DOI: 10.1111/j.1475-6773.2005.00400.x

The Quality of the Quality Indicator of Pain Derived from the Minimum Data Set

Ning Wu, Susan C. Miller, Kate Lapane, Jason Roy, and Vincent Mor

Objective. To examine facility variation in data quality of the level of pain documented in the minimum data set (MDS) as a function of level of hospice enrollment in nursing homes (NHs).

Data Source. Clinical assessments on 3,469 nonhospice residents from 178 NHs were merged with On-line Survey Certification and Reporting data of 2000, Medicare Claims data of 2000 and the MDS of 2000–2002.

Study Design. Using the same assessment protocol, NH staff and study nurses independently assessed 3,469 nonhospice residents. Study nurses' assessments being gold standard, we quantified and compared quality of NH staff's pain rating across NHs with high, medium, or low hospice use. Multilevel models were built to assess the effect of NH hospice use levels on the occurrence of false positive (FP) and false negative (FN) errors in NH-rated "severe pain."

Principal Findings. Of 178 NHs, 25 had medium and 41 high hospice use. NHs with higher hospice use had lower sensitivities. In multilevel analysis, we found a significant facility-level variation in the probability of FP and FN errors in facility-rated "severe pain." Resident characteristics only explained 4 and 0 percent of the facility variation in FP and FN, respectively; characteristics and locations (state) of NHs further explained 53 and 52 percent of the variance. After controlling for resident and NH characteristics, staff in NHs with medium or high hospice use were less likely to have FP or FN errors in their MDS documentation of pain than were staff in NHs with low or no hospice use.

Conclusions. The examination of data quality of pooled MDS data from multiple NHs is insufficient. Multilevel analysis is needed to elucidate sources of heterogeneity in the quality of MDS data across NHs. Facility characteristics, e.g., hospice use or NH location, are systematically associated with overrated/underrated pain and may bias pain quality indicator (QI) comparisons. To ensure the integrity of QI comparison in the NH setting, the government may need to institute regular audits of MDS data quality.

Key Words. Multilevel analysis, nursing home, minimum data set, pain, hospice, quality indicator, reliability

Quality indicators (QIs) are developed by government and researchers to measure and monitor the performance of health care providers (Starfield 1998). To be considered useful for monitoring providers' quality of care, a QI must be clinically meaningful and have good reliability and validity (Berg et al. 2002). Existing studies have addressed some of the measurement features a QI should process, including the stability of QIs over time, sample size issues, the responsiveness of QIs to interventions, and the use of risk adjustment in QI comparisons (Starfield 1998; Kritchevsky et al. 1999; Rosen et al. 1999; Gandjour et al. 2002; Mor et al. 2003; Rantz et al. 2004). However, the extent and causes of ascertainment bias have been studied less extensively (Mor et al. 2003).

Ascertainment bias refers to systematic errors in the assessment and documentation of the true prevalence of a phenomenon (e.g., a clinical condition) because of the difference in assessors' measurement skills or adherence to measurement protocols (Carr 2004). Ascertainment bias in the data used for QI calculation biases provider–performance comparisons. If the comparisons are to guide the regulators to reward or punish performance outliers, then invalid results may have a serious adverse impact on both health care providers and consumers. In the nursing home (NH) setting, QIs are calculated from a facility-generated data set—the minimum data set (MDS). Ascertainment bias occurs when facility staff under- or overestimate the severity of residents' conditions relative to either other residents (i.e., at a resident level) or other NHs (i.e., at a facility level). To-date, fewer studies have examined, or compared, the extent of ascertainment bias at the facility level than at the resident level (Simmons et al. 2004). One reason is the difficulty in obtaining a "gold standard" measure from multiple providers to which provider-generated data can be compared. Even when "gold standard" data are available, examination of data quality has often been based on pooled data of all the providers, ignoring the variation in the data quality across providers (Gambassi et al. 1998; Fisher et al. 2002; Schnelle et al. 2003). One exception is the study conducted by Mor et al. (2003) that found significant variation in the facility-specific κ's of QI components derived from facility-generated MDS.

Address correspondence to Ning Wu, M.D., Ph.D., Health Services Research and Evaluation, Abt Associates Inc., 55 Wheeler St., Cambridge, MA 02138. Susan C. Miller, Ph.D., is with the Center for Gerontology and Health Care Research, Brown University School of Medicine, Providence, RI. Kate Lapane, Ph.D., and Vincent Mor, Ph.D., are with the Department of Community Health, Brown University School of Medicine, Providence, RI. Jason Roy, Ph.D., is with the Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY.

Multilevel analysis may be an appropriate approach to examine the existence and extent of ascertainment bias of QI data used to compare provider performance. Multilevel analysis has been used in studies evaluating the performance of health care providers, since to make inference or draw conclusions about providers, not individual patients, traditional statistical methods that ignore the hierarchical data structure are not appropriate (Leyland and Goldstein 2001). Similar to provider performance, the quality of a provider's data is uniformly affected by the provider's characteristics, e.g., staffing level and resident case mix, etc. Multilevel analysis takes into account the clustered data structure, allows researchers to estimate the effect of providers' characteristics on the extent and direction of ascertainment bias, generates less biased estimates for standard errors of effect, and allows the consideration of how such bias may influence the QI comparison.

Multilevel analysis can also assist in developing QIs. A good QI should not only be clinically meaningful but also not be subject to measurement error. The utility of an important QI can be compromised if it is difficult to collect valid and/or reliable data (Huff 1997). To date, most of the emphasis in the development of QIs has been on experts' opinions and insufficient emphasis has been devoted to empirical and quantitative analyses on the data quality of QIs (Huff 1997). With multilevel analysis, researchers can compare the susceptibility of different data items to measurement errors. If alternatives exist and both items are considered to be clinically meaningful, it would be preferable to select the more reliable item for QI calculation.

In this article, using the largest reliability data file collected in the NH setting to date and using multilevel analysis, we illustrate how NH characteristics may affect NH staff's adherence to assessment protocols, and in turn, the reliability and validity of the MDS-derived variable "severe pain" for the calculation of the pain QI now being publicly reported. We also create another MDS-derived pain variable and compare its data quality to that of "severe pain."

Specifically, we examine the impact of NH hospice use on the quality of the MDS-derived pain variables in nonhospice residents. Previous studies have suggested that residents enrolled in hospice received better assessment and management of pain than do residents not enrolled in hospice (Miller, Gozalo, and Mor 2000). In addition, nonhospice residents in NHs with a higher level of hospice use may receive more comprehensive pain assessments than nonhospice residents in NHs with no or low hospice use, probably because of NH staff's acquiring skills in detecting pain via their collaboration with hospice staff, a phenomenon often referred to as the hospice spill over

effect (Wu et al. 2003). If the comparison of the pain QI among NHs is unduly influenced by the differential assessment of pain as a function of the level of hospice use in NHs, conclusions are likely to be invalid. To further test the presence of hospice spill over effect in NH settings, we built multilevel models to assess the effect of NH hospice use on the occurrence of false positive (FP) and false negative (FN) errors in MDS-derived pain variables.

METHOD

Subjects

Subjects are 3,469 nonhospice NH residents from 178 NHs who participated in the "National Study to Validate the Long-Term and Post-Acute Care Quality Indicators Derived from MDS" (referred as the QI validation study). Detailed information about the study design and data collection procedure is published elsewhere (Morris et al. 2002; Mor et al. 2003). Briefly, 209 NHs from six states (California, Illinois, Missouri, Ohio, Pennsylvania, and Tennessee) participated the study in late 2001 and early 2002. In each NH, study nurses selected up to 30 residents with a recently completed MDS, and conducted an additional and abbreviated MDS assessment. The NH resident selection strategy was to ensure that the time interval between the two MDS assessments be short enough to examine interrater agreement between NH staff and study nurses. Of the 3,799 residents whose interval between the two assessments were less than or equal to 30 days, we excluded 63 (1.7 percent) because of missing data on at least one pain assessment and 78 (2.1 percent) who were on hospice when either one or both of the assessments were conducted. To obtain stable estimates for the effect of facility characteristics on data quality, we further excluded 31 NHs with less than 10 eligible participating residents. The remaining 3,469 nonhospice residents composed our study sample.

Data Sources

To determine residents' demographic characteristics and clinical conditions we merged the QI validation data with the 2000–2002 MDS Natural Repository file. Information on NH characteristics were obtained from 2001 On-line Survey Certification and Reporting data. The extent of hospice use in an NH was determined by using the 2000 Medicare claims data for all the residents in the study NHs. All the Medicare hospice claims during the year 2000 from the 178 participating NHs were abstracted.

Measurements

The MDS version 2.0 documents both the frequency and intensity of pain a resident experienced in 7 days prior to the assessment date. We used these two measures to create a scale to represent the severity of pain, a scale used in previous studies (Teno et al. 2001; Wu et al. 2003). Seven scores were given to indicate the increasing severity of pain: 0, no pain; 1, less than daily mild pain; 2, daily mild pain; 3, less than daily moderate pain; 4, daily moderate pain; 5, less than daily excruciating pain; and 6, daily excruciating pain. The CMS defines the pain QI as the prevalence of residents with daily moderate pain or excruciating pain at any frequency. Accordingly, we created a dichotomized variable indicating the existence of such pain (referred to as "severe pain"). In addition, based on the same seven-point pain scale we created another dichotomized pain variable to indicate the presence of pain which should be managed: no pain or mild less than daily pain versus mild daily pain or moderate/excruciating pain at any frequency (referred to as "mild daily or worse pain"). We reasoned that mild, persistent pain can have a significant negative impact on residents' quality of life. Hence, residents with mild daily pain should also receive appropriate pain management. Furthermore, we expect this pain variable be more reliable than "severe pain" since nurses would be more likely to agree on presence/absence of pain than on the ratings of pain levels (Manfredi et al. 2003).

In the process of assessment and documentation of pain, both study nurses and NH staff were expected to follow the same assessment protocol, contained in the Resident Assessment Instrument (RAI) Users' Manual (Morris et al. 2002). In accordance with the RAI manual, study nurses and NH staff were instructed to use all sources of information to complete the MDS, including residents, facility staff, attending physicians, and medical charts (study nurses were prohibited of course from reviewing the facility MDS record). No additional assessment instructions were given to study nurses. Study nurses completed the MDS assessments of sampled residents before attending to any other data collection in facilities. Thus, we are confident we can attribute the difference in pain ratings between NH staff and study nurses to random errors, or differential adherence to RAI protocols. The MDS assessments by NH staff were not specifically collected for the QI validation study, but rather were part of the routine assessments conducted by NHs and submitted to CMS to fulfill the government's mandate. Therefore, these can be presumed to reflect the real quality of MDS data compiled by CMS. The assessment data collected by study nurses were considered the gold standard because the study nurses were

experienced, uniformly trained to follow the RAI protocol, and certified before they were sent to NHs to collect data (Morris et al. 2002). Evidence of study nurses' interrater reliability to one another on pain assessments revealed κ's for pain frequency and intensity over 0.75 (Mor et al. 2003).

Hospice concentration is defined as the total (unduplicated) number of an NH's residents enrolled in hospice divided by the total (unduplicated) number of residents living in an NH within a predefined time period (e.g., 1 calendar year) (Miller, Gozalo, and Mor 2000). For the calculation of the hospice concentration, because of the unavailability of Medicare claims data for the year 2001, we linked the MDS Natural Repository file and the 2000 Medicare Claims data. The numerator was operationalized as the number of individuals with at least one hospice claim during their NH stays in 2000, and the denominator was operationalized as the number of individuals with at least one MDS assessment during the same period of time. We categorized the 178 NHs into three groups by their hospice concentration: no or low (hospice concentration <3 percent), medium (hospice concentration = 3–5 percent), and high (hospice concentration >5 percent) hospice use.

Analytic Approach

We compared residents' and facilities' characteristics across the three NH hospice use categories. To examine the facility variation in the quality of MDS pain data, we calculated sensitivity, specificity, γ and κ for the two dichotomized pain variables based on pooled data and data from individual NHs, assuming study nurses' rating as the gold standard. γ is the weighted difference between the FP and FN rates (Roy and Mor 2004). γ's value reflects the direction of the NH staff's measurement error: a γ larger than 0 implies that, on average, NH staff in the facility over report residents' pain relative to study nurses; smaller than 0 means underreporting; and equal to 0 is equivalent to no over-or underreport. κ is the chance corrected rater agreement (Brennan and Hays 1992). We compared results from pooled analysis and from analyses at the facility level. We also compared the distribution of these four facility-level statistics on data quality among the three hospice use categories. An existing study suggests NHs with very low rates of pain had more FN errors, and therefore, the low pain QIs may indicate underreporting rather than adequate management of pain (Cadogan et al. 2004). To further explore the association between the pain QI and quality of pain data, we plotted the facility-specific prevalence of pain and γ's.

Multilevel models with multinominal outcomes were built in *MLwiN* to extract the "pure" effect of NH hospice use on the validity of facility

documentation of "severe pain" and "mild daily or worse pain," respectively (Yang et al. 2001; Leyland and Goldstein 2001). By comparing with the gold standard (study nurses' assessments), facility-generated pain ratings were categorized into three groups: (1) FN, (2) FP, and (3) ratings that agree. For the ith resident living in jth NH, we modeled the probability of the resident's facility pain rating being in one of the three data quality categories. Consider a two-level model with one explanatory variable at individual level and one at facility level. The model can be expressed as

$$\log(\pi_{ij}^{(s)}/\pi_{ij}^{(r)}) = \beta_0^{(s)} + \beta_1^{(s)} x_{1ij} + \beta_2^{(s)} x_{2j} + u_{0j}^{(s)}$$

where $s = 1, 2$ for FN and FP, respectively; $r = 1$ for ratings that agree; u_0 indices random intercepts at facility level. Two sets of regression coefficients are estimated simultaneously for FP versus agreed ratings and FN versus agreed ratings, respectively.

We first fitted null models with random intercepts to examine the between-facility variation of average probabilities of FP or FN. Then we sequentially added resident characteristics, NH characteristics and the state in which study NHs were located in the models and observed the reduction in the unexplained facility-level variation of the average probabilities of FP and FN. Lastly we evaluated the impact of medium or high hospice use in an NH on the likelihood of occurrence of FP and FN errors relative to correct ratings. In the final models, we kept variables that either significantly improve the model fitting or change the estimated effect of NH hospice use on the quality of MDS pain data by over 10 percent.

RESULTS

Both NH staff and study nurses were less likely to record "severe pain" or "mild daily or worse pain" for nonhospice residents in NHs with medium or high hospice use compared with in NH with no or low hospice use (Table 1). In the pooled analysis of data quality, the dichotomized pain variables "severe pain" and "mild daily or worse pain" had good sensitivity and specificity, acceptable κ's and γ's close to zero (Table 2). There was a significant heterogeneity in the quality of the dichotomized pain variables across NHs (Figure 1). Although not statistically significant, on average NHs with high hospice use had lower sensitivity, γ and κ for "severe pain" than did NHs with less hospice use (Figure 1a). For the dichotomized pain variable "mild daily or worse pain" the data quality at facility level is relatively more homogenous across the

Table 1: Characteristics of Residents and Nursing Homes by Hospice Use Concentration Categories

	Hospice Concentration Categories		
Resident Characteristics	*No or Low (0–3%) (n = 2,285)*	*Medium (3–5%) (n = 453)*	*High (5%+) (n = 731)*
Having severe pain			
Rated by study nurses (%)	27.13	15.23	17.92
Rated by NH staff (%)	31.03	14.79	13.27
Having mild daily or worse pain			
Rated by study nurse (%)	44.90	28.48	28.86
Rated by NH staff (%)	46.87	28.92	25.58
Age categories (years)			
<65 (%)	13.30	12.14	8.34
75–85 (%)	36.46	29.80	30.64
85–95 (%)	28.10	35.32	42.13
95+ (%)	3.85	5.74	8.89
Nonwhite (%)	19.12	11.92	12.72
Men (%)	34.66	23.18	27.36
Not married (%)	72.04	80.57	83.31
Cancer (%)	11.07	4.86	6.84
Diabetes (%)	25.16	26.27	21.89
Osteoporosis (%)	12.15	15.67	11.22
Arthritis (%)	27.88	27.81	20.25
Any pressure ulcer (%)	22.76	18.54	15.18
Hip fracture in the past 6 months (%)	5.47	4.86	6.57
Wound infection (%)	7.75	3.97	3.56
Dementia (%)	20.61	30.68	31.19
Cognitive impairment			
No or mild (CPS = 0, 1) (%)	49.50	32.45	22.44
Moderate (CPS = 2, 3) (%)	31.29	39.96	47.20
Severe (CPS = 4–6) (%)	19.21	27.59	30.37
Nursing Home Characteristics	*No or Low (0–3%) (n = 110)*	*Medium (3–5%) (n = 25)*	*High (5%+) (n = 43)*
Hospice concentration			
Median (range)	0.67 (0–2.95)	3.63 (3.08–4.99)	6.88 (5.19–13.08)
Facility ran for profit (%)	38.47	59.38	56.36
Free-standing facility (%)	44.20	85.43	96.99
Part of a chain (%)	52.78	53.20	42.27
NH occupancy rate >80% (%)	67.83	67.11	82.22
Facility has special care unit	10.24	25.83	33.93
Percent of residents paid by Medicaid			
Mean ± SD	38.13 ± 35.39	60.64 ± 27.16	57.2 ± 22.64

continued

Table 1: *Continued*

Nursing Home Characteristics	*No or Low (0–3%) (n = 110)*	*Medium (3–5%) (n = 25)*	*High (5%+) (n = 43)*
Percent of residents paid by Medicare			
Mean ± SD	36.65 ± 34.87	14.91 ± 21.51	8.17 ± 13.69
Percent of residents with dementia			
Mean ± SD	1.76 ± 6.28	7.82 ± 15.13	8.84 ± 19.58
Number of nurse hour per resident			
Mean ± SD	4.55 ± 1.85	3.33 ± 1.11	3.04 ± 0.98
State in which NH was located			
California (%)	21.82	8.00	6.98
Illinois (%)	20.91	8.00	23.36
Missouri (%)	3.64	24.00	27.91
Ohio (%)	10.91	24.00	18.60
Pennsylvania (%)	18.18	32.00	18.60
Tennessee (%)	24.55	4.00	4.65

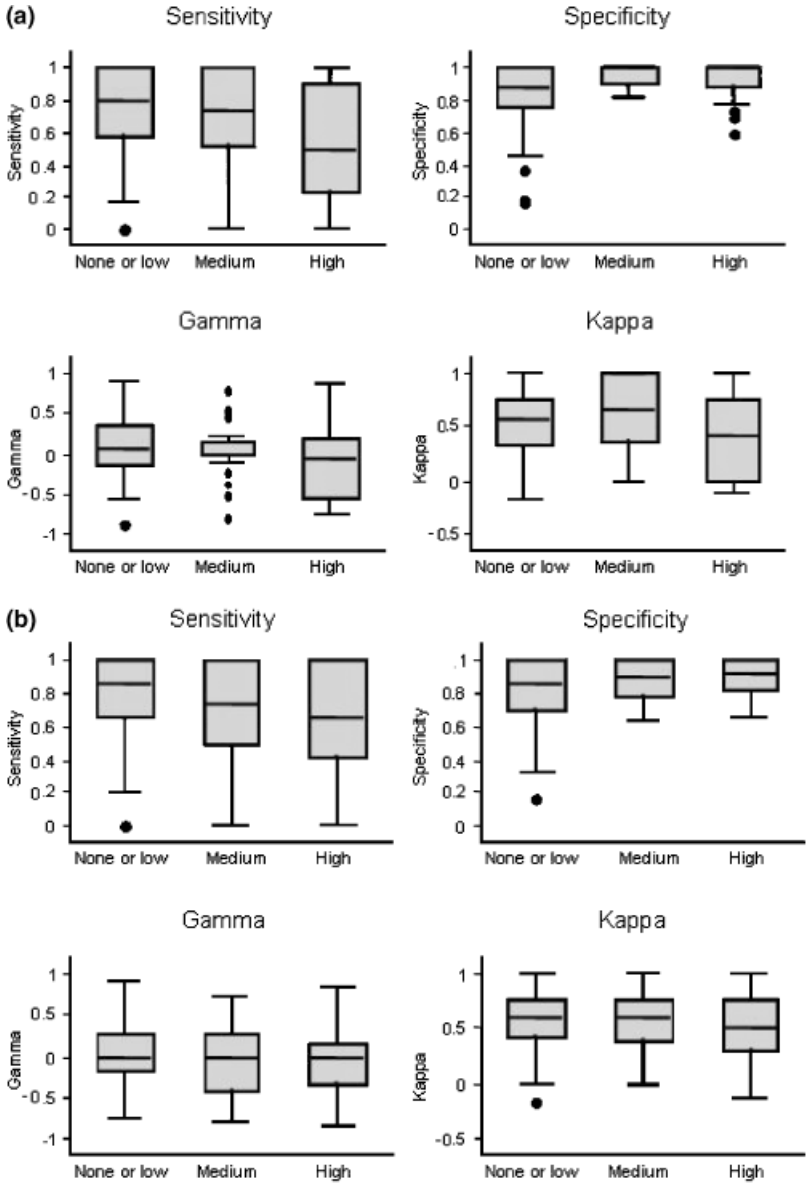
hospice use categories as compared with "severe pain" (Figure 1b). We did not find a strong association between the facility-specific prevalence of pain and γ's (Figure 2). In NHs with very low prevalence of "severe pain," about half of the NHs underrecord and the other half overrecord pain; in NHs with very high prevalence of "severe pain," most overrecord pain.

The intra-class correlation (ICC) of the probability of facility documented "severe pain" being FP and FN was 0.19 and 0.19, respectively (Table 3). There was a small covariance between the outcomes of FP and FN, suggesting an NH's probability of having FP errors is not correlated with its probability of

Table 2: Pooled Analysis on Quality of Dichotomized Pain Variables

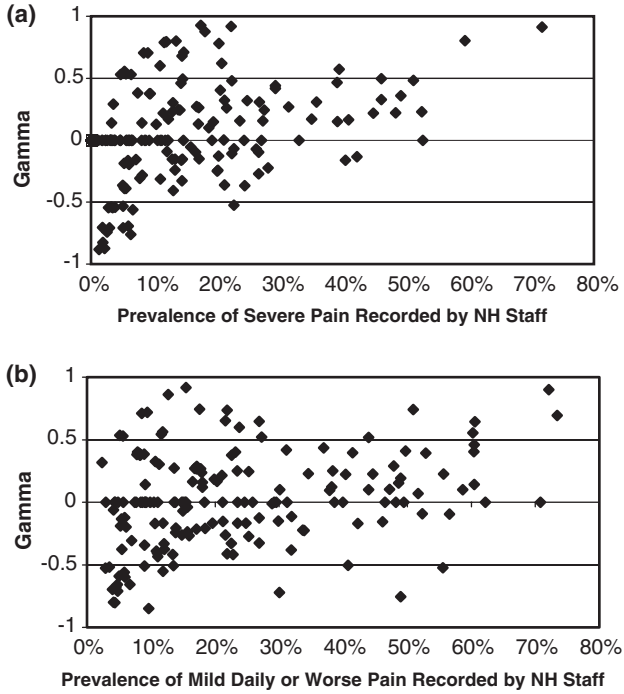
	N	*Sensitivity*	*Specificity*	γ	κ
Quality of the variable "Severe pain"					
All residents	3,469	0.72	0.89	0.05	0.60
Nursing home hospice use					
No or low	2,285	0.77	0.86	0.11	0.60
Medium	453	0.70	0.95	−0.02	0.65
High	731	0.50	0.95	−0.18	0.49
Quality of the variable "Mild daily or worse pain"					
All residents	3,469	0.80	0.86	0.02	0.65
Nursing home hospice use					
No or low	2,285	0.83	0.83	0.05	0.66
Medium	453	0.75	0.90	0.01	0.64
High	731	0.65	0.91	−0.10	0.58

Figure 1: Distribution of Facility-Level Statistics on the Quality of Dichotomized Pain Variables: (a) For "Severe Pain" and (b) For "Mild Daily and Worse Pain"



X-axis: categories of nursing home hospice concentration (none or low: 0–3 percent; medium: 3–5 percent; high: 5+ percent)

Figure 2: The Association between Facility-Specific Prevalence of Pain and γ: (a) "Severe Pain" and (b) "Mild Daily or Worse Pain"



having FN errors. With the inclusion of NH resident characteristics in the model, we observed a 4 percent reduction in the estimate of the facility-level variance for FP and none for FN. That is, only a small fraction of the between-facility difference in the facility-specific proportion of FP was explained by case mix; and these resident characteristics were unable to explain the variation of FN among NHs. After adding NH characteristics in the model (except for states), the facility-level variance in FP further decreased by 40 percent, in FN reduced by 23 percent. Finally, in the full model with resident and NH characteristics as well as state, the facility-level variance decreased by 53 and 50 percent (Table 3). The state in which the study NH was located explained a substantial amount of between-facility variation of FP and FN. However, unexplained facility-level variances for both outcomes remained highly significant meaning there are still unmeasured facility-level factors that may contribute to the heterogeneity in the quality of MDS pain assessments.

Table 3: The Association between Hospice Concentration and the Occurrence of Disagreement in Pain Rating on Nonhospice Residents between NH and Study Nurses: Severe Pain versus Nonsevere Pain (Cutoff = 4)

	False Positive (FP)	*False Negative (FN)*
Null models with random intercepts		
Unexplained variance at facility level		
Variance	0.87 (0.17)*	0.88 (0.18)
Covariance	−0.12 (0.12)	
Intraclass correlation	0.19	0.19
Full models with random intercepts		
Hospice concentration (%)		
0–3	—†	—
3–5	0.39 (0.21–0.73)	0.54 (0.29–1.00)
5+	0.56 (0.33–0.95)	1.00 (0.61–1.65)
Age (centered)		
Linear term	0.94 (0.78–1.13)	1.03 (0.85–1.25)
Quadratic term	0.99 (0.92–1.07)	1.00 (0.91–1.10)
Cognitive impairment		
No or mild (CPS = 0 or 1)	—	—
Moderate (CPS = 2 or 3)	0.57 (0.42–0.79)	0.80 (0.57–1.13)
Severe (CPS = 4, 5, or 6)	0.34 (0.21–0.55)	0.72 (0.47–1.11)
Having cancer	1.04 (0.69–1.57)	1.55 (1.02–2.35)
Having dementia	1.02 (0.68–1.52)	1.06 (0.73–1.54)
Male	0.75 (0.57–1.00)	0.85 (0.62–1.15)
Nurse hours/resident	1.04 (0.86–1.26)	1.00 (0.81–1.23)
Occupancy rate <80%	1.50 (1.01–2.25)	0.75 (0.48–1.18)
Part of a chain	1.00 (0.71–1.41)	1.29 (0.90–1.86)
Having 100+ beds	0.81 (0.53–1.25)	0.66 (0.43–1.02)
Number of health deficiencies >50th percentile in the state	0.96 (0.65–1.43)	0.99 (0.66–1.49)
Location of the study nursing home		
California	2.07 (1.11–3.86)	0.52 (0.24–1.11)
Illinois	1.25 (0.66–2.36)	1.40 (0.74–2.65)
Missouri	1.76 (0.79–3.93)	4.07 (1.99–8.30)
Ohio	2.37 (1.25–4.49)	1.21 (0.62–2.39)
Pennsylvania	1.39 (0.75–2.58)	1.41 (0.77–2.58)
Tennessee	—	—
Unexplained variance at facility level		
Variance	0.41 (0.12)	0.44 (0.14)
Covariance	−0.07 (0.10)	

*Standard error of variance at the level of nursing home.

†Referent group.

After controlling for resident and NH characteristics, facility-rated "severe pain" from NHs with medium hospice use were more likely to agree with gold standards, i.e., were half as likely to have FP or FN errors, in comparison

with no or low hospice use (Table 3). In NHs with high hospice use, facility pain ratings were as likely to be FN as in NHs with low or no hospice use, whereas they were less likely to be FP.

Table 4: The Association between Hospice Concentration and the Occurrence of Disagreement in Pain Rating on Nonhospice Residents between NH and Study Nurses: Mild Daily or Worse Pain versus No Pain or Mild Less than Daily Pain (Cutoff = 1)

	False Positive (FP)	*False Negative (FN)*
Null models with random intercepts		
Unexplained variance at facility level		
Variance	0.60 (0.13)*	0.47 (0.12)
Covariance	−0.09 (0.09)	
Intraclass correlation	0.10	0.06
Full models with random intercepts		
Hospice concentration (%)		
0–3	—†	—
3–5	0.62 (0.37–1.04)	0.72 (0.44–1.20)
5+	0.60 (0.37–0.97)	0.82 (0.54–1.27)
Age (centered)		
Linear term	1.05 (0.88–1.25)	1.30 (1.10–1.54)
Quadratic term	0.96 (0.88–1.05)	1.05 (0.96–1.14)
Cognitive impairment		
No or mild (CPS = 0 or 1)	—	—
Moderate (CPS = 2 or 3)	0.64 (0.47–0.86)	0.81 (0.59–1.12)
Severe (CPS = 4, 5, or 6)	0.46 (0.30–0.69)	1.04 (0.71–1.51)
Having cancer	0.99 (0.66–1.49)	1.24 (0.82–1.88)
Having dementia	1.14 (0.80–1.61)	0.94 (0.68–1.32)
Male	0.87 (0.66–1.14)	1.06 (0.80–1.40)
Nurse hours/resident	0.89 (0.74–1.08)	0.95 (0.80–1.14)
Occupancy rate <80%	1.57 (1.08–2.30)	0.94 (0.64–1.38)
Part of a chain	0.77 (0.56–1.06)	1.09 (0.80–1.49)
Having 100+beds	1.13 (0.76–1.67)	1.01 (0.69–1.46)
Number of health deficiencies >50th percentile in the state	1.06 (0.74–1.51)	1.02 (0.73–1.44)
Location of the study nursing home		
California	1.71 (0.95–3.09)	0.73 (0.39–1.35)
Illinois	1.11 (0.61–2.02)	1.76 (1.04–2.97)
Missouri	1.48 (0.71–3.07)	3.05 (1.66–5.60)
Ohio	2.72 (1.52–4.87)	0.95 (0.51–1.75)
Pennsylvania	1.40 (0.80–2.46)	1.12 (0.67–1.88)
Tennessee	—	—
Unexplained variance at facility level		
Variance	0.36 (0.11)	0.25 (0.10)
Covariance	0.01 (0.08)	

*Standard error of variance at the level of nursing home.

†Referent group.

We observed a much smaller facility-level variation in the probability of FP and FN for "mild daily or worse pain" than for "severe pain." The ICC was 0.10 and 0.06 for FP and FN, respectively (Table 4). The inclusion of resident characteristics in the multilevel model reduced the variation of FP by 12 percent and the variation of FN by 3.2 percent; the inclusion of facility characteristics (except for states) further reduced the variations by 8 percent for both. In the full model, 40 and 47 percent of facility variance in FP and FN was explained by the covariates (Table 4).

DISCUSSION

Ascertainment bias in the identification of clinical conditions threatens the validity of QI comparisons of health care providers' performance. This is the first study that uses multilevel analysis to explore the impact of NH characteristics on the quality of facility-generated MDS data.

There are four major findings in our study. First, it is insufficient to examine the data quality using pooled data over multiple providers. The variations in data quality across providers may not be identified by examining the overall quality of the pooled data because errors could average out, or appear insignificant. Stratified analysis on pooled data provided some insight on the variation of data quality in NHs with different levels of hospice use. However, the comparison of data quality across hospice use strata is confounded by factors that are simultaneously associated with both data quality and NH hospice use. Our findings underline the importance of multilevel analysis in assessing the quality of provider-generated data.

Second, in our study we found facility characteristics to be more important than resident characteristics in explaining facility variations in the probability of disagreement between facility-rated pain and our gold standard nurse raters. One reason is that the pain ratings given by study nurses may also be biased by residents' characteristics given the assessment protocol and limited time to observe and interact with residents. On the other hand, the study nurses were uniformly trained and were from outside the participating NHs. Therefore, even though study nurses' assessments also had errors, the errors were largely because of study nurses' differential assessments on residents, rather than because of the facility culture or management/organizational structure. In contrast, NH staff's assessments were influenced by both resident and facility characteristics. Hence, the design of the QI validation study enabled us to separate the impact of individual and facility characteristics on the quality of MDS pain data. It is

important to note that one purpose of the QI validation study was to collect the best-quality MDS data under realistic conditions. The implication is if we were to implement a regular data quality audit in NHs, given the available resources (e.g., inadequate protocols and limited financial resources), the data collected for the purpose of audit might also be biased by resident characteristics. Yet such a quality audit may still identify important provider-level factors that systematically bias QI calculation and QI comparisons.

Actually, in the comparison of the NH pain QI, ascertainment bias related to resident characteristics may be partially corrected by adjusting for resident characteristics, e.g., cognitive function or age, assuming that the existence and extent of such biases are ubiquitous in all the NHs. On the other hand, the facility-related ascertainment bias is more serious, albeit less studied. Our results suggest that the local environment, or culture, in the facility, e.g., administrative emphasis on data quality, paces of daily tasks being conducted, available time nurses have to interact with residents, and interaction between nurses and nurse assistants are important in causing the disparity in data quality. At present, no simple adjustment strategy is available, or deemed appropriate, to address facility-related ascertainment bias in the QI comparison. More research is needed to address this problem. In the long run, to obtain high-quality data from all NHs, strategies need to be developed to minimize the impact of variation in NHs' organizational structure and resources on measurement acuity.

Third, we did not find a clear association between the pain QIs and the direction of ascertainment bias. In Figure 2a, facilities with extremely low pain QIs had a wide range of γ's, with a few more facilities at the negative end suggesting pain was underdocumented by NH staff. The reverse was observed in facilities with extremely high pain QI. In our study, the pain QI derived from facility MDS can be a mixed result of both prevalence and care process of pain in a facility, as well as measurement errors and sampling errors (i.e., small random samples of residents per facility and random variation of QI over time). It is difficult to attribute the extreme QIs to any one of the four reasons. When interpreting QIs released by CMS, sampling errors and care process may play a less important role than did in our study, since the QIs are often derived from a larger sample over 3 months, and the facilities may have stopped recording pain to justify analgesic use. Still it is necessary to collect validation data to judge the validity of observed QIs.

Finally, although derived from the same MDS-derived pain scale, the two dichotomized pain variables created with different cut-offs were found to have different reliability and validity. Compared with that of "severe pain,"

the quality of "mild daily or worse pain" is not only better in the pooled analysis, but also less subject to ascertainment bias with regard to NH and resident characteristics. It is likely that raters may agree on the presence of pain, but may assign different values of pain level based on raters' experience and/or expectation. Because QIs are always derived from dichotomized variables (e.g., presence of a symptom or receipt of a medical procedure), where the QI is dichotomized may have a significant impact on the validity of QI comparison. In promulgating QIs one has to balance between clinical meaningfulness and data quality.

According to the results of the multilevel models, the quality of MDS documentation of pain on nonhospice residents is better in NHs with medium or high levels of hospice use than in NHs with no or low hospice use. This finding is compatible with what previous studies have suggested (Miller, Gozalo, and Mor 2000; Wu et al. 2003). We also found that the hospice spill over effect is the strongest in NHs with medium hospice use and less strong in NHs with high hospice use, indicating there could be different mechanisms at work in these two types of NHs. We assume the hospice spill over effect results from NH staff acquiring the knowledge and experience in detecting pain via their collaboration with hospice staff. However, in NHs with high hospice use, seriously dying residents may tend to be in one floor or wing of an NH, or the division of care responsibility between hospice and NH staff may be so complete that NH staff are not responsible for symptom detection or involved in the assessment process. Hence, it may be that in NHs with high hospice use, NH staff do not have as much opportunity to learn as in NHs with medium hospice use to improve their ability to correctly measure pain.

In our study, on average, study nurses' assessments were 13 days later than facility MDS (SD = 8, range: 0–30). The experience of pain in residents could have varied and residents may have received pain interventions before study nurses' assessment. If hospice spill over effect manifests as better pain management, then residents in NHs with higher hospice use would be more likely to have received adequate medications if "severe pain" was detected. These residents would also be less likely to be in "severe pain" subsequently when study nurses assessed them for a second time. Hence, in our study some FP errors may actually reflect appropriate pain management, and because of this, we would overestimate the effect of NH hospice use on the likelihood of FP.

Limitations

By treating study nurses' documentation as the gold standard, we assumed their assessments reflected the real experience of pain in participating resi-

dents. This may not be true. Although research nurses were strictly trained and experienced, they were not as familiar with the pain behaviors in residents with cognitive impairment as were NH nurses, and may have been less likely to "catch" the pain episodes because of the time limited nature of these assessments. Furthermore, our results showed that for residents with moderate or severe cognitive impairment, there is less likelihood of rating disagreement between NH and study nurses than for residents without cognitive impairment—the regression coefficients for FP and FN are both negative in relation to moderate or severe cognitive impairment (Table 3). This may be an indication that study nurses may have filled out the MDS by consulting the same source of information as NH staff. Thus, the documentation of pain on MDS would be equally affected by these sources, and we would underestimate the variation between raters. However, this fact would not substantially bias the estimated impact of NH hospice use on data quality with the inclusion of residents' cognitive function (as a surrogate for study nurses' reliance on medical charts and care-givers when completing MDS) in the multilevel models.

To calculate the hospice concentration, ideally we would have used MDS and claims data of the year 2001, during which the QI validation study was conducted. However, we only had access to the 2000 Medicare Claims file. Studies have shown that NH hospice use is relatively stable over a short period of time (Miller, Gozalo, and Mor 2000). To evaluate the stability of hospice concentration in an NH over a 6-month period, we calculated and compared hospice concentrations for the 178 NHs between January–June 2000 and July–December 2000. We found little difference and the Pearson correlation between the pairs was very high (data not shown).

Implications

We need further studies on the impact of facility factors on the quality of MDS data since, in this study, a significant amount of facility variation in data quality remained unexplained. However, it is difficult and costly to have uniformly trained research staff collect repeated measures in multiple sites and large-scale studies are difficult to conduct by a single or a few research institutes. At present, the government does not have a plan to regularly monitor the data quality of MDS data. To ensure the integrity of QI comparisons, our results suggest that CMS examine the possibility of instituting regular audits of MDS data quality. We found great variation in the quality of MDS-derived "severe pain" QI. Similar problems may exist for the QIs on pressure ulcers, incon-

tinence and weight loss, mood, etc. Furthermore, ascertainment bias is not unique to the NH setting. QIs are now used in numerous provider settings and many of these indicators are derived from provider-collected data (Mor et al. 2003). Indeed, in some settings, nonstaff contractors are responsible for medical chart abstraction. Although it is cautioned by CMS that the values of QIs of a provider are only suggestive of poor or high quality of care, and thus, warrant further investigation, not knowing the complicated technical barriers and statistical complexity behind the numbers, consumers are likely to use the QIs as the only source of information. Therefore, ensuring data integrity through regular monitoring of data quality is important if QIs are to be a useful quality screen for consumers.

ACKNOWLEDGMENTS

This study is partially supported by the NIA grant # AG11624. The authors would like to thank Dr. Joseph Angelelli from Pennsylvania State University and Ms. Margaret Bryn at Hebrew Research Center for compiling the data file.

REFERENCE

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.

Brennan, P. F., and B. J. Hays. 1992. "The Kappa Statistic for Establishing Interrater Reliability in the Secondary Analysis of Qualitative Clinical Data." *Research in Nursing and Health* 15 (2): 153–8.

Cadogan, M. P., J. F. Schnelle, N. Yamamoto-Mitani, G. Cabrera, and S. F. Simmons. 2004. "A Minimum Data Set Prevalence of Pain Quality Indicator: Is It Accurate and Does It Reflect Differences in Care Processes?" *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 59 (3): 281–5.

Carr, S. M. 2004. "Ascertainment Bias" [accessed on July 7, 2004]. Available at http://www.mun.ca/biology/scarr/Ascertainment_bias.htm

Fisher, S. E., L. D. Burgio, B. E. Thorn, R. Allen-Burge, J. Gerstle, D. L. Roth, and S. J. Allen. 2002. "Pain Assessment and Management in Cognitively Impaired Nursing Home Residents: Association of Certified Nursing Assistant Pain Report, Minimum Data Set Pain Report, and Analgesic Medication Use." *Journal of the American Geriatrics Society* 50 (1): 152–6.

Gambassi, G., F. Landi, L. Peng, C. Brostrup-Jensen, K. Calore, J. Hiris, L. Lipsitz, V. Mor, and R. Bernabei. 1998. "Validity of Diagnostic and Drug Data in

Standardized Nursing Home Resident Assessments: Potential for Geriatric Pharmacoepidemiology." *Medical Care* 36 (2): 167–79.

Gandjour, A., F. Kleinschmit, V. Littmann, and K. W. Lauterbach. 2002. "An Evidence-Based Evaluation of Quality and Efficiency Indicators." *Quality Management in Health Care* 10 (4): 41–52.

Huff, E. D. 1997. "Comprehensive Reliability Assessment and Comparison of Quality Indicators and Their Components." *Journal of Clinical Epidemiology* 50 (12): 1395–404.

Kritchevsky, S. B., B. I. Braun, P. A. Gross, C. S. Newcomb, C. A. Kelleher, and B. P. Simmons. 1999. "Definition and Adjustment of Cesarean Section Rates and Assessments of Hospital Performance." *International Journal for Quality in Health Care* 11 (4): 283–91.

Leyland, A. H., and H. Goldstein. 2001. *Multilevel Modelling of Health Statistics.* New York: John Wiley & Sons.

Manfredi, P. L., B. Breuer, D. E. Meier, and L. Libow. 2003. "Pain Assessment in Elderly Patients with Severe Dementia." *Journal of Pain and Symptom Management* 25 (1): 48–52.

Miller, S. C., P. Gozalo, and V. Mor. 2000. "Outcome and Utilization for Hospice and Non-Hospice Nursing Facility Decedents." In *Synthesis and Analysis of Medicare's Hospice Benefit.* Washington, DC: Office of Disability, Aging, and Long Term Care Policy in the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services.

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., J. Angelelli, R. Jones, J. Roy, T. Moore, and J. N. Morris. 2003. "Inter-Rater Reliability of Nursing Home Quality Indicators in the US." *BMC Health Services Research* 3 (1): 20.

Morris, J. N., T. Moore, R. Jones, V. Mor, J. Angelielli, K. Berg, C. Hale, S. Morris, K. M. Murphy, and M. Rennison. 2002. "Validation of Long-Term and Post-Acute Care Quality Indicators." CMS Contract No: 500-95-0062/T.O. #4

Rantz, M. J., L. Hicks, G. F. Petroski, R. W. Madsen, D. R. Mehr, V. Conn, M. Zwygart-Staffacher, and M. Maas. 2004. "Stability and Sensitivity of Nursing Home Quality Indicators." *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 59 (1): 79–82.

Rosen, A. K., D. R. Berlowitz, J. J. Anderson, A. S. Ash, L. E. Kazis, and M. A. Moskowitz. 1999. "Functional Status Outcomes for Assessment of Quality in Long-Term Care." *International Journal for Quality in Health Care* 11 (1): 37–46.

Roy, J., and V. Mor. "The Effect of Provider-Level Ascertainment Bias on Profiling Nursing Homes." *Statistics in Medicine* 2004 (in press).

Schnelle, J. F., M. P. Cadogan, D. Grbic, B. M. Bates-Jensen, D. Osterweil, J. Yoshii, and S. F. Simmons. 2003. "A Standardized Quality Assessment System to Evaluate Incontinence Care in the Nursing Home." *Journal of the American Geriatrics Society* 51 (12): 1754–61.

Simmons, S. F., M. P. Cadogan, G. R. Cabrera, N. R. Al-Samarrai, J. S. Jorge, L. Levy-Storms, D. Osterweil, and J. F. Schnelle. 2004. "The Minimum Data Set Depression Quality Indicator: Does It Reflect Differences in Care Processes?" *The Gerontologist* 44 (4): 554–64.

Starfield, B. 1998. "Quality-of-Care Research: Internal Elegance and External Relevance." *Journal of the American Medical Association* 280 (11): 1006–8.

Teno, J. M., S. Weitzen, T. Wetle, and V. Mor. 2001. "Persistent Pain in Nursing Home Residents." *Journal of the American Medical Association* 285 (16): 2081.

Wu, N., S. C. Miller, K. Lapane, and P. Gozalo. 2003. "The Problem of Assessment Bias When Measuring the Hospice Effect on Nursing Home Residents Pain." *Journal of Pain and Symptom Management* 26 (5): 998–1009.

Yang, M., J. Rasbash, H. Goldstein, and M. Barbosa. 2001. "MLwiN Macros for Advanced Multilevel Modelling" [accessed on July 7, 2004]. Available at http://www.multilevel.ioe.ac.uk/download/advmacma.pdf