Nursing Home Performance

Improving Quality Assessment through Multilevel Modeling: The Case of Nursing Home Compare

Greg Arling, Teresa Lewis, Robert L. Kane, Christine Mueller, and Shannon Flood

Objective. To demonstrate how multilevel modeling and empirical Bayes (EB) estimates can improve Medicare's Nursing Home Compare quality measures (QMs).

Data Sources/Study Setting. Secondary data from July 1 to September 30, 2004. Facility-level QMs were estimated from minimum data set (MDS) assessments for approximately 31,000 Minnesota nursing home residents in 393 facilities.

Study Design. Prevalence and incidence rates for 12 nursing facility QMs (e.g., use of physical restraints, pressure sores, and weight loss) were estimated with EB methods and risk adjustment using a hierarchical general linear model. Three sets of rates were developed: Nursing Home Compare's current method, unadjusted EB rates, and risk-adjusted EB rates. Bayesian 90 percent credibility intervals (CIs) were constructed around EB rates, and these were used to flag facilities for potential quality of care problems.

Data Collection/Extraction Methods. MDS assessments were performed by nursing facility staff, transmitted electronically to the Minnesota Department of Health, and provided to the investigators.

Principal Findings. Facility rates and rankings for the 12 QMs differed substantially using the multilevel models compared with current methods. The EB estimated rates shrank considerably toward the population mean. Risk adjustment had a large impact on some QM rates and a more modest impact on others. When EB CIs were used to flag problem facilities, there was wide variation across QMs in the percentage of facilities flagged.

Conclusions. Multilevel modeling should be applied to Nursing Home Compare and more widely in other health care quality assessment systems.

Key Words. Nursing home, quality, risk adjustment, multilevel models

Many health care quality assessment systems rely on data from patient-level outcomes in order to draw inference about provider performance. Providers having higher rates of adverse outcomes (e.g., infections, pressure sores, or mortality) are assumed to be delivering poorer quality care. Despite the

hierarchical nature of these data (e.g., patients nested within health care facilities), there have been relatively few efforts to model health care quality explicitly from a multilevel perspective. Ignoring the multilevel nature of these data can lead to erroneous inferences about care quality. We apply multilevel modeling to a health care quality assessment system—Medicare's Nursing Home Compare—to show how it contrasts with conventional methods and leads to better estimates of care quality.

The Nursing Home Compare website is designed to provide consumers with quality-related information on the nation's approximately 17,000 nursing facilities to help them make an informed choice about nursing home care. The nursing home quality measures (QMs) are a major component of this system in addition to results from the facility's licensure and certification report. The 12 chronic care QMs reported on this website are expressed as facility rates of the incidence or prevalence of potentially problematic care processes (e.g., restraints or catheters) or untoward outcomes (e.g., mobility decline or urinary tract infections). The Nursing Home Compare reporting system compares each facility's QM rate to the state and national averages. Facilities with higher than average QM rates are presumed to offer poorer quality care.

The QMs could be of great benefit to consumers in helping to identify the best quality facilities, as well as to providers who could compare themselves against peers and identify areas for quality improvement. However, the current QM rates have fundamental problems (Mor, Angelelli et al. 2003; Mor, Berg et al. 2003; Zimmerman 2003; Arling et al. 2005). First, risk adjustment is used sparingly; by giving only minimal attention to differences in resident acuity or risk among facilities, the QMs may generate unfair or misleading facility comparisons. Facilities with residents at higher risk of poor outcomes (i.e., more medically unstable, functionally dependent, or cognitively impaired) are at a disadvantage when compared with facilities taking care of lower risk residents. Second, the QMs fail to adequately deal with

Address correspondence to Greg Arling, Ph.D., IU Center for Health Policy, School of Public and Environmental Affairs, IUPUI, 324 N Senate Ave., Indianapolis, IN 46204. Dr. Arling is with the Indiana University Center for Health Policy, School of Public and Environmental Affairs, Indiana University–Purdue University Indianapolis and with the Indiana University Center for Aging Research. Teresa Lewis, B.A., is with the Minnesota Department of Human Services, Nursing Facility Rates and Policy Division, St. Paul, MN and also with the Division of Health Policy and Management, School of Public Health, University of Minnesota, Minneapolis, MN. Robert L. Kane, M.D., and Shannon Flood, B.A., are with the Division of Health Policy and Management, School of Public Health, University of Minnesota, Minneapolis, MN. Christine Mueller, R.N., Ph.D., is with the School of Nursing, University of Minnesota, Minneapolis, MN.

estimation error. QM rates are presented in Nursing Home Compare with no information about their accuracy or precision. For example, no consideration is given to facility size even though QM estimates for small facilities are less precise and are more likely to take on extreme values than are estimates for large facilities. The General Accounting Office (2002) criticized the Nursing Home Compare QMs along these lines.

This study proposes a better method for estimating facility QM rates. We develop and test multilevel models that produce risk-adjusted empirical Bayes (EB) estimated QM rates with credibility intervals (CIs). Our objectives are to (1) describe the multilevel modeling approach and why it is particularly well-suited to this application; (2) describe a set of multilevel models and produce new fully risk-adjusted EB QM estimates and CIs; (3) assess outcomes from these models—change in facility QM rate, rank, and likelihood of being flagged as having quality problems; (4) draw conclusions about effectiveness of the current versus new QM rates in minimizing estimation error and discriminating between facilities in the quality of their care; and (5) consider the practical implications of this approach.

QUALITY MEASURES (QMs)

Twelve chronic care QMs were selected by the National Quality Forum from among a pool of nursing quality indicators (QIs) developed by researchers at the Center for Health Systems Research and Analysis (CHSRA) at the University of Wisconsin (Zimmerman et al. 1995) and by other researchers primarily at Abt Associates and Brown University (Morris et al. 2003). Table 1 lists the 12 QMs and their definitions. For the resident, the QM is a binary variable defined by the presence or absence of the care process or outcome. The facility QM rate is a proportion based on the number of residents experiencing a care process or outcome divided by the number of residents at risk. All residents in the facility may be at risk (no exclusions), or the QM may be calculated for only a subset of residents who meet certain conditions. Rates are calculated and reported each calendar quarter.

RISK ADJUSTMENT

Risk adjustment has received considerable attention in the general health services research literature (Iezzoni 2004; Stafford et al. 2004; Blumenthal

Table 1: Nursing Home Compare Chronic-Care QM Definitions and Study Risk Adjusters

QM	Definition	Risk Adjusters
Activities of daily living (ADL) decline	QM: Decline of two ADL items (bed mobility, transfer, eating, toileting) between assessments. Exclusions: Totally ADL dependent at prior, comatose, end-state disease, hospice	Cognitive Performance Scale (CPS), Alzheimer's disease, gender, age, length of stay (LOS)
Bedfast	QM: Bedfast on target assessment. Exclusions: Comatose	End-stage disease, gender, age, LOS
Increasing depression or anxiety	QM: Worsening Mood Scale between assessments. Exclusions: Mood Scale at maximum at prior, comatose	CPS, stroke, Alzheimer's disease, dementia other than Alzheimer's, gender, age, length of stay
Low-risk incontinence	QM: Bowel or bladder incontinence. Exclusions: Severely cognitively impaired, totally dependent in mobility ADLs (bed mobility, transfer, locomotion on unit), comatose, catheter, ostomy	Personal Severity Index (PSI) subset 1: diagnoses, [†] PSI subset 2: nondiagnoses, [†] Resource Utilization Groups (RUG) late loss ADL, [†] bed mobility, transfer, locomotion on unit, CPS, gender, age, LOS
Indwelling catheters	QM: Indwelling catheter. Exclusions: None	Bowel incontinence at prior,* pressure sores at prior,* stroke, paraplegia, quadriplegia, gender, age, LOS
Mobility decline	QM: Decline in locomotion on unit. Exclusions: Totally dependent in on-unit locomotion at prior, comatose, end-state disease, hospice	Recent falls,* eating assistance,* toileting assistance,* PSI subset 1: diagnoses, [†] toileting dependence, eating dependence, arthritis, paraplegia, hemiplegia, stroke, hip fracture, gender, age, LOS
Moderate or severe pain	QM: Moderate daily pain or any severe pain. Exclusions: None	Independence or modified independence in daily decision making,* CPS, [†] long term memory, [†] cancer, arthritis, hip fracture, gender, age, LOS
Physical restraints	QM: Trunk, limb, or chair restraints. Exclusions: None	Physically abusive, gender, age, LOS

High-risk pressure sores	<i>QM</i> : Stages 1–4 pressure sore and in high-risk group (impaired transfer or bed mobility, comatose, or malnourished). <i>Exclusions</i> : None	RUG Nursing Case Mix Index, [†] RUG Clinically Complex, [‡] PSI subset 1: diagnoses, [‡] end-stage disease, history of resolved ulcers, gender, age, LOS
Low-risk pressure sores	<i>QM</i> : Stages 1–4 pressure sore and not qualified for high-risk group. <i>Exclusions</i> : None	RUG Nursing Case Mix Index, [‡] end-stage disease, history of resolved ulcers, gender, age, LOS
Urinary tract infections	<i>QM</i> : Urinary tract infection. <i>Exclusions</i> : None	Locomotion on unit, stroke, paraplegia, quadriplegia, gender, age, LOS
Weight loss	<i>QM</i> : Weight loss. <i>Exclusions</i> : Hospice	PSI subset 1: diagnoses, CPS, cancer, gender, age, LOS

*Nursing Home Compare adjuster.

[†]Abt Associates adjuster; all other items are Kane adjusters.

QM, quality measure.

et al. 2005) and more specifically with nursing home QIs (Arling et al. 1997; Mukamel 1997; Kidder et al. 2002). On the one hand, there is general agreement that differences in resident acuity or risk between providers should be taken into account. On the other hand, risk adjustment if not appropriately carried out might “let providers off the hook” by adjusting away true differences in quality or setting lower levels of acceptable performance for certain resident groups. After much debate, members of the National Quality Forum took a conservative approach by selecting QMs for Nursing Home Compare that were only minimally adjusted (National Quality Forum 2002).

Table 1 shows QM risk adjustment methods and variables. Exclusion removes residents from the QM denominator if they have particular characteristics (e.g., end-stage disease or coma). Half of the QMs have residents excluded from calculations. The second method, stratification, divides residents into risk groups for a given QM and reports their rates separately. Exclusion and stratification are simple to perform and understand. However, their ease of use involves a tradeoff, as they are more prone to measurement error when risk categories are small or when high or low risk residents are concentrated in certain facilities (Mor, Angelelli et al. 2003; Mor, Berg et al. 2003). Two QMs (prevalence of pressure sores and of incontinence) are stratified by risk group, although high-risk incontinence is not reported on the Nursing Home Compare website.

Statistical adjustment through logistic regression, a third method, is applied to three Nursing Home Compare QMs: indwelling catheter, mobility decline, and pain. The probability of a resident triggering a QM (i.e., having a care process or experiencing an outcome) is modeled as a function of the resident’s individual characteristics (Berg et al. 2002). Individual QM probabilities are then averaged across the facility and compared with the observed QM rate. If the facility’s predicted rate is lower than its observed rate, its estimated rate will be adjusted upward. If its predicted rate is higher than observed its estimated rate will be adjusted downward. A multiple regression method allows many risk factors to be considered simultaneously and it arrays residents on a continuum from low to high risk rather than assigning them to very broad categories. On the other hand, statistical adjustment is the most complex of the three methods and it may be difficult for stakeholders to understand (Zimmerman 2003). In addition, a conventional regression model does not deal effectively with problems of estimation error or account for clustering of residents in facilities (Normand, Glickman, and Gatsonis 1997).

ESTIMATION ERROR

Estimation error is a well-recognized problem in health care quality measurement systems (Iezzoni 1994; Greenland, Schwartzbaum, and Finkle 2000). Reporting averages without accompanying confidence or prediction intervals can seriously disadvantage some providers, particularly when they vary in size or when their performance falls within a relatively narrow range (Marshall and Spiegelhalter 2001; Ash, Shwartz, and Pekoz 2003). Although this problem has been recognized for nursing home QIs (Arling et al. 2005), very little has been done about it. Even among moderate-sized facilities, confidence intervals around QI rates can be quite wide (Mor, Angelelli et al. 2003; Mor, Berg et al. 2003). Yet, none of the QI systems (e.g., CHSRA, Abt-Brown, or Nursing Home Compare) report confidence intervals. Nursing Home Compare deals with estimation error by failing to report chronic care QMs for facilities with fewer than 30 residents in the QM denominator. This method results in missing quality information for many small facilities and it ignores the range of estimation error among larger facilities.

MULTILEVEL MODELS AND EB ESTIMATES

When data are collected and analyzed at more than one level (e.g., resident and facility) multilevel models (e.g., HLM) are more appropriate for risk adjustment and estimation than conventional methods, e.g., OLS regression or logistic regression (Greenland 2000; Raudenbush and Bryk 2002). Conventional approaches allow two main choices: aggregate QMs to the facility level, thereby losing information within facilities about relationships between risk factors and quality; or analyze the data entirely at the resident level, making it difficult to draw conclusions about quality differences between facilities. Multilevel modeling, on the other hand, recognizes explicitly that residents are nested in facilities with sources of variation at both levels, i.e., among residents within each facility as well as between facilities. Multilevel models are more likely than conventional methods to produce standard errors with the appropriate degrees of freedom for independent variables at each level.

Multilevel models also provide a better framework for estimation than single level models. Multilevel models can produce EB estimates which generally are more accurate than estimates based on classical or “frequentist” assumptions (Greenland 2000). EB estimates combine Bayesian with classical

approaches. Bayesian estimates rely on a prior distribution describing the investigator's subjective beliefs and the uncertainty about those beliefs before new data are collected. In EB estimates, an empirical distribution serves as the prior. After new data are collected, referred to as the "Likelihood," the prior distribution is revised in light of new findings to arrive at a posterior distribution reflecting new information gained from gathering additional data. A set of EB estimation intervals, referred to as "CIs," can be constructed around the EB estimates from the posterior distribution. CIs will typically be narrower than traditional confidence intervals due to the additional information provided by the EB prior. The Bayesian CI also allows us to make a statement about the probability that the true population parameter lies within the interval. For example, with a 90 percent CI we can say that there is a 90 percent probability that the true facility QM rate lies within the interval.¹

We have employed multilevel models in producing EB estimates for facility QMs. The QM distribution in the total nursing home population serves as the empirical prior, the QM distribution in a particular facility is the Likelihood, and the EB estimate is based on the posterior distribution. Before observing QM data for a facility our best guess of its QM rate would be the population average. After collecting data and observing the facility's QM rate (Likelihood) we revise our prior to arrive at the posterior QM estimate. The influence of the facility's observed QM rate on the posterior estimate will depend on the size of the facility and the amount of QM variation within and between facilities. The QM rates in larger facilities will be more certain (e.g., have lower standard errors) than in smaller facilities and, thus, will have greater weight or influence on the overall posterior (EB) estimate. Also, QMs with less variation between facilities have a more certain empirical prior (population average QM rate), which then has a greater influence on the posterior. As the prior tends to pull the posterior estimate toward the population mean, EB estimates are referred to as "shrinkage" estimates. The QM rates for smaller facilities that are quite distant from the population average typically experience the most shrinkage, and these facilities are less likely to be identified as outliers. This is reasonable because of the greater uncertainty associated with their estimates. Although the EB point estimate (EB QM rate) has a conservative bias, it is more precise (smaller standard error) than a conventional estimate resulting in greater overall accuracy, i.e., lower mean squared error (Greenland 2000).

Risk adjustment has a further effect on EB estimates. If the adjustment model is properly specified, i.e., it includes relevant risk factors that are

independent of care quality, the model can result in a more informative prior than a model without adjusters. Raudenbush and Bryk (2002) caution, however, that poorly specified adjustment models can lead to biased EB estimates posing a threat to the validity of inferences about quality or other indicators of performance. We return to the issue of bias in the discussion section.

METHODS

Data Source, QM Construction, and Risk Adjustment

Analyses were conducted using minimum data set (MDS) information on all 31,000 residents in Minnesota's 393 Medicaid-certified nursing facilities for the third calendar quarter of 2004. The MDS is an assessment mandated for all Medicare and Medicaid-certified facilities (Morris et al. 1990), and is completed for residents at admission, every calendar quarter (90 days) thereafter, and in case of significant change in health status. We followed the same method as Nursing Home Compare in constructing QMs. Prevalence QMs were based on the resident's target assessment, i.e., annual, significant change, or quarterly MDS assessment closest to the end of the calendar quarter. Incidence QMs were based on the target assessment and a prior assessment (annual, significant change, quarterly, or admission) conducted approximately 90 days before the target assessment. Nursing Home Compare standardizes the three statistically risk-adjusted QMs according to a national average QM rate. We employed the state rate for standardization; in all other respects we duplicated the Nursing Home Compare method for calculating the currently reported QMs. For the multilevel models with full risk adjustment, we used additional resident-level MDS-based adjusters proposed by Abt Associates (Kidder et al. 2002) and developed in our own work on QIs (Kane et al. 2004, Table 1). Both sets of adjusters were chosen because of their ability to predict risk and the minimal chance for nursing home causation or intervention.

Estimation Methods

We compared three methods of estimating rates for the 12 chronic care QMs: (1) currently reported Nursing Home Compare QM rates; (2) multilevel EB QM rates and CIs without additional risk adjustment; and (3) multilevel EB QM rates and CIs using full risk adjustment. Multilevel modeling was performed with *HLM* 6.02 statistical software (Raudenbush, Bryk, and Congdon 2004). Because QMs are binary outcomes, analysis was based on a hierarch-

ical general linear model, assuming a Bernoulli sampling distribution with a logit link function and a linear structural model. Random intercept models were developed and tested with resident-level risk adjuster variables and facility treated as a random effect. We tested for variation in slopes of risk adjuster variables but found insufficient evidence for random slope models, which can be quite complex and difficult to estimate.²

We dropped a facility from a QM rate calculation if it had fewer than 20 residents in the denominator, especially affecting three of the QMs as described below. This contrasts with Nursing Home Compare's threshold of <30 residents. We selected a lower threshold (<20) because of the capabilities of the EB method to contend with estimation error. Facilities dropped from QM calculations tended to be smaller (fewer total residents); however, they did not differ from other facilities in terms of ownership, location, or other characteristics.

RESULTS

We first describe sample means and standard deviations (SD) for the facility QM rates. Next, we examine absolute and percentage differences in facility QM rates and ranks between methods. Third, we compare alternative ways of flagging facilities for potential quality problems. Facilities were ranked from low to high according to their QM rates with each method. Having a QM rate at the 90th percentile or higher is used to flag facilities with rates based on the current method, whereas an EB CI with a lower limit above the population mean is used for the unadjusted and adjusted EB estimates. Finally, we illustrate the effects of EB shrinkage and risk adjustment through scatter diagrams.

Facility QM Rates with Alternative Methods

Alternative QM rates were calculated for each of the 393 facilities having at least 20 eligible residents during the calendar quarter. Table 2 shows the number of facilities, median number of residents, and means and SD for facility QM rates. Almost all facilities met the >20 cutoff for nine of the QMs. However, only 83 percent (325 facilities) met the >20 cutoff for the depression/anxiety QM, 67 percent (265 facilities) met the cutoff for the moderate/severe pain QM, and 65 percent (257 facilities) met the cutoff for the low-risk pressure sore QM. These QMs had correspondingly low numbers of residents in the QM denominators, with medians ranging from 29 to 36 residents. Facility medians for other QM denominators ranged from 41 to 64 residents.

Table 2: Facility Average Reported, Empirical Bayes (EB), and EB-Adjusted QM Rates

Quality Measure	Residents in QM Denominator		Reported QM Rate		EB QM Rate		EB-Adjusted QM Rate	
	Facilities with 20+ Residents	Median Residents/Facility	Mean	SD	Mean	SD	Mean	SD
ADL decline	393	55	0.165	0.076	0.160	0.046	0.158	0.045
Bedfast	393	64	0.012	0.017	0.009	0.006	0.008	0.005
Increasing depression/anxiety	325	36	0.133	0.086	0.126	0.047	0.117	0.045
Bladder or bowel incontinence	388	48	0.486	0.135	0.486	0.100	0.477	0.112
Indwelling catheter	393	64	0.059	0.039	0.053	0.013	0.047	0.012
Mobility decline	390	46	0.157	0.076	0.149	0.040	0.147	0.035
Moderate/severe pain	265	29	0.272	0.147	0.261	0.108	0.257	0.104
Physical restraints	393	64	0.045	0.051	0.041	0.037	0.039	0.037
High-risk pressure sores	365	41	0.088	0.058	0.085	0.025	0.070	0.017
Low-risk pressure sores	257	32	0.025	0.030	0.021	0.006	0.017	0.003
Urinary tract infections	393	64	0.072	0.042	0.069	0.016	0.065	0.014
Weight loss	393	60	0.081	0.046	0.079	0.018	0.068	0.014

QM, quality measure; SD, standard deviations; ADL, activities of daily living.

Average reported QM rates ranged from 0.01 (bedfast) to 0.49 (low-risk incontinence). Applying the EB method resulted in a small change in mean facility rates. However, the facility SDs for most QMs decreased considerably, which is an indication of EB shrinkage. The EB method with risk adjustment resulted in slight changes in means and modest changes in SDs. In an analysis available from the authors on request, we compared coefficients of variation (CV), or the SD/mean values, for the three adjustment methods. The CVs for reported rates ranged widely from 1.40 for the bedfast QM to 0.28 for the low-risk incontinence QM. Using the EB method with adjustment produced CV reductions ranging from 15 to 82 percent. The greatest shrinkage occurred for the bedfast (CV change from 1.40 to 0.59), high-risk pressure sore (CV change from 0.66 to 0.25) and low-risk pressure sore (CV change from 1.19 to 0.21) QMs.

Changes in Facility Rates and Ranks

We also assessed effects of alternative estimation methods by calculating the absolute value of the difference in facility rates and ranks. The numbers reported in Table 3 indicate how much the average facility's rate and ranking shifted when EB estimation and risk adjustment were applied. Compared with the reported method, the EB estimation method produced the greatest rate and ranking changes for the low-risk pressure sore (77 percent mean absolute difference in rates and 18 mean absolute difference in ranks), bedfast (74 percent rate and 35 rank change), and indwelling catheter (34 percent rate and 20 rank change) QMs. The EB method with risk adjustment led to additional rate and rank changes for all QMs, with the low-risk pressure sore (84 percent rate and 25 rank change), bedfast (79 percent rate and 36 rank change), high-risk pressure sore (39 percent rate and 32 rank change), and low-risk incontinence (16 percent rate and 66 rank change) QMs most affected.

Flagging Problem Facilities

Three methods were used to flag facilities for potential quality problems: reported QM rate at the 90th percentile or higher; EB 90 percent CI exceeding the population mean QM rate; and risk-adjusted EB 90 percent CI exceeding the population mean QM rate. The Nursing Home Compare website does not flag outlier facilities; however, other applications such as the CHSRA QIs have adopted a 90th percentile threshold. We chose a 90 percent CI rather than 95 percent so that our QM thresholds would have greater sensitivity, helping to balance the conservative bias inherent in the EB estimates.

Table 3: Differences between Reported, EB, and EB-Adjusted Facility QM Rates and Ranks

Quality Measure	Mean Reported Rate	Reported versus EB			Reported versus EB-Adjusted		
		Mean Abs Rate Difference	Mean Abs % Difference	Mean Abs Rank Difference	Mean Abs Rate Difference	Mean Abs % Difference	Mean Abs Rank Difference
ADL decline	0.165	0.023	14	8.8	0.026	16	24.2
Bedfast	0.012	0.009	74	34.5	0.009	79	36.1
Increasing depression/anxiety	0.133	0.030	23	8.3	0.034	25	17.4
Bladder or bowel incontinence	0.486	0.027	6	6.0	0.077	16	66.2
Indwelling catheter	0.059	0.020	3	20.0	0.022	38	31.4
Mobility decline	0.157	0.029	19	10.7	0.034	22	22.4
Moderate/severe pain	0.272	0.034	12	6.1	0.037	14	17.4
Physical restraints	0.045	0.012	26	14.4	0.013	29	18.0
High-risk pressure sores	0.088	0.026	30	13.7	0.034	39	32.3
Low-risk pressure sores	0.025	0.019	77	18.3	0.021	84	24.6
Urinary tract infections	0.072	0.021	28	16.8	0.023	31	25.0
Weight loss	0.081	0.022	27	15.8	0.027	34	39.5

EB, empirical Bayes; Abs, absolute; ADL, activities of daily living.

Table 4 shows the number of facilities flagged with each individual method and the agreement in flagging between methods based on the κ statistic. The 90th percentile rule flagged more facilities than either EB approach with the exception of the incontinence, pain, and physical restraints QMs, where EB methods flagged considerably more facilities. Only the activities of daily living (ADL) decline and depression/anxiety QMs displayed a degree of consistency; 25–35 facilities flagged on these QMs regardless of method. In contrast, there was little difference between the two EB methods, although risk adjustment tended to flag fewer facilities. The risk-adjusted EB CIs flagged 87 facilities on the restraints QM and 59 facilities on the low-risk incontinence QM, whereas fewer than 20 facilities (5 percent) were flagged on the bedfast, indwelling catheter, high- and low-risk pressure sore, UTI, and weight loss QMs. The κ statistics indicate low levels of agreement between the reported and EB flagged facilities, particularly the bedfast, indwelling catheter, low-risk pressure sore, and UTI QMs. The κ 's for the reported compared with the EB-adjusted flags were generally lower than for the EB method alone. The κ for the incontinence QM had a steep drop from 0.64 to 0.30. Finally, there was a relatively high level of agreement in flagging between the EB CIs alone and with risk adjustment. This suggests that outlier status was more sensitive to estimation error than to differences in facility risk characteristics.³

Illustration of Key Concepts

We chose the incontinence QM to illustrate the key concepts: EB shrinkage, risk adjustment effects, and CIs. Figure 1 shows facilities by their size and the difference between reported and EB incontinence rates. Smaller facilities with fewer residents in the QM denominator were much more likely to experience shrinkage toward the population mean. Figure 2 presents a scatter diagram of facilities by reported and EB-adjusted rates for the incontinence QM. Two patterns are evident in this figure. First, facilities with reported rates of incontinence below the population average (0.47) tend to shrink upward, while facilities with rates above the mean tend to shrink downward. Second, risk adjustment has a strong effect on rates for facilities at all levels of reported incontinence. Finally, Figure 3 illustrates the EB-adjusted rates and CIs for a 5 percent random sample of nursing facilities. This graph presents a visual image of estimation error (size of the CI) while pointing to those facilities with EB estimates significantly above (210, 258, 392, 397, and 453) or below the population average (34, 109, 280, 297, 307, and 393).

Table 4: Number of Facilities Flagged and Agreement between Adjustment Methods

Quality Measure	Total N	Number of Facilities Flagged				Agreement between Methods (κ)		
		Reported Flag	EB Flag	EB-Adjusted Flag	Reported and EB	Reported and EB-Adjusted	EB and EB-Adjusted	
ADL decline	393	40	39	38	0.73	0.64	0.85	
Bedfast	393	40	19	17	0.46	0.53	0.79	
Increasing depression/anxiety	325	33	36	36	0.75	0.77	0.81	
Bladder or bowel incontinence	388	39	63	59	0.64	0.30	0.47	
Indwelling catheter	393	40	16	18	0.43	0.37	0.82	
Mobility decline	390	40	28	20	0.64	0.53	0.77	
Moderate/severe pain	265	27	48	40	0.67	0.65	0.87	
Physical restraints	393	39	85	87	0.64	0.54	0.89	
High-risk pressure sores	365	37	19	16	0.58	0.42	0.79	
Low-risk pressure sores	257	27	4	0	0.06	–	–	
Urinary tract infections	393	38	15	13	0.46	0.44	0.85	
Weight loss	393	40	21	19	0.55	0.55	0.89	

κ statistics could not be computed for QMs with no agreement between methods.

EB, empirical Bayes; QM, quality measure; ADL, activities of daily living.

Figure 1: Facilities by Difference in Incontinence Rate (Empirical Bayes [EB] Rate—Reported Rate) and Number of Residents

DISCUSSION

Multilevel modeling resulted in much different nursing home QM rates than those reported in the current Nursing Home Compare system. Current QM rates are likely to be inaccurate and misleading because they fail to take into account estimation error and involve only limited risk adjustment. The EB rates from the multilevel models deal effectively with estimation error and include CIs, which allow stronger inference about care quality. Extensive risk adjustment resulted in significant changes in rates for some QMs, e.g., incontinence, but had a less pronounced effect on others. We recommend that multilevel modeling techniques be applied to Nursing Home Compare and other nursing home QI systems. The EB estimated rates should be reported for all QMs, either as a substitute for or alongside current rates, there should be a lower facility cutoff (e.g., <20 residents), and CIs should be reported for each rate. Furthermore, QMs that have very high estimation error (e.g., prevalence of low risk pressure sores) should be dropped from the reporting system or treated as sentinel events.

Figure 2: Facilities by Reported and Empirical Bayes (EB)-Adjusted Rate of Incontinence

We have no “gold standard” for validating our risk adjustment method or EB estimates. Instead, our conclusions must be based on theoretical grounds, primarily the well-established fields of estimation and multilevel modeling theory, as well as on sound clinical evidence for the selection of risk adjusters. Second, our results are derived from the nursing home population of a single state. However, we have no reason to believe that the basic pattern of results would differ significantly from other states or the nation. Rather, this methodology could be applied nationally, regionally, or within individual states. However, our study raises several additional issues that must be addressed when applying multilevel models to EB estimation and risk adjustment.

Conservative Bias

The EB approach involves tradeoffs between sensitivity and specificity. While it corrects for estimation error due primarily to variation in facility size, it also

Figure 3: Facilities by Empirical Bayes (EB)-Adjusted Rates and 90 percent Credibility Intervals for Prevalence of Incontinence (Random 5 Percent Facility Sample)

introduces a conservative bias where rates for small outlier facilities, in particular, are shrunk toward the population mean. The greater specificity of the EB rates may be achieved at the cost of letting some facilities “off the hook.” Smaller facilities delivering care that is truly of very low quality could be “false negatives” with EB-adjusted rates that look much better than their observed rates. Conversely, smaller facilities delivering truly high-quality care could have EB-adjusted rates that make them appear mediocre. The validity of the EB estimates could be improved by increasing the number of data points, i.e., tracking QM rates over calendar quarters, or by looking for similar patterns across related QMs. One must also keep in mind that the QMs are general indicators of care quality that should be followed up with in-depth investigation.

Risk Adjustment

Extensive risk adjustment could also address provider concerns about the special populations they are caring for. However, risk adjustment should

proceed with caution because adjustment adds another layer of complexity to the interpretation of QM rates, and it can threaten the validity of the QMs if estimates are conditioned on poorly specified risk adjustment models. First, risk adjusters should be carefully evaluated to ensure that they are clinically relevant, outside provider control and only minimally influenced by care quality (Zimmerman 2003). For example, catheters put residents at risk for UTIs and weight loss puts them at risk for pressure sores or ADL decline, yet neither variable would be an appropriate risk adjuster. Second, adjustment should make a substantive difference compared with unadjusted rates. For example, adjustment has a large impact on facility EB rates for the incontinence QM but only a minimal impact on the mobility decline QM.

Flagging Problem Facilities

The EB CIs offer a powerful and statistically sound basis for flagging problem facilities. One or two residents having an adverse outcome might place an otherwise average facility above the 90th percentile threshold on a low prevalence QM (e.g., weight loss or pressure sores). The EB approach with CIs (Figure 3) allows for a more thoughtful evaluation of a facility's QM rate in relation to its estimation error. In addition, when compared with 90th percentile flags, the EB CI flags led to very different and we would argue more statistically valid results, further calling into question the percentile method. The large numbers of facilities with EB CI flags on the low-risk incontinence and restraints QMs may reflect the pervasiveness of these care problems in the nursing home population. The low-risk incontinence QM may be particularly useful as a regulatory and quality improvement tool as the Centers for Medicare and Medicaid Services undertakes its new initiative on care of incontinent residents. In contrast, the pressure sore QMs for low- and high-risk populations had few facilities flagged with EB CIs, suggesting that the conventional 90th percentile flag for these QMs could be very misleading.

QM Reporting

Can EB-adjusted QM rates and CIs be presented in a way that stakeholders of diverse backgrounds can understand and find useful? The methodology is arguably opaque; however, the gains in accuracy, reliability, and inclusiveness should outweigh the reduced transparency (Mor, Angelelli et al. 2003; Mor, Berg et al. 2003). Improved graphics may help stakeholders grasp key concepts. For example, Figure 3 quickly identifies facilities performing better or worse than expected, while the width of the CIs shows the reliability of each facility's

estimate. The EB estimates with CIs also lend themselves to regulation or facility quality improvement because they involve explicit criteria for targeting quality problems and tracking quality over time.

Future Research

We have recommended elsewhere a thorough reengineering of the QMs with an evolution from the current QMs to new QMs that are more comprehensive, rigorously tested, and that subject themselves to continuous quality improvement (Arling et al. 2005). Future research should explore alternative model specifications and EB estimation methods for QMs. Sensitivity analysis should be performed on the effects of risk adjustment. Estimation methods based on statistical sampling or simulation, such as Monte Carlo Markov Chains, involve fewer distributional assumptions and may produce better estimates than analytical methods such as EM Laplace (Hox 2002). Estimation might also be improved with more formal Bayesian approaches involving a subjective prior, based on clinical judgment or explicit utility functions, either alone or in combination with an empirically derived EB prior (Greenland 2000; Spiegelhalter, Abrams, and Myles 2004). For example, Berlowitz et al. (2002) used Bayesian hierarchical modeling to risk adjust pressure sore development in nursing homes, selecting as prior a threshold rate they judged would indicate a care problem. Finally, multilevel analysis can be applied in studying facility-level contextual variables (e.g., facility acuity, staffing, or management practices) that operate alone or in interaction with resident characteristics to affect care outcomes. By understanding the organizational context of care quality we will be better able to improve it.

ACKNOWLEDGMENTS

This study was supported by a grant from the National Institute on Aging (1 R01 AG021985) and by a contract with the Minnesota Department of Human Services. Two anonymous reviewers and the editors of *HSR* offered very thoughtful comments. The authors, however, are solely responsible for any errors or omissions as well as the opinions expressed.

NOTES

1. The conventional confidence interval based on "frequentist" assumptions is less intuitive. No statement can be made about the probability of the true parameter

falling into a particular confidence interval. Instead, we must say that in a long series of confidence intervals, constructed from independent samples, the intervals will contain the true parameter 90 percent of the time (Spiegelhalter, Abrams, and Myles 2004).

2. Estimation was performed with the expectation–maximization Laplace method in *HLM* 6.02, which produces accurate approximation to maximum likelihood estimates, particularly in correcting bias that arises when variance is large and probability of outcomes is small, two conditions present in many QMs. The *HLM* 6.02 software also produces EB residuals and posterior variances, which we used to compute the multilevel EB facility rates and credibility intervals. A technical description of HGLM estimation is available in Raudenbush and Bryk (2002, chapter 14) and Raudenbush, Bryk, and Congdon (2004, chapters 5 and 6).
3. When we compared these flagging results with a parallel analysis employing the Nursing Home Compare cutoff of <30 residents, we found that the number of facilities with reportable QMs remained high (>380 facilities) for six QMs; however, it declined to 326 for depression/anxiety, 328 for low-risk incontinence, 331 for mobility decline, 128 for moderate/severe pain, 279 for high-risk pressure sore, and 145 for low-risk pressure sore. Facilities flagging at the 90 percent EB CIs declined proportionately.

REFERENCES

- Arling, G., R. L. Kane, T. Lewis, and C. Mueller. 2005. "Future Development of Nursing Home Quality Indicators." *Gerontologist* 45 (2): 147–56.
- Arling, G., S. L. Karon, F. Sainfort, D. R. Zimmerman, and R. Ross. 1997. "Risk Adjustment of Nursing Home Quality Indicators." *Gerontologist* 37 (6): 757–66.
- Ash, A. S., M. Shwartz, and E. A. Pekoz. 2003. "Comparing Outcomes across Providers." In *Risk Adjustment for Measuring Health Care Outcomes*, edited by L. I. Iezzoni, pp. 297–333. Ann Arbor, MI: Health Administration Press.
- Berg, K., V. Mor, J. Morris, K. M. Murphy, T. Moore, and Y. Harris. 2002. "Identification and Evaluation of Existing Nursing Homes Quality Indicators." *Health Care Financing Review* 23 (4): 19–36.
- Berlowitz, D. R., C. L. Christiansen, G. H. Brandeis, A. S. Ash, B. Kader, J. N. Morris, and M. A. Moskowitz. 2002. "Profiling Nursing Homes Using Bayesian Hierarchical Modeling." *Journal of the American Geriatrics Society* 50 (6): 1126–30.
- Blumenthal, D., J. S. Weissman, M. Wachterman, E. Weil, R. S. Stafford, J. M. Perrin, T. G. Ferris, K. Kuhlthau, R. Kaushal, and L. I. Iezzoni. 2005. "The Who, What, and Why of Risk Adjustment: A Technology on the Cusp of Adoption." *Journal of Health Politics, Policy and Law* 30 (3): 453–73.
- General Accounting Office. 2002. *Public Reporting of Quality Indicators Has Merit, but National Implementation Is Premature* (No. GAO-03-187). Washington, DC: United States General Accounting Office.
- Greenland, S. 2000. "Principles of Multilevel Modeling." *International Journal of Epidemiology* 29 (1): 158–67.

- Greenland, S., J. A. Schwartzbaum, and W. D. Finkle. 2000. "Problems Due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis." *American Journal of Epidemiology* 151 (5): 531–9.
- Hox, J. J. 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Iezzoni, L. I. 1994. "Using Risk-Adjusted Outcomes to Assess Clinical Practice: An Overview of Issues Pertaining to Risk Adjustment." *Annals of Thoracic Surgery* 58 (6): 1822–6.
- . 2004. "Risk Adjusting Rehabilitation Outcomes: An Overview of Methodologic Issues." *American Journal of Physical Medicine and Rehabilitation* 83 (4): 316–26.
- Kane, R. L., S. Flood, B. Bershadsky, and G. Keckhafer. 2004. "Effect of an Innovative Medicare Managed Care Program on the Quality of Care for Nursing Home Residents." *Gerontologist* 44 (1): 95–103.
- Kidder, D., M. Rennison, H. Goldberg, D. Warner, B. Bell, L. Hadden, J. Morris, R. Jones, and V. Mor. 2002. "MegaQI Covariate Analysis and Recommendations: Identification and Evaluation of Existing Quality Indicators That Are Appropriate for Use in Long-Term Care Settings. CMS Final Report" [accessed on May 20, 2005]. Available at www.cms.hhs.gov/
- Marshall, E. C., and D. J. Spiegelhalter. 2001. "Institutional Performance." In *Multilevel Modeling of Health Statistics*, edited by A. H. Leyland and H. Goldstein, pp. 727–42. Chichester, UK: John Wiley & Sons.
- Mor, V., J. Angelelli, D. Gifford, J. Morris, and T. Moore. 2003. "Benchmarking and Quality in Residential and Nursing Homes: Lessons from the US." *International Journal of Geriatric Psychiatry* 18 (3): 258–66.
- Mor, V., K. Berg, J. Angelelli, D. Gifford, J. Morris, and T. Moore. 2003. "The Quality of Quality Measurement in US Nursing Homes." *Gerontologist* 43 (Special No. 2): 37–46.
- Morris, J. N., C. Hawes, B. E. Fries, C. D. Phillips, V. Mor, S. Katz, K. Murphy, M. L. Drugovich, and A. S. Friedlob. 1990. "Designing the National Resident Assessment Instrument for Nursing Homes." *Gerontologist* 30 (3): 293–307.
- Morris, J. N., T. Moore, R. Jones, V. Mor, J. Angelelli, K. Berg, C. Hale, S. Morris, K. M. Murphy, and M. Rennison. 2003. "Validation of Long-Term and Post-Acute Care Quality Indicators. CMS Final Report" [accessed on November 12, 2004]. Available at www.cms.hhs.gov/
- Mukamel, D. B. 1997. "Risk-Adjusted Outcome Measures and Quality of Care in Nursing Homes." *Medical Care* 35 (4): 367–85.
- National Quality Forum. 2002. "Nursing Home Performance Measures: Meeting Summary" [accessed on September 7, 2005]. Available at www.qualityforum.org/txNHmeetingsummaryweb6-15-03.pdf
- Normand, S. T., M. E. Glickman, and C. A. Gatsonis. 1997. "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications." *Journal of the American Statistical Association* 92 (439): 803–14.
- Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

- Raudenbush, S. W., A. S. Bryk, and R. Congdon. 2004. *HLM: Hierarchical Linear and Nonlinear Modeling Statistical Software (Release 6.02) Software*. Lincolnwood, IL: Scientific Software International.
- Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, UK: John Wiley & Sons.
- Stafford, R. S., D. Li, R. B. Davis, and L. I. Iezzoni. 2004. "Modelling the Ability of Risk Adjusters to Reduce Adverse Selection in Managed Care." *Applied Health Economics and Health Policy* 3 (2): 107–14.
- Zimmerman, D. R. 2003. "Improving Nursing Home Quality of Care through Outcomes Data: The MDS Quality Indicators." *International Journal of Geriatric Psychiatry* 18 (3): 250–7.
- Zimmerman, D. R., S. L. Karon, G. Arling, B. R. Clark, T. Collins, R. Ross, and F. Sainfort. 1995. "Development and Testing of Nursing Home Quality Indicators." *Health Care Financing Review* 16 (4): 107–27.