherent in the design, implementation, and dissemination of health care quality performance measures. **Design and Methods:** Using national and state-specific Minimum Data Set data from 1999, we examined sample size, measure stability, creation of ordinal ranks, and risk adjustment as applied to aggregated facility quality indicators. **Results:** Nursing home Quality Indicators now in use are multidimensional and quarterly estimates of incidence-based measures can be relatively unstable, suggesting the need for some averaging of measures over time. **Implications:** Current public reports benchmarking nursing homes' performances may require additional technical modifications to avoid compromising the fairness of comparisons.

Key Words: *Quality indicators, Quality measurement, Assessment methodology, Quality of care, Benchmarking, Performance measures, Nursing homes*

Health care providers' and insurers' accountability for the services that they render is increasingly a subject of concern to regulators, advocates, and consumers (Epstein, 1998). As efforts to contain costs while increasing competition in the health care field have advanced, concerns about deteriorating quality of care now receive even more attention than

health care costs. Measuring health care quality and comparing providers' performance has emerged as the most hopeful strategy for holding them accountable for the care they provide (Jencks, 1994).

Quality measurement, performance monitoring, and quality improvement are a constant refrain in the United States (Cohn, Corrigan, and Donaldson, 2000). Hospitals regularly produce statistics regarding their performance in selected clinical areas, and most are now surveying their patients about their satisfaction with the care they receive (Edgman-Levitan & Cleary, 1996; Rosenthal et al., 1998). Managed care companies are routinely compared with how well they ensure that preventive health services are delivered to their subscribers (Scanlon, Darby, Rolph, & Doty, 2001). The performance of surgeons is routinely monitored in terms of mortality and complication rates, whereas ambulatory practices' performance in holding down waiting times is touted, particularly in highly competitive markets (Hannan, Siu, Kumar, Kilburn, & Chassin, 1995; Marshall, Shekelle, Leatherman, & Brook, 2000). Finally, nursing homes in the United States are being compared with numerous quality indicators derived from inspections, as well as routinely collected clinical data (Phillips, Zimmerman, Bernabei, & Jonsson, 1997).

There are numerous pressures stimulating the wholesale adoption of quality measurement and benchmarking. Regulators, particularly the Centers for Medicare and Medicaid Services (CMS), have been aggressively pushing quality measurement. Purchasers and managed care companies have been forced to prove that their networks provide quality care, and the National Committee for Quality Assurance has promoted performance comparisons among managed care organizations across the country, making such comparative information available to interested parties on the Internet (Schneider, Riehl, Courte-Wienecke, Eddy, & Sennett, 1999). Increasingly, computerized data systems in hospitals, ambulatory office claims, and the recently nationally implemented Minimum Data

This study was supported in part by a MERIT grant from the National Institute on Aging (AG-11624) and Contract 500-98-026 from the Centers for Medicare and Medicaid Services to Abt Associates, with subcontracts to Brown University and the Hebrew Rehabilitation Center for the Aged, Boston.

Address correspondence to Vincent Mor, PhD, Department of Community Health, Brown University School of Medicine, Box G-A418, Providence, RI 02192. E-mail: Vincent_Mor@brown.edu

¹Brown University Center for Gerontology and Health Care Research, Providence, RI.

²McGill University School of Physical and Occupational Therapy, Montreal, Quebec, Canada.

³Rhode Island Quality Partners, Providence, RI.

⁴The Research and Training Institute, the Hebrew Rehabilitation Center for the Aged, Boston, MA.

⁵Abt Associates, Inc., Cambridge, MA.

Set (MDS) used to assess nursing home residents, all provide the data from which measures of quality performance can be created at minimal cost.

Purpose and Methods

Using the nursing home industry and measures of nursing home quality as an example, this paper examines the conceptual assumptions and empirical issues that underpin the construction, application, and dissemination of performance measures. Although the paper focuses more on the problems than the solutions, suggestions of strategies for overcoming these issues are periodically offered. Data are presented from a series of analyses conducted using information drawn from various sources, most consistently from the CMS repository of MDS data that have been routinely collected and archived since late 1998.

A uniform resident assessment system was mandated under the OBRA 1987 Nursing Home Reform legislation as one of several recommendations from the 1986 Institute of Medicine report on Nursing Home Quality. An assessment process and summary recording format, the MDS, was mandated in all Medicare/Medicaid nursing facilities as of 1991. The assessment is done on admission and annually thereafter with documentation of any changes in status on a quarterly basis. Since October 1998, all MDS records have been transmitted via state public health agencies to a national repository maintained by CMS and are to be used to assist the existing Survey and Certification process in monitoring the quality of care provided to residents of Medicare/Medicaid certified nursing facilities.

Conceptual Issues Inherent in Applying Performance Measures

Establishing quality measures and interpreting and communicating their meaning to those who might make decisions based on them requires a shared understanding of a number of conceptual issues. First, it requires a shared understanding of quality among those creating, applying, and living under the stricture of performance measures. Second, the linkage between *process* and *outcome* quality measures is not always well understood, meaning that good facility performance (*viz.* a process standard) may not guarantee good performance on a presumably related outcome measure (Ramsey, Sainfort, & Zimmerman, 1995). Third, a standard set of quality performance measures assumes that all providers have the same goals and seek to treat the same kinds of patients. This may not be the case; nursing homes collaborating with hospices for terminal care management may not seek to achieve functional improvement, making this an inappropriate performance measure for some of the residents

they serve (Miller, Gozalo, & Mor, 2001). Fourth, assessing provider performance based on patient outcomes implies that providers are *accountable* for a substantial degree of variation in residents' outcome, even though facilities may actually have limited ability to influence all outcome measures. Fifth, although it seems obvious that performance measures do not measure provider quality per se, it is often the case that the measures are reified as signifying the actual quality of care provided in ways that are relevant to the potential consumer of the service. Finally, regardless of who uses quality measures, they need to be readily understood. Whether they necessarily need to be fully transparent and replicable by all parties is another issue, probably more related to the trust that consumers have in providers and that both have in the government or entity that manages the quality measurement enterprise.

These conceptual issues are inherent in the use of performance measures intended to differentiate the quality of care providers offer. Even if all the technical issues enumerated herein could be solved, these conceptual issues would remain. Although obvious, they are important to keep in mind in evaluating the appropriateness of such measures.

Technical Complications in Executing Performance Measures

Although various forms of performance measurement are published on almost a daily basis in the United States, numerous complex technical problems remain that may undermine their validity. In the paragraphs that follow, these are discussed using examples drawn from the realm of quality monitoring measures in the U.S. nursing home industry. The issues addressed include data reliability and validity; the sample size on which stable measures can be based; the multidimensionality of quality measures; the implications of transforming data into ranks; the complexity of risk adjustment; the influence of patient selection bias; and the differences in providers' assessment practices and how censoring, or discharges, influences performance measures.

Data Reliability and Validity

If the data used to construct performance measures are not consistently obtained and are not related to quality in the manner expected, the resulting measures are unlikely to be broadly accepted as meaningful. Reports in the literature generally support the contention that when nurse assessors are properly trained, their MDS records resemble those that "gold standard" research nurses would have done had they assessed the resident and recorded the data in the MDS (Hawes et al., 1995; Morris et al., 1997). Furthermore, epidemiological evidence suggests that

measures of function, diagnostic information, and treatment data included in the MDS are valid and relate to resident characteristics as expected (Gambassi et al., 1999). Despite these positive indications of the potential for reliable MDS data for use in constructing quality measures, to validly compare nursing homes' performance, similarly reliable and valid data must be present in all facilities. Audits by the Office of the Inspector General revealed that whereas systematic procedures are in place in most facilities to train staff, disagreements among facility and audit assessors were found in more than 17% of data elements used to define case-mix group membership, higher in the functional status measures (Office of the Inspector General, 2000). Although many states report the existence of data integrity audit procedures, these are not done uniformly and are not coordinated with federal initiatives in this area (General Accounting Office, 2002).

One way to overcome some of the negative consequences of less than perfect interfacility reliability in the calculation of quality measures is to base those measures on multi-item scales or on change in status. Most current quality measures are precisely defined based on a specific MDS item. Were some measures defined as relative change in an activity of daily living scale, measures of interrater reliability such as the weighted kappa, would be more positive than an exact degree of agreement. Because numerous elements in the MDS have been found to form psychometrically reliable and valid summary scales of cognitive, physical, and emotional functioning, as well as pain severity, the basis for such outcome-based quality measures is present (Fries, Simon, Morris, Flodstrom, & Bookstein, 2001; Morris et al., 1994; Morris, Fries, & Morris, 1999).

Sample Size and Measure Stability

In many instances, the indicators of quality being measured are rare, or should be. Postsurgical infection rates, life-threatening medication errors, etc., are likely to be small (McClellan & Staiger, 1999). Similarly, the number of new nursing home-acquired pressure ulcers in a 3-month period is hopefully quite small (Berlowitz, Bezerra, Brandeis, Kader, & Anderson, 2000). This means that many of the performance measures of interest will have very large standard errors, or bands, around which the true estimate of the event rate might actually lie. For example, if the true 3-month incidence of pressure ulcers is 5%, the 95% confidence interval around the estimate for any given facility would range from 1% to 11% in a facility with 100 residents in the denominator. Not until the number of observations exceeds 200 do the confidence intervals around the observed rate drop to less than twice the size of the point estimate.

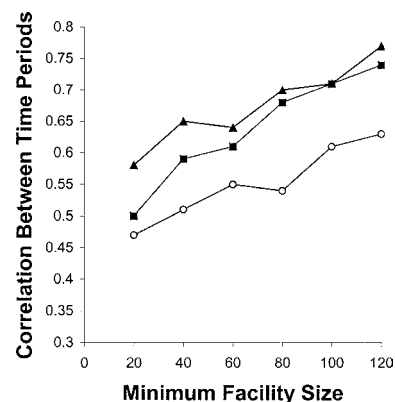


Figure 1. Effect of facility size and number of observation quarters on bowel incontinence quality measures (BOW) interquarter correlations. Simulation is based upon Minimum Data Set data —○— Bow 1 quarter; —■— Bow 2 quarters; —●— Bow 3 quarters.

In some circumstances, using the 95% confidence interval results in estimates that are too conservative. However, even using an 80% confidence interval with a small number of observations means that potentially meaningful differences might be missed. Another approach to increasing sample size is to increase the duration of the observation period included to calculate the denominator or numerator. Unfortunately, the more a performance measurement relies on historical data, the less it may truly reflect the current performance of that provider. This increased stability may have the paradoxical effect of labeling “poor” a provider that has overcome historical problems, whereas labeling “good” another provider that has let standards slide.

The implications of small sample size for the stability of quality measures, particularly those focused on incidence or change, can be readily seen in Figure 1. Examining the impact of facility size and the number of months of observation in constructing the performance measure of *change in bowel functioning independence*, we used data from all nursing homes in Massachusetts in 1999. Three types of measures of change in independent bowel functioning were constructed, one based on just the immediate past quarter, another based on the past two quarters of observation, and the last based on the past three quarters. Each was created for two successive quarters going forward, meaning that there was overlap between the quarters included in the performance measures based on multiple quarters. We correlated adjacent measures (Quarter 1 with Quarter 2) for all three types of measures. As can be seen, the correlation between the first and second quarters' performance measures was under .5 for the smallest facilities. As facility size increases, the correlation among measures increases, but more for those measures that have the longer look-back period, suggesting that stability is both a function of size and the way the performance measure is constructed.

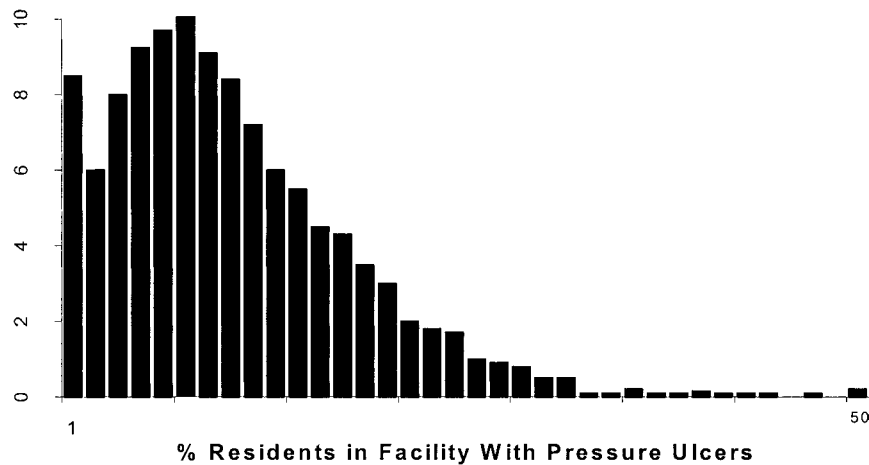


Figure 2. National facility distribution of pressure ulcer prevalence. Percent of Facilities With Pressure Ulcers: OSCAR 2000. From OSCAR = Online Survey Certification and Reporting.

Because many pertinent quality measures refer to only a subset of the nursing home population and more than one quarter of nursing facilities have fewer than 60 beds, to increase the stability of quality measures some increase in the duration of the observation period will be necessary. However, even this may not be sufficient for small facilities with little turnover among residents without including repeated measures on the same resident in the summary. For this, a more sophisticated approach using Generalized Estimating Equations is technically possible and correct, but complex to implement broadly and not transparent to the various users of these measures. Regardless of how stability is increased, it does make sense to have a minimum denominator size before any information about the performance of a facility is reported.

Transforming Performance Measures into Ranks

In many instances, the metric in which a performance measure is expressed (e.g., incidence rate or percentage of the population deteriorating) results in a highly skewed distribution, because most providers have relatively low event rates. However, these raw rates are often transformed into percentiles for ease of interpretation and to allow for comparisons across different performance measures. The resulting ranks range from 0 to 100, with allowance for ties.

While ordinal numbers (first, 31st, etc.) are readily understood, depending on the underlying distribution, they can be quite misleading. Figure 2 presents the distribution of U.S. nursing facilities in terms of the prevalence of pressure ulcers in 2000. The highest rate (approximately 50%) would have the highest (worst) rank. However, the difference between the 70th percentile and the 25th percentile is actually quite small in terms of the actual pressure ulcer rate (approximately 9% vs. 5%). Creating the appearance of difference when there might be little

meaningful difference in performance can be misleading to the public, as well as to other users of such information.

Rather than convert rates to ranks, where possible, it would be desirable to establish “benchmarks,” or standards, based on consensus or even what can be attained under optimal conditions by a good home. Experiments in nursing facilities have revealed that numerous clinical problems can be prevented or kept to a very low level, and some practices—like physical restraints—can be eliminated in the proper supportive environment. Even in the absence of consensus about a “standard,” a probabilistic approach is possible. Under the current CMS public reporting of nursing home quality measures demonstration program, the state of Rhode Island has chosen to classify facilities based on whether their performance is superior or inferior to the median facility using a confidence interval of 50% (rather than the 95% confidence interval). Correcting for the confidence interval around a home with an average of 100 beds, this is basically equivalent to determining whether the facility is in the top or the bottom 15% of homes. Although the cutpoint is still based on an empirical distribution, which can be quite skewed, it avoids the problem of users attributing meaning to a 10 percentage point difference in the facility rank.

Complexity of Risk Adjustment

A great deal has been written about the importance of risk adjustment to compare apples to apples and not to oranges. Many performance measures will be very sensitive to the kinds of patients providers serve. Patients with a pressure ulcer before nursing home admission, a known risk factor for acquiring a future pressure ulcer, are not randomly distributed across nursing homes in a given community, perhaps because hospital discharge planners

and physicians know that some nursing facilities excel at pressure ulcer care and prevention, and so refer their at-risk patients there. If a performance measure does not take into account this difference in the case-mix of the providers being compared, it will be penalizing the provider with the reputation for excellence in pressure ulcer care (Berlowitz et al., 2000).

Although this kind of risk adjustment is fairer because it levels the playing field without rewarding facilities for skimming low-risk patients, it is possible to overadjust (Zimmerman et al., 1995). Sticking with the pressure ulcer example, incontinence of urine and feces is also a risk factor for acquiring a pressure ulcer. However, incontinence, like pressure ulcers, could arise as a result of the same underlying phenomenon (e.g., poor nursing home care). Adjusting for incontinence when long-stay residents' incontinence may have been caused by the facility improves facilities' observed performance in a way that is unfair. Unfortunately, there are no easy answers to the whole issue of under- or overadjustment, and each performance measure must be carefully considered individually.

Other risk adjustment strategies are to exclude cases from the denominator. For example, residents with a psychiatric diagnosis are not included in the denominator used to calculate the rate of antipsychotic use, because their diagnosis means that they have a presumptive reason for use of antipsychotics. Excluding comatose residents or those determined to be terminally ill from the calculation of outcome-based quality measures is also done.

Whether to risk adjust via stratification, direct standardization or using regression-based approaches is beyond the scope of this article. Risk adjustment using statistical modeling using regression results in less than transparent rates since, depending on the degree of adjustment; it is difficult to cross-walk between the observed and adjusted measures. On the other hand, stratification places greater pressure on the precision of measurement of both the quality indicator, as well as the stratification variable(s), and reduces the sample size on which the measure is calculated for any given home. If transparency can be sacrificed, direct standardization that allows for multiple strata, but results in a consolidated measure, has the advantage of being more readily understood than a regression model and yet suffers less from small sample size.

The Multidimensionality of Quality

Research from the literature on hospital performance, health care plan performance, and also the performance of nursing homes suggests that there is only a low level of correlation among the various measures of quality (Rosenthal et al., 2000; Sloss et al., 2000). Hospitals with low rates of mortality for patients presenting with an acute myocardial

Table 1. Characteristics of Facilities Ranking Among the Worst One Half on Three Performance Indicators in Ohio: 1999

Facility Characteristic	Ranked in Bottom Half on All Three QIs	N	Mean	SD
% Medicaid	No	470	.6260	.1934
	Yes	75	.6912	.1545
Total deficiencies	No	470	8.24	6.83
	Yes	75	10.48	9.17
Health deficiencies	No	470	5.40	5.55
	Yes	75	7.92	8.43

Note: QI = quality indicator.

infarction may have higher rates of some other type of undesirable outcome. The National Committee for Quality Assurance that assembles data on health plan performance on a standard set of measures finds relatively low levels of correlation across the actual performance of health plans on many measures pertinent to primary care and prevention.

Using a large number of MDS-based performance measures from numerous domains, we examined data from more than 1,500 facilities in five states (New York, South Dakota, Maine, Mississippi, and Kansas) using standard data reduction techniques such as factor analysis and multiple regression analysis. Although we did find that worsening bowel and worsening bladder incontinence measures were reasonably associated, the average intercorrelations among performance measures were low. For example, the correlation between the proportion of patients taking antipsychotics, after excluding those with psychiatric diagnoses, and the proportion of residents in physical restraints is only .04.

We went one step further by actually applying risk-adjusted performance measures for pressure ulcers, antipsychotic use, and inadequate pain management to identify those facilities in Ohio in 1999 that were in the bottom half of the distribution on all three measures. (We initially used the bottom 20% of facilities, but found that there was virtually no overlap.) Table 1 presents the results of these analyses and compares the 75 facilities who met this criteria to the remainder in terms of the numbers of survey and certification deficiencies (based on state inspections) they were found to have, as well as the proportion of Medicaid residents in the home. As can be seen, the differences are relatively small in light of the extreme selection of facilities performing poorly on all three measures.

These results fly in the face of a common sense understanding of quality indicators and our natural expectation that good nursing homes will be able to achieve good outcomes consistently across most measures of performance that are important. They also defy the common expectation among consumers that there is such as thing as the best overall

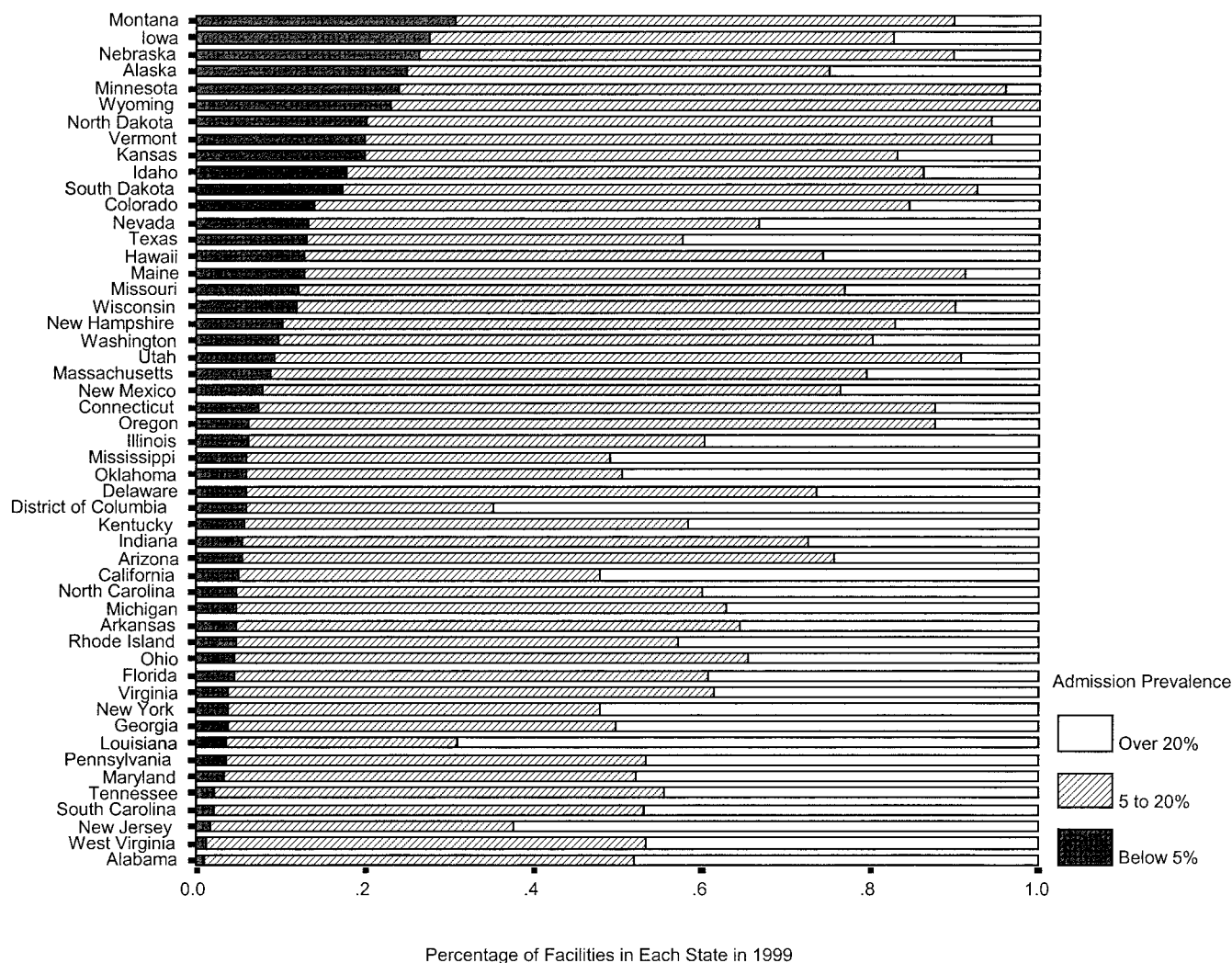


Figure 3. Within- and between-state variation in facility pressure ulcer prevalence at time of admission.

provider. Nonetheless, these findings are not inconsistent with providers' knowledge that achieving excellence across a wide array of parameters is very difficult and that doing one set of functions well does not necessarily generalize if a different technology or approach is required. From this perspective, good providers will want to know their performance in many different quality domains. On the other hand, consumers will likely want less complex information that best summarizes the multiplicity of measures. This is an area that requires considerably more research to properly inform how this information can be best transmitted.

Selection Bias

America's nursing homes are used for many purposes, in many cases going beyond the common public perception of these facilities as a place of residence for highly impaired elders. Nursing homes have become increasingly specialized, with some focusing on postacute care patients who remain in

the facility for only a short period, others focusing on patients with psychiatric disorders, and still others providing hospice care to a significant number of their patients (Banazak-Holl, Zinn, Brannon, Castle, & Mor, 1997). This means that, in many markets, hospitals may refer their postacute referrals to selected facilities; knowledgeable agents will recommend facilities differentially on the basis of the patients' needs.

We examined the relationship between the percentage of new admissions to nursing homes with a pressure ulcer at the time of admission and the prevalence of pressure ulcers among all residents of that home who had already remained in the facility for more than 6 months. We found that these two measures were correlated .34 in all 566 Massachusetts nursing facilities in 1999.

Figure 3 presents the distribution of facilities in each state in terms of the proportion of residents admitted with a pressure ulcer, categorized into three groups, those with more than 20% of admissions with pressure ulcers, those with below 5%, and

Table 2. Percentage of Cancer Patients in Nursing Homes in Six States Rated as Having Daily Pain

Patient Status	Kansas	Maine	Miss	New York	Ohio	South Dakota
New admits (% pain)	39.5%	48.1%	23.3%	26.7%	26.2%	45.4%
Resident in July 1996	32.3	42.3	19.3	16.6	20.3	36.4

those in between. As is clear, there are substantial within- and between-state differences in the proportion of residents admitted to facilities with a pressure ulcer. Some facilities admit many residents with pressure ulcers, whereas other facilities may still see very few of this type of resident, thus suggesting that there is natural triage underway.

Patients entering a nursing home with a pre-existing pressure ulcer, if they survive and remain in the facility, are at increased risk of acquiring another ulcer in the future. Facilities that admit many patients with a pressure ulcer are bound to have some of them remain in the facility long enough to be incorporated into the denominator of long-stay residents on whom the prevalence, or even incidence, of pressure ulcers is calculated. Without some adjustment for this selection phenomenon, facilities that specialize in these kinds of patients may be unduly penalized by merely excluding the short-stay residents from the calculation of the pressure ulcer rate. It is not sufficient to merely exclude from the denominator newly admitted patients, because patients entering a home with a pre-existing pressure ulcer who become chronic and long-stay residents will nonetheless still be at increased risk of developing a new ulcer. Rather, some adjustment for long-stay residents' status on admission (at least for some period of time) might be needed to better account for residents' predisposition based on earlier experience. This approach may be applicable to quality measures beyond pressure ulcers, but needs to be carefully considered and examined in the future.

Ascertainment Bias

Related to the issue of risk adjustment and selection is how to cope with interfacility differences in assessment practices. For example, Bernabei and his colleagues (1998) have shown that the prevalence of pain is strongly related to selected demographic and clinical characteristics of the patients. There is also evidence suggesting that the prevalence of pain varies as a function of state, even for the more homogeneous patients with a diagnosis of cancer, most of whom do not survive 6 months (Teno, Weitzen, Wetle, & Mor, 2001). Table 2 contrasts the rates of daily pain assessed among cancer patients newly admitted to nursing homes in each of six different states. As can be seen, in some states, otherwise similar patients are much more likely to be rated as being in pain than in others. This

differential is just as true among patients newly admitted to the facilities as it is of longer stay residents with a cancer diagnosis. It is unlikely that these large differences are attributable to the underlying biological differences in the patient populations. Rather, the standards and expectations for assessing the presence of pain among nursing home patients must be presumed to vary as a function of different approaches to training in the assessment of pain (and other clinical issues) that have been instituted by facilities.

Ascertainment has the perverse effect of penalizing those who seek to identify and document the problem. This phenomenon can be presumed to influence quality measurement in all sectors of the health care industry, because chart audits are only as complete as the completeness of the information they contain, and coding for Medicare claims is also potentially biased by upcoding to maximize Diagnosis Related Groups reimbursement, something presumably done differentially. There is no quick fix to this methodological concern. However, the design and selection of quality measures, particularly for public reporting, should focus on those areas in which the phenomenon is less pervasive, particularly avoiding things like pain and mood disturbance that are difficult to audit in the charts. This is not to say that these measures not be used to inform facilities' quality improvement efforts nor to guide state inspectors, but caution should be exercised before they are reported to the public.

Censoring Observations Related to Longitudinal Outcomes

In many instances, the desired outcome of interest is a measure of the patient's change in status or the incidence of an untoward event. These types of longitudinal measures require data about the resident at several points in time, or at least require a monitoring system to determine whether an event of interest (e.g., rehospitalization) has occurred (Porell, Caro, & Silva, 1998). Using administrative records, such as Medicare claims, is generally a relatively good source for monitoring things like hospitalization or mortality. Measuring the incidence of pressure ulcers requires observing patients for a period. Those patients discharged to hospital, or leaving the facility for another reason, are censored: by using current approaches for constructing this measure, they are not counted in the calculation of a rate. If some facilities are inclined

to discharge patients because they acquire pressure ulcers but before they have to be assessed, they effectively censor their clinical problems by passing them on to a hospital or another provider. Because those facilities apt to discharge their problems will appear to have superior outcomes to those that retain their clinically trying cases, this censoring introduces bias into the comparisons. Although problematic for calculating longitudinal quality measures for long-stay residents, this problem of differential censoring is particularly severe in the case of short-term, postacute patients who are in a nursing home only long enough to receive a few days of therapy and are then discharged before a follow-up assessment is done. This results in censoring the "successes." Unless the denominator is taken into consideration, it could look like the most successful facilities have the worst performance on measures such as functional improvement.

For both short- and long-stay resident-based measures, it is important to test facilities for the comparability of their censoring rates. To the extent that censoring seems to affect the observed rates of things ranging from pressure ulcer rates to functional improvement, a facilities' performance should not be reported because it is biased relative to the facilities that have lower rates of censoring. Although dropping such facilities may not be satisfying, it does not communicate possibly erroneous information about quality performance.

Applications of Quality Measures

There are four basic audiences for performance measures: providers themselves, regulators, purchasers, and consumers. The value and purpose of the information contained in performance measures varies as a function of these audiences (Frankenfield, Marciniak, Drass, & Jencks, 1997). Providers have been the staunchest advocates for using quality measures to identify care problems that can be addressed as a part of a continuous quality improvement program. The establishment of quality indicators based on uniformly available data, no matter how technically compromised, has provided the basis for quality improvement programs now being undertaken in many U.S. nursing facilities (Castle, 1999; Rantz et al., 2000).

CMS, which is responsible for setting standards and regulating nursing homes in the United States, has recently adopted the practice of using performance measures to supplement and guide the traditional facility survey and certification process that is accomplished by state Departments of Health throughout the country. Reports of a facility's performance on numerous dimensions of quality are provided to the regulatory inspectors to guide the inspection process to focus on identified quality problems. The quality measures are used as a starting

point that may be then validated based on the inspection protocol. As such, the measures are treated as mere indicators of potential quality problems of successes, rather than being interpreted as ipso facto measures of quality.

Purchasers of health care (e.g., insurance companies, employers) have begun to urge providers to compete on both price and quality. Unfortunately, there is little evidence that, despite the increasing availability of performance measures, purchasers and employers actually use them to decide which provider or insurer to select (Marshall et al., 2000). Finally, consumers, their families, and advocates have periodically called out for the public release of quality data so that they can actively select the providers that best meet their needs (Edgman-Levitan & Cleary, 1996). Last year CMS decided to disseminate and publish established quality indicators on the Internet and via other broadly distributed print media, beginning with nursing homes (HHS press release, October 2001). In April 2002, a demonstration program was begun in six states, and the plan is for the system to be expanded nationally later in 2002. How consumers and their advocates respond to the availability of this information will be very important to the future of the whole topic of quality measurement.

Summary

Benchmarking of health care providers is underway in virtually every sector of the health care industry in the United States. The oft-noted concerns about its meaning and utility do not appear to have slowed the rush toward the measurement of quality performance and the publication of the results (Hofer et al., 1999; Shojania & Wachter, 2000; Wennberg, 2000). In the long-term care arena, the universal availability of the MDS has served as a stimulus to develop and implement quality measurement and reporting. CMS is driving the long-term care field by its commitment to publish, for universal Internet access, summary quality indicators about each nursing facility in the country.

It is still too early to tell what the consequences of this movement will mean, both in general and in light of the numerous technical and conceptual problems inherent in interpreting the resulting quality ratings as currently constructed. Some believe that the publication of surgeon-specific cardiovascular surgery mortality rates contributed to the significant reduction in postsurgical mortality observed in New York and Pennsylvania (Hannan et al., 1998). Others have argued that the observed changes were merely part of a pre-existing trend and that publication of the report cards had few benefits (Epstein, 1998; Schneider & Epstein, 1996). To date, there are no data on how the nursing home industry

is responding to the publication and dissemination of this information (Hawryluk, 1999).

There is some concern that there will be a backlash among providers that may undermine the acceptability of the quality monitoring process. This appears to have happened in the ambulatory care sector, in which physicians feel that small samples (inadequate or costly), data collection, and selection bias all serve to undermine the perceived validity and fairness of any report cards of physicians' practice performance (Greene, Barlow, & Newman, 1996; Greenfield, Kaplan, Kahn, Ninomiya, & Griffith, 2002; Hofer et al., 1999; Shojania & Wachter, 2000). Although data in the long-term care field are clearly not the problem (although there are clearly problems of uniformity and comprehensiveness), rare events and selection bias because of the increased specialization of nursing homes in the United States can compromise the fairness of performance comparisons.

Despite the problems, a variety of new statistical models that address some of the worst technical problems are now under development and may be more generally available for application to performance benchmarking efforts (McClellan & Staiger, 1999, 2000; Normand, Glickman, & Gatsonis, 1997). Some of these new techniques may make the resulting information less transparent, and others will require policy makers to balance the need for measure stability and sensitivity. This and similar issues related to the multidimensionality of quality will require more thought as to their implications and how the end users of the information understand and use it. The entire health care industry is still in the relatively early stages of this revolution of accountability; this is clearly true for nursing homes (Chassin, 1998). It will be interesting to see how this movement evolves over the coming decades in the United States and whether and how it will be exported to other countries concerned with these issues.

References

- Banaszak-Holl, J., Zinn, J. S., Brannon, D., Castle, N. G., & Mor, V. (1997). Specialization and diversification in the nursing home industry. *Health Care Managing Review*, 3, 91-99.
- Berlowitz, D. R., Bezerra, H. Q., Brandeis, G. H., Kader, B., & Anderson, J. J. (2000). Are we improving the quality of nursing home care: The case of pressure ulcers. *Journal of the American Geriatrics Society*, 48, 59-62.
- Bernabei, R., Gambassi, G., Lapane, K., Landi, F., Gatsonis, C., Dunlop, R., et al. (1998). Management of pain in elderly patients with cancer. SAGE Study Group. Systematic Assessment of Geriatric Drug Use via Epidemiology. *Journal of the American Medical Association*, 279(23), 1877-1882.
- Castle, N. G. (1999). Outcomes measurement and quality improvement in long-term care. *Journal of Healthcare Quality*, 21, 21-25.
- Chassin, M. (1998). Is health care ready for six sigma quality? *Milbank Quarterly*, 76, 565-591.
- Cohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: National Academy Press.
- Edgman-Levitan, S., & Cleary, P. D. (1996). What information do consumers want and need? *Health Affairs* (Millwood), 15, 42-56.
- Epstein, A. (1998). Rolling down the runway: The challenges ahead for quality report cards. *Journal of the American Medical Association*, 279, 1691-1696.
- Frankenfield, D. L., Marciniak, T. A., Drass, J. A., & Jencks, S. (1997). Quality improvement activity directed at the national level: Examples from the Health Care Financing Administration. *Quality Managing Health Care*, 5, 12-18.
- Fries, B. E., Simon, S. E., Morris, J. N., Flodstrom, C., & Bookstein, F. L. (2001). Pain in U.S. nursing homes: Validating a pain scale for the minimum data set. *The Gerontologist*, 41, 173-179.
- Gambassi, G., Lapane, K. L., Landi, F., Sgadari, A., Mor, V., & Bernabei, R. (1999). Gender differences in the relation between comorbidity and mortality of patients with Alzheimer's disease. Systematic Assessment of Geriatric drug use via Epidemiology (SAGE) Study Group. *Neurology*, 53, 508-516.
- General Accounting Office. (2002). *Nursing homes: Federal efforts to monitor resident assessment data should complement state efforts* (Publication No. GAO-02-279). Washington, DC: Author.
- Greene, B. R., Barlow, J., & Newman, C. (1996). Ambulatory care groups and the profiling of primary care physician resource use: Examining the application of case mix adjustments. *Journal of Ambulatory Care Management*, 19(1), 86-89.
- Greenfield, S., Kaplan, S. H., Kahn, R., Ninomiya, J., & Griffith, J. L. (2002). Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results. *Annals of Internal Medicine*, 136, 111-121.
- Hannan, E. L., Siu, A. L., Kumar, D., Kilburn, H., & Chassin, M. R. (1995). The decline in coronary artery bypass graft mortality in New York State. *Journal of the American Medical Association*, 273, 209-213.
- Hawes, C., Morris, J. N., Phillips, C. D., Mor, V., Fries, B. E., & Nonemaker, S. (1995). Reliability estimates for the minimum data set for nursing home resident assessment and care screening (MDS). *The Gerontologist*, 35, 172-178.
- Hawryluk, M. (1999). Quality indicators help focus survey. QIs to pinpoint areas for further investigation. *Provider*, 25, 26-27, 29-30, 33-34 passim.
- Hofer, T. P., Hayward, R. A., Greenfield, S., Wagner, E. H., Kaplan, S. H., & Manning, W. G. (1999). The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *Journal of the American Medical Association*, 281, 2098-2105.
- Jencks, S. (1994). The government's role in hospital accountability for quality of care. *Joint Commission Journal on Quality Improvement*, 20, 364-369.
- Marshall, M. N., Shekelle, P. G., Leatherman, S., & Brook, R. H. (2000). The public release of performance data: What do we expect to gain? A review of the evidence. *Journal of the American Medical Association*, 283, 1866-1874.
- McClellan, M., & Staiger, S. (2000). *The changing hospital industry: Comparing not-for-profit and for-profit institutions*. Chicago: The University of Chicago Press, pp. 93-112.
- McClellan, M., & Staiger, S. (1999). The quality of health care providers. NBER Working Paper No. 7327, Cambridge, MA.
- Miller, S. C., Gozalo, P., & Mor, V. (2001). Hospice enrollment and hospitalization of dying nursing home patients. *American Journal of Medicine*, 111, 38-44.
- Morris, J. N., Fries, B. E., Mehr, D. R., Hawes, C., Phillips, C., Mor, V., et al. (1994). MDS cognitive performance scale. *Journal of Gerontology: Medical Sciences*, 49, M174-M182.
- Morris, J. N., Fries, B. E., & Morris, S. A. (1999). Scaling ADLs within the MDS. *Journal of Gerontology: Medical Sciences*, 54, M546-M553.
- Morris, J. N., Nonemaker, S., Murphy, K., Hawes, C., Fries, B. E., Mor, V., et al. (1997). A commitment to change: Revision of HCFA's RAI. *Journal of the American Geriatrics Society*, 45, 1011-1016.
- Normand, T., Glickman, M. E., & Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, 92, 803-814.
- Office of the Inspector General. (2000). *Nursing home resident assessment quality of care* (Publication No. OEI-02-99-00040). Washington, DC: Author.
- Phillips, C. D., Zimmerman, D., Bernabei, R., & Jonsson, P. V. (1997). Using the resident assessment instrument for quality enhancement in nursing homes. *Age and Ageing*, 26(Suppl. 2), 77-81.
- Porell, F., Caro, F. G., Silva, A., & Monane, M. (1998). A longitudinal analysis of nursing home outcomes. *Health Service Research* 33 (4 Pt 1), 835-865.
- Ramsay, J. D., Sainfort, F., & Zimmerman, D. (1995). An empirical test of the structure, process, and outcome quality paradigm using resident-based, nursing facility assessment data. *American Journal of Medical Quality*, 10, 63-75.
- Rantz, M. J., Petroski, G. F., Madsen, R. W., Mehr, D. R., Popejoy, L., Hicks, L. L., et al. (2000). Setting thresholds for quality indicators

- derived from MDS data for nursing home quality improvement reports: An update. *Joint Commission Journal on Quality Improvement*, 26, 101–110.
- Rosenthal, G. E., Baker, D. W., Norris, D. G., Way, L. E., Harper, D. L., & Snow, R. J. (2000). Relationships between in-hospital and 30-day standardized hospital mortality: Implications for profiling hospitals. *Health Service Research*, 34, 1449–1468.
- Rosenthal, G. E., Hammer, P. J., Way, L. E., Shipley, S. A., Donar, D., Wojtalay, B., et al. (1998). Using hospital performance data in quality improvement: The Cleveland Health Quality Choice experience. *Joint Commission Journal on Quality Improvement*, 24, 347–360.
- Scanlon, D. P., Darby, C., Rolph, E., & Doty, H. E. (2001). The role of performance measures for improving quality in managed care organizations. *Health Service Research*, 36, 619–641.
- Schneider, E. C., & Epstein, A. M. (1996). Influence of cardiac-surgery performance reports on referral practices and access to care. *New England Journal of Medicine*, 335, 251–256.
- Schneider, E. C., Riehl, V., Courte-Wienecke, S., Eddy, D. M., & Sennett, C. (1999). Enhancing performance measurement: NCQA's road map for a health information framework. National Committee for Quality Assurance. *Journal of the American Medical Association*, 282, 1.
- Shojania, K. G., & Wachter, R. M. (2000). Unreliability of physician "report cards" to assess cost and quality of care. *Journal of the American Medical Association*, 283, 52; discussion 53–54.
- Sloss, E. M., Solomon, D. H., Shekelle, P. G., Young, R. T., Saliba, D., MacLean, C. H., et al. (2000). Selecting target conditions for quality of care improvement in vulnerable older adults. *Journal of the American Geriatrics Society*, 48, 363–369.
- Teno, J. M., Weitzen, S., Wetle, T., & Mor, V. (2001). Persistent pain in nursing home residents. *Journal of the American Medical Association*, 285, 2081.
- Wennberg, J. E. (2000). Unreliability of physician "report cards" to assess cost and quality of care. *Journal of the American Medical Association*, 283, 52–53; discussion 53–54.
- Zimmerman, D. R., Karon, S. L., Arling, G., Clark, B. R., Collins, T., Ross, R., et al. (1995). Development and testing of nursing home quality indicators. *Health Care Finance Review* 16, 107–127.

Received May 14, 2002

Accepted October 16, 2002

Decision Editor: Laurence G. Branch, PhD