Benchmarking and quality in residential and nursing homes: lessons from the US

Vincent Mor*, J. Angelelli, D. Gifford, J. Morris and T. Moore

Brown University, Department of Community Health, and Center for Gerontology and Health Care Research Providence, Rhode Island, USA

SUMMARY

Background  Performance measurement and benchmarking are common concerns in the delivery of long term care. It is common to measure the performance of providers and to publicly report these data. This paper examines selected technical challenges facing those who design, implement and disseminate health care quality performance measures.

Method  Review of the application of measures of performance in the US nursing home sector.

Results  Using examples drawn from the skilled nursing home arena, problems ranging from data reliability and validity, the multi-dimensional nature of quality measures and selection bias as well as differential measurement abilities are discussed.

Conclusions  Benchmarking of performance is an inherently complex issue. However, to ensure that such comparisons are both fair and valid requires measures to be more technically sophisticated and sensitive to real changes attributable to changes in care. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS — quality assurance; nursing homes; performance measurement; standards of care

INTRODUCTION

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 in many countries, concerns about deteriorating quality of care have began to 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).

Performance measurement based upon a comparison of various indicators of quality is generally characterised as benchmarking. This can take the form of comparing one provider’s performance to another, to an average or to a level of performance based upon an externally stipulated standard. The choice of comparative approach reflects how well understood the indicator of interest; comparisons of providers to one another or a group average is often done when there is no established standard, or performance target, that can be generally expected to be met by ‘good’ providers. In the industrial quality control sector from whence health care and other service industries borrowed performance measurement, the observed production error rate can be readily compared to the minimum acceptable rate. As sources of systematic error in the production system are identified and corrected, the minimum acceptable rate may be lowered in concert with the ethos of continuous quality improvement. While this approach has been standard in many manufacturing industries, it has only emerged as an accepted way of thinking in
Quality measurement, performance monitoring and quality improvement is a constant refrain in the entire sector in the US (Cohn et al., 2000). Hospitals now 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 and Cleary, 1996; Rosenthal et al., 1998). Insurers, particularly managed care companies, are routinely compared on how well they ensure that preventive health services are delivered to their subscribers. The performance of surgeons and surgical groups is routinely monitored in terms of mortality and complication rates while ambulatory practices’ performance in holding down waiting times and providing patients’ with positive experiences is compared, particularly in highly competitive markets (Hannan et al., 1994; Marshall et al., 2000). Finally, nursing homes in the US are now being compared on numerous quality indicators derived from inspections as well as routinely collected clinical data (Phillips et al., 1997).

The origins of the current interest in quality comparisons are varied but perhaps the most influential was the publication of a series of papers documenting the huge geographic variation in practice patterns that were unrelated to the ‘outcomes’ experienced by the populations living in those areas (Wennberg et al., 1989). Wennberg and his colleagues at Dartmouth have continued to document this variation revealing its association to the availability of health care resources. When the observed variation in treatment patterns was considered in light of what was known to be ‘effective’ treatment, evidence based guidelines for selected treatment of common medical conditions became the criterion against which variation could be compared making it possible to compare rates of adherence to accepted clinical practice guidelines. The logical extension of such efforts has been the publication of the Dartmouth Atlas of Health Care and the recent paper which compares treatment patterns received by Medicare beneficiaries in all 50 states on 22 different quality indicators, most of which are predicated on adherence to accepted clinical practices (Wennberg, 1999; Jencks et al., 2000).

In the US there are numerous pressures stimulating the wholesale adoption of quality measurement and benchmarking. Regulators, particularly the Health Care Financing Administration (HCFA), the agency that is responsible for administering the Medicare programme which sets standards and finances the Medicaid programme of health care for the poor, have been aggressively pushing performance measurement. For example, much to the chagrin of hospitals throughout the country, nearly a decade ago HCFA published hospital specific mortality rates, adjusted for disease severity (Iezzoni, 1997). The hue and cry that accompanied this publication set the tone for future performance comparisons produced by state governments as well as insurers and advocacy organizations. The National Committee for Quality Assurance (NCQA) has promoted performance comparisons among managed care organizations across the country, making such comparative information available to interested parties on the internet (Schneider et al., 1999). Finally, health care providers convinced of their superiority in quality have also been pushing comparative measurement as a way of giving them a competitive advantage (Rantz et al., 1997). Thus, while initiated by a few providers and professionals convinced that this approach would improve care and reduce inappropriate practice variation by the end of the 20th century, performance measurement and benchmarking is well on its way to wholesale adoption by virtually all segments of the industry (US General Accounting Office, 1994).

Types of performance measures
Comparing health care providers can be done in a variety of ways depending upon the purpose and the availability of data. If ‘standards of care’ exist about which there is no ambiguity, as noted above, performance on these standards can be used as a benchmark against which one or many providers can be compared. In other instances measures of the processes of care used can be compared probabilistically since professional opinion and normative information would suggest that in most cases the treatment should be provided. Finally, the ‘outcomes’ that patients experience can be compared, albeit only rarely against an absolute standard. Here again a probabilistic comparative approach can be used. In the paragraphs below, examples of these measures and their application are briefly reviewed.

Standard of care measures have been used by all sectors of health care. For example, NCQA compares MCOs, providers of healthcare, on the basis of the proportion of women 50 and over who’ve had a mammogram in the last two years (Rogers, 1998) and the extent to which children enrolled as members receive ‘on time’ immunisations. Recent applications rating hospital care quality include things like
anti-coagulant administration at the time of emergency department admission with cardiac symptoms as well as the proportion of post acute myocardial infarction patients discharged with a prescription for an anti-coagulant or anti-platelet agent (Jencks et al., 2000; Wennberg, 1999). In the nursing home arena there is evidence that prescription of high dose, long acting benzodiazepines causes falls and so making sure they are not prescribed is currently a standard of care (Zimmerman et al., 1995).

Probabilistic process measures are often based upon a combination of research evidence and professional opinion and have been applied in numerous settings. For example, an axilllary node dissection to definitively stage women with early stage breast cancer is not always needed but is generally considered necessary to properly treat patients. The emergence of primary care sensitive hospitalisations or the rate of asthmatic children admitted to the emergency department are used to evaluate the comprehensiveness of the chronic disease management programme present in a paediatric practice (Bindman et al., 1995). The incidence of falls or hip fractures of nursing home residents is another example of an event that should probably not be reduced to zero but higher than average rates might signify a problem (Zimmerman et al., 1995). Finally, use of physical restraints, while it should be rare, might not always be avoided, but again outliers may be assumed to be managing behaviour problems inappropriately (Zinn et al., 1993).

Outcome based performance measures can themselves be sub-divided into various classes. There are prevalence measures such as the proportion of patients in untreated pain or surgical infection rates (Bernabei et al., 1998). While not normally thought of in this way, surveys of consumer satisfaction or patients and family members’ experiences of the care received, are also prevalence measures since these ‘readings’ of provider performance tend to be taken on a ‘snap-shot’ basis. Some measures of outcome truly reflect the incidence of selected conditions or events such as hospital or nursing home acquired infections, falls, emergency department visits or hospitalisation (Intrator et al., 1999; Berlowitz et al., 2000; Rosenthal et al., 2000). Obviously, mortality and survival time fall into this class of outcome measures. Finally, in some instances, changes in patient condition (perception/behaviour) reflect outcomes that require measuring patients’ condition on two separate occasions (Porell et al., 1998; Rosen et al., 1999).

While arbitrary, this classification scheme is useful in that it reflects different levels of effort and complexity associated with measuring provider’s performance.

Technical complications in executing performance measures

Although various forms of performance measurement and provider comparisons are published on almost a daily basis in the US, numerous complex technical problems remain which may undermine their validity. In the paragraphs below each of these is discussed using examples drawn from the realm of quality monitoring measures in the US nursing home industry. Obviously, examples could be readily drawn from virtually every other health care sector currently comparing providers on the basis of performance measures since they all suffer from the same kinds of generic problems.

Aggregating data to represent providers

All performance measures require systematic data characterising phenomena as different as the delivery of certain services to the experiences of all patients or a subset of them. In almost all instances these data must be aggregated from the micro level at which they are collected in order to be reported in a manner so as to characterise the performance of the provider of interest. This is because performance measures are based upon reported statistics which generally have an explicitly defined denominator and a particular numerator. Thus, a denominator might be the number of residents in a nursing facility on a given day and the numerator might be the number of residents in that denominator who were physically restrained. This instantaneous ‘point prevalence’ physical restraint rate might differ considerably from one predicated on a more dynamic, time sensitive numerator and denominator. For example, the proportion of person days on which residents are physically restrained might be a much more sensitive and meaningful performance measure than one based strictly on point prevalence. However, to capture it the denominator would have to be defined as the number of person days ‘at-risk’ (under observation) during a given calendar period (e.g. 3 months) and the numerator would be defined as the number of patient days on which a physical restraint was applied. While the data system necessary to generate a commonly defined and implemented point prevalence measure across multiple providers is complex, to have a system in place to
record daily person time would require an extraordinarily sophisticated information system. Without a uniformly implemented information system across all providers being compared, it is very likely that the information will not be collected comparably. Indeed, a recent validation study examined the accuracy of the information MCOs provided to the National Committee on Quality Assurance (NCQA) which then compared them on various performance measures. The results identified major problems with the consistency and reported reliability of the data obtained from numerous primary care groups and clinic settings and then aggregated up to reflect the ‘performance’ of the MCO whose health insurance subscribers were being served by that network (Kippen et al., 1997). Even without significant data consistency problems, comparing aggregated measures between providers or groups of providers ignores very significant intra-provider variation that might be very useful in understanding the consistency of practice within an organisation.

Small sample sizes

Regardless of the structure of the performance measure, in many instances the phenomena of interest are fairly rare, or should be. The number of patients who acquire nosocomial infections following particular types of surgery in a given hospital is likely to be small as is the number of acute myocardial infarctions. Similarly, the number of nursing home acquired pressure ulcers in a three month period is hopefully quite small (McClellan and Staiger, 1999; Berlowitz et al., 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. Table 1 presents standard error and confidence intervals around the estimate of a particular proportion based quality indicator assuming that the ‘true’ rate in the facility was 0.05.

As is evident, it is not until the number of observations exceeds 200 that the confidence intervals around the observed rate are less than twice its size. Depending upon the purpose of the performance measure comparisons, there are certainly circumstances in which using the 95% confidence interval (CI) results in estimates that are too conservative, but even using an 80% CI with a small number of observations means that potentially meaningful differences might be missed and potentially trivial differences might be highlighted. It is obviously possible to increase the duration of observation or the number of months or years of data included in the denominator or numerator in order to increase the stability of the measures. However, the more a performance measurement relies upon historical data, the less it may truly reflect the current performance of that provider. This increased stability can result in a provider that has overcome historical problems continuing to score poorly while another provider that has let standards slide may continue to look like the best for sometime after their shine has tarnished. Finding the right balance between stability and sensitivity is the challenge that still needs to be faced.

Complexity of risk adjustment

A great deal has been written about the importance of risk adjustment, often necessary in order to compare ‘apples to apples’ and not to oranges (Iezzoni, 1997). Many performance measures will be very sensitive to the kinds of patients served by the providers being compared. Diabetes, a known risk factor for acquiring pressure ulcers, is 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 penalising the provider with the reputation for excellence in pressure ulcer care (Berlowitz et al., 2000). While this kind of risk adjustment is fairer because it ‘levels the playing field’ without rewarding ‘skimming’ low-risk patients, it is possible to over adjust (Zimerman et al., 1995). Sticking with the pressure ulcer example, incontinence of urine and faeces is also a risk factor for acquiring a pressure ulcer, even stronger than diabetes. However, incontinence, like pressure ulcers, could arise as a result of the same underlying

Table 1. Relationship between sample size and the standard error of estimate for a hypothetical quality indicator with incidence of 0.05

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>Standard error</th>
<th>95% CI (Binomial exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.09</td>
<td>0.002-0.444</td>
</tr>
<tr>
<td>20</td>
<td>0.05</td>
<td>0.001-0.245</td>
</tr>
<tr>
<td>30</td>
<td>0.04</td>
<td>0.008-0.223</td>
</tr>
<tr>
<td>50</td>
<td>0.03</td>
<td>0.01-0.16</td>
</tr>
<tr>
<td>100</td>
<td>0.02</td>
<td>0.01-0.11</td>
</tr>
<tr>
<td>200</td>
<td>0.01</td>
<td>0.02-0.09</td>
</tr>
<tr>
<td>500</td>
<td>0.009</td>
<td>0.03-0.07</td>
</tr>
</tbody>
</table>

Copyright © 2003 John Wiley & Sons, Ltd.

Int J Geriatr Psychiatry 2003; 18: 258–266.
A phenomenon, e.g. poor nursing home care. Adjusting for incontinence when some residents’ (particularly long stay patients’) incontinence was caused by the facility, improves their observed performance in a way that might be unfair. Unfortunately, there are no easy answers to the whole issue of under or over adjustment and each performance measure must be carefully considered individually.

Whether to adjust for demographic factors or for clinical factors that might be empirically related to the event of interest, although it need not be, is another matter of considerable interest. For example, Table 2 presents selected information on the prevalence of pain among nursing home residents with cancer (Bernabei et al., 1998). As can be seen, older patients and minorities are less likely to be rated as being in daily pain, the former possibly for biological reasons, the latter perhaps due to inadequate assessment.

We also find that cognitively impaired patients are less likely to be assessed as in pain, presumably due to the difficulties inherent in assessing those individuals for the presence and severity of pain. Whether any or all of these factors should be adjusted for in calculating performance measures for long term care patients’ pain and its management is not obvious, but certainly is crucial.

### Ascertained bias

Related to the issue of adjustment is how to cope with measurement bias that is related to performance in a particular domain of quality. For example, we have seen 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 fairly homogeneous group of nursing home patients admitted with a diagnosis of cancer, most of whom do not survive six months (Bernabei et al., 1998). Table 3 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 due to the underlying biological differences in the patient populations. Rather, the standard and expectation for assessing the presence of pain among nursing home patients must be presumed to vary as a function of different approaches to training that have been instituted.

It is important to recognise that this problem of a differential in the completeness of the quality data is not restricted to nursing homes. Comparing the rates of blood pressure control among hypertensive patients in an ambulatory clinic population may be biased by the fact that some clinics may be far more systematic about the routine measurement of patients’ blood pressure. Furthermore, comparing the rates of anti-depressant use among depressed patients is obviously going to be very sensitive to whether there is a systematic effort to identify patients with depressive symptoms. For many different clinical signs and syndromes, if the system does not explicitly seek out the problem, the prevalence observed will be lower than in a system that does seek it out.

### Censoring observations related to longitudinal outcomes

In many instances the desired outcome of interest is a measure of patient’s change in status or the incidence of an untoward event. As noted earlier, these types of longitudinal measures require data about the patient at several points in time, or at least, require that a monitoring system is in place to determine whether an event of interest (e.g. re-hospitalisation) has occurred (Porell et al., 1998). Using administrative records such as Medicare claims is generally a relatively good

---

**Table 2.** Demographic and clinical factors associated with the presence of daily pain

<table>
<thead>
<tr>
<th></th>
<th>Daily pain (n = 2698)</th>
<th>No pain (n = 6725)</th>
<th>Odds ratio</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>85+ Years Age</td>
<td>785</td>
<td>2460</td>
<td>0.58</td>
<td>(0.51–0.67)</td>
</tr>
<tr>
<td>African–American</td>
<td>118</td>
<td>570</td>
<td>0.55</td>
<td>(0.44–0.68)</td>
</tr>
<tr>
<td>Other minority</td>
<td>74</td>
<td>298</td>
<td>0.74</td>
<td>(0.55–0.99)</td>
</tr>
<tr>
<td>Low cognitive</td>
<td>693</td>
<td>2451</td>
<td>0.54</td>
<td>(0.49–0.61)</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 3.** Percent of cancer patients in nursing homes in six states rated as having daily pain

<table>
<thead>
<tr>
<th>% Pain</th>
<th>Kansas</th>
<th>Maine</th>
<th>Miss</th>
<th>NY</th>
<th>Ohio</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Admits</td>
<td>39.5%</td>
<td>48.1%</td>
<td>23.3%</td>
<td>26.7%</td>
<td>26.2%</td>
<td>45.4%</td>
</tr>
<tr>
<td>July 1996</td>
<td>32.3</td>
<td>42.3</td>
<td>19.3</td>
<td>16.6</td>
<td>20.3</td>
<td>36.4</td>
</tr>
</tbody>
</table>

---

*Copyright © 2003 John Wiley & Sons, Ltd.*

*Int J Geriatr Psychiatry* 2003; 18: 258–266.
source for monitoring things like hospitalisation or mortality. However, increasingly in some parts of the US, older patients are joining MCOs that are not required to report event data in a form that is compatible with health service utilisation claims generated for all service use events in the ‘fee-for-service’ Medicare system. Thus, patients jumping from Medicare to managed care, or vice versa, may not have their outcome events observed and so are ‘censored’. Even for a quality measure as simple as having received the complete complement of childhood immunisations, families that switch insurance companies cannot be tracked to make sure that the children have, indeed had all the required immunisations, at least not from centralised computerised records. Insurance company switching compromises the completeness of ambulatory and preventive health practice quality indicators because the families leaving an insurer may be at greatest risk of not having had complete vaccination but they are censored since they do not contribute to the denominator of an indicator.

In the case of nursing homes, measures such as the incidence of pressure ulcers requires that patients are observed for a period. Those patients discharged to hospital or leaving the facility for another reason are censored and cannot be counted in the calculation of a rate. If some facilities are inclined to discharge patients as 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. Since 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.

**Conceptual issues underlying benchmarking that require resolution**

Not only are there myriad technical issues that need to be addressed in order to benchmark health care providers on the basis of quality, there remain important conceptual issues pertinent to how the benchmarks and comparisons will be used. Here we will address three of the varied conceptual issues that influence the meaning and utility of benchmarks. First, the multidimensionality of quality has implications for ‘ranking’ providers and for the feasibility of finding the ‘best’ provider. Second, the nature of the audience to whom performance measures and benchmarks are reported have enormous implications for how they are constructed and communicated. Finally, does it make a difference if benchmarks are negatively or positively framed, viewed as the extent to which a ‘gold standard’ is achieved or as a violation of a minimum standard?

**The multidimensionality of quality**

Research from the literature on hospital performance, health care plan performance and also the performance of nursing homes all suggests that the variety of different performance measures advocated as meaningfully measuring quality suggests that there is only a low level of correlation among the various measures of quality (Rosenthal et al., 2000; Sloss et al., 2000). That is, hospitals with low rates of mortality for patients presenting with an acute myocardial infarction have higher rates of some other type of undesirable outcome. The NCQA 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. Finally, our own research and that of others in the field reveals only a very low level of statistical evidence for consistent and coherent dimensions of quality that can usefully summarise the multiplicity of nursing home quality 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 fly in the face of the common expectation among consumers that there is such as thing as the best ‘overall’ provider. Since different individuals are likely to value different aspects of health care differently, it is unlikely that a single measure of quality will be suitable for all individuals, even though each individual might be very clear about her/his own preferences. However, although performance quality might be multi-dimensional, it would be chaotic to imagine different measures for different types of people. How we will resolve the conflict between a desire to identify the best provider with the fact of there being multiple dimensions of quality that are only weakly correlated is a major challenge.

**Positively versus negatively framed measures and benchmarks**

Many of the measures used to compare the performance of providers tend to focus on the prevalence of performance errors or problems (Zimmerman et al., 1995). This is most common in the nursing home industry where the incidence of pressure ulcers
or the prevalence of restraints of problematic drug prescription patterns are commonly applied indicators. On the other hand, many of the ambulatory measures are calibrated based upon ‘standards of care’ for which there is good evidence for the appropriateness of the standard or convincing expert opinion (Hofer et al., 1999; Garnick et al., 1994). These somewhat more positively framed measures carry the impetus as well as the instructions for how to improve since the measure is based upon a standard for which there is a rationale. On the other hand, the negatively framed measures can be important if violation of certain accepted quality standards such as justified implementation of restraints is clearly a problem.

**Different audiences for performance 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 (Frankenfeld et al., 1997). For example, providers can benefit, and indeed have been the staunchest advocates for using performance measures as the first step toward identifying care problems that can be addressed as a part of a continuous quality improvement programme. Hospitals throughout the country routinely seek to identify areas of performance that they deem to be important and in which they wish to improve (Rosenthal et al., 1998). In the nursing home field, the establishment of quality indicators based upon uniformly available data, no matter how technically compromised, has provided the basis for numerous different quality improvement programmes that are now reportedly being undertaken routinely in a large proportion of US nursing facilities (Castle, 1999). Since facilities are most interested in understanding how their performance in a particular area changes over time, many of the technical complications associated with using performance measures for comparative purposes become moot since the facility is only being compared to itself.

The HCFA, which is responsible for setting standards and regulating nursing homes in the US, has recently decided to pursue a new regulatory approach that relies upon the use of performance measures routinely calculated for each US facility from available clinical/administrative data (Hawryluk, 1999). 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 those aspects of performance that appear to be most deficient, based upon the performance measures.

In this instance, the performance data is used as a ‘starting point’ that may be then ‘validated’ based upon the inspection protocol which means that theoretically the quantitative performance measures could be contradicted upon closer inspection of actual patients and records. Thus, the performance measures continue to be treated as mere indicators of potential quality problems or successes, rather than being interpreted as *ipso facto* measures of quality.

Purchasers of health care such as insurance companies or even employers who pay for the health care of their employees have been urged to force providers to compete on both price and quality. Quality is measured based upon performance measures such as those promulgated by the NCQA. In this way providers that are consistently poor performers might be excluded from purchase of service contracts or not included as providers by a managed care company. HCFA, the biggest purchaser of health care is now considering how to encourage states which administer the Medicaid health insurance programme for the poor to begin applying prudent purchaser models that would emphasise purchasing based upon quality as well as price (Jencks et al., 2000). Unfortunately, there is little evidence that in spite of the increasing availability of performance measures, purchasers and employers are actually considering them in their decision as to 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 performance measure data so that they can actively select the providers and insurers they feel best meets their needs (Edgman-Levitan and Cleary, 1996). Several problems have been found in making these data available. First, consumers often find the resulting published information to be difficult to understand, absorb and act upon and secondly, only a very small minority of consumers and their families and advocates actually avail themselves of the information that is available. In spite of these problems, the vision of having informed consumers selecting providers continues to influence the design of performance measure reporting systems and the policies implicit in a system forced to compete on the basis of both cost and quality.

**Summary**

Benchmarking of health care providers is underway in virtually every sector of the health care industry in the US. 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 and Wachter, 2000; Wennberg, 2000). In the long term care arena, the universal adoption and computerisation of the nursing home resident assessment instrument (Minimum Data Set (Morris et al., 1995)) has catapulted this sector into the lead, in its potential for the measurement of meaningful outcomes, if not the reality. Just as HCFA published hospital specific mortality statistics over a decade ago, it too 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. Indeed, HCFA has articulated a vision statement about the application of performance measures, which derives from the recommendations of the Presidents’ Commission on Consumer Protection and Quality in Health Care Industry released in 1998. This vision commits HCFA to adopt consumer and purchaser driven indicators of quality using standardised data; i.e. a model consistent with that currently being implemented in the nursing home industry.

It is still too early to tell what the consequences of all this disclosure of data will mean, both in general and in light of the numerous technical and conceptual problems inherent in interpreting the resulting performance ratings. Some have argued that the publication of surgeon specific cardio-vascular surgery mortality rates contributed to the significant reduction in post-surgical mortality observed in the years following publication of these data in New York and Pennsylvania (Hannan, 1994; Hannan et al., 1995; Chassin et al., 1996). Others have noted that the observed changes were merely part of a pre-existing trend and that publication of the ‘score cards’ had few benefits (Schneider and Epstein, 1996; Epstein, 1998). To date there is no data on how the nursing home industry is responding to the publication and dissemination of this information, although several commentaries from the field suggest that industry leaders welcome the opportunity to demonstrate their quality superiority (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 where physicians feel that small samples, inadequate, or costly, data and selection bias all serve to undermine the perceived validity and fairness of any ‘report cards’ of physicians’ practice performance (Greene et al., 1996; Hofer et al., 1999; Shojania and Wachter, 2000). While 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 due to the increased specialisation of nursing homes in the US can compromise the ‘fairness’ of performance comparisons.

Nonetheless, in spite of 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 (Normand et al., 1997; McLellan and Staiger, 1999; McLellan and Staiger, 2000). Some of these new techniques will require new data and others will require policy makers to adopt an appropriate balance between the stability of a provider’s performance measure and the sensitivity of the measure to real changes brought about by improvements in care. This and similar issues related to the multi-dimensionality of quality will require considerably more thought as to their implications and how the users of the information understand it and use it. The entire health care industry is still in the relatively early stages of this revolution of accountability and this is even more true for nursing homes. It will be interesting to see how this movement evolves over the coming decades in the US and whether and how it will be exported to other countries concerned with these issues.

ACKNOWLEDGEMENTS

Supported in part by a MERIT grant from the National Institute on Aging (AG-11624) and contract from the Health Care Financing Administration to Abt Associates with a major subcontract to Brown University.

REFERENCES


Hannan EL, Siu AL, Kumar D, et al. 1995. Facility characteristics asso-
ciation with mortality among Medicare beneficiaries in Boston and 30-day standardized hospital mortality: implications for profiling hospitals. Health Serv Res 34(7): 1449–1468.

Hawryluk M. 1999. Quality indicators help focus survey. QIs to pin-

Hofer TP, Hayward RA, Greenfield S, et al. 1999. Functional sta-


