

Reliability of the Nursing Home Survey Process: A Simultaneous Survey Approach

Robert H. Lee, PhD,¹ Byron J. Gajewski, PhD,³ and Sarah Thompson, PhD²

Purpose: We designed this study to examine the reliability of the nursing home survey process in the state of Kansas using regular and simultaneous survey teams. In particular, the study examined how two survey teams exposed to the same information at the same time differed in their interpretations. **Design and Methods:** The protocol for simultaneous surveys consists of having one in-region and one out-of-region team survey a facility together. **Results:** The regular and simultaneous survey teams generally agreed about the number of deficiencies. The intraclass correlation coefficient was 0.87 for total deficiencies and 0.76 for deficiencies with scores of G or higher. But in a substantial number of instances the teams did not agree about the scope and severity of the deficiency or about what regulation the nursing home had breached. **Implications:** The survey process is reliable when assessing aggregate results, but it is only moderately reliable when examining individual citations. Stakeholders (i.e., consumers, policy makers, nursing home administrators) should be aware of the limitations of the survey process. It needs to be modified to reduce variability.

Key Words: Federal citations, F tags, Quality of care, Deficiencies

In order to participate in Medicare and Medicaid, nursing facilities must meet conditions of participation set by the Centers for Medicare and Medicaid Services (CMS; for a review, see Mullan & Harrington, 2001). In order to ensure compliance with 189 federal regulations, state survey agencies must inspect each nursing facility every 9 to 15

months (CMS, 2005). These regulations fall into several categories: resident rights, quality of life, quality of care, resident assessment, services, dietary, pharmacy, rehabilitation, dental and physician, physical environment, and administration. Surveyors cite deficiencies when a facility does not substantially comply with a regulation. Although the regulations and survey process are federally mandated, state agencies carry out the survey process.

Dissatisfaction with the survey process is widespread. Resident advocacy groups stress that state survey teams often miss important problems with care and fail to respond to complaints quickly. A Government Accountability Office (GAO; 2004) study identified several reasons for these shortcomings: insufficient and inexperienced survey staff, confusion about the regulations, inadequate state oversight of the survey process, and the predictable timing of surveys. Surveyors question the integrity of the inspection, political pressures to water down inspection findings, and the effectiveness of the enforcement process (Grassley, 2004). Industry representatives argue that the current survey and enforcement system "is an entirely subjective, process-oriented snapshot inspection system that focuses on punishment—not quality improvement" (Ousley, 2001 p. 1).

An ongoing concern for all of these stakeholders is that the number of deficiencies varies substantially between states (GAO, 2003). For example, in 2001 the proportion of deficiency-free nursing homes ranged from 33.5% in Virginia to 0% in Nevada, and the mean number of deficiencies ranged from a high of 14.2 per facility in Nevada to a low of 1.9 per facility in New Jersey (Office of the Inspector General, 2003).

Variation also exists within states. For example, the state of Kansas is composed of 6 survey regions. In 2001 facilities in the Northeast Region averaged 11.64 deficiencies, nearly three times as many as facilities in the West Region (3.69 deficiencies). Furthermore, deficiencies in the Northeast Region tended to be assigned higher scope and severity. Administrators and directors of nursing tended to think this heterogeneity reflected differences in the survey process; surveyors thought it reflected differences in facility characteristics. Although they did

Address correspondence to Robert H. Lee, Department of Health Policy and Management, University of Kansas Medical Center, Mail Stop 3044, 3901 Rainbow Boulevard, Kansas City, KS 66160. E-mail: rlee2@kumc.edu

¹Department of Health Policy and Management, University of Kansas School of Medicine, Kansas City.

²University of Kansas School of Nursing, Kansas City.

³Schools of Allied Health and Nursing, Center for Biostatistics and Advanced Informatics, University of Kansas Medical Center, Kansas City.

Table 1. Scope and Severity Matrix

Severity of the Deficiency	Scope of the Deficiency, Rating (State Share)		
	Isolated	Pattern	Widespread
Immediate jeopardy to resident health or safety	J (0.2%)	K (0.0%)	L (0.0%)
Actual harm that is not immediate jeopardy	G (5.8%)	H (0.0%)	I (0.0%)
No actual harm with potential for more than minimal harm that is not immediate jeopardy	D (45.0%)	E (34.0%)	F (9.7%)
No actual harm with potential for minimal harm	A (0.0%)	B (0.9%)	C (4.3%)

Notes: The State Share is the percentage of deficiency citations with this scope and severity cited in surveys of free-standing Kansas nursing homes in 2003. F, H, I, J, K and L deficiencies may constitute substandard quality of care. Fines may be levied or restrictions on participation in Medicare and Medicaid may be imposed.

not resolve this question, our earlier analyses found statistically significant regional differences ($p < .001$) even after controlling for size, case mix, nursing hours per resident day, and ownership (Forbes-Thompson et al., 2003). The reliability of the survey process appears to be worthy of careful study.

The purpose of this study was to evaluate in some depth how and why Kansas survey teams varied in their assessments. More specifically, our aim was to compare the findings of two survey teams exposed to the same information at the same point in time. We addressed this aim using a mixture of quantitative and qualitative methods.

An overview of the survey process provides a context for our study. Surveys entail standard procedures plus flexibility once a team enters a nursing facility. The process begins with presurvey preparation that includes a review of the facility's quality indicators (Arling, Kane, Lewis, & Mueller, 2005), history of complaints, and previous survey results. The team then proceeds to an entrance conference with the administrator and an initial tour. After this the team selects a group of residents, based on presurvey information and the initial tour, for a more in-depth review. Using protocols established by CMS, the survey team gathers information in a number of ways, including medical record reviews, observations of direct resident care, resident interviews, family interviews, and observations of events such as activities and meals. Each phase of the survey process has detailed written guidelines, and as information is gathered, the team reviews it and sharpens the focus of the survey on potential problem areas.

This structure allows teams to react to and explore problems identified during data collection. It also allows for prioritization of problems while on

site. However, this flexibility may also increase the variability of the survey process, because surveys of apparently similar facilities may focus on quite different aspects of care. How detailed a survey becomes also may depend on the observational skills of the surveyors, the clinical and management skills of the surveyors, or the number of problems found.

On the last day of the survey, surveyors meet to interpret their findings and to identify the number, scope, and severity of deficiencies that they found. The survey team then meets with the administrative staff and shares its preliminary findings. In Kansas, a quality improvement coordinator reviews these findings before the team submits the final survey report to the Department on Aging.

We should note that the final survey report may not be "final." Nursing homes can appeal any deficiencies or penalties through an informal dispute resolution process. Reductions in the number, scope, and severity of citations are common (GAO, 2003a).

Some deficiencies identify more serious problems than others, and some deficiencies allow for the imposition of more serious penalties. Table 1 outlines the scope and severity of deficiencies that surveyors may cite. Ratings A through C indicate substantial compliance with recommendations, so only Category 1 remedies are permitted (Office of the Inspector General, 2005). These remedies include development of a plan to correct the problem, enhanced monitoring by the survey agency, or mandatory training. Teams often do not cite such deficiencies. There were 0 A citations in Kansas in 2003, 21 B citations, and 96 C citations.

Citations that are rated D, E, or G permit imposition of Category Two remedies. These remedies include fines, denials of payment for new admissions, or denials of payment for all residents. These are the most common types of citations. More than 1,700 D and E deficiencies were cited in Kansas in 2003. G deficiencies are far less frequent; only 129 were issued in 2003.

Deficiencies that are rated F, H, I, J, K, or L can result in Category Three remedies. These include fines, termination from Medicare and Medicaid, and temporary management by an individual chosen by the state agency. F deficiencies are fairly common; more than 200 were cited in 2003. In contrast, H–L deficiencies are uncommon. A total of 5 J deficiencies were cited in 2003.

In most instances, the Department on Aging imposes Category Two or Three penalties only when a nursing home has failed to make corrections by the time of its resurvey. As a result, Category Two or Three penalties are not common. During the second and third quarters of 2003, the Kansas Department on Aging imposed fines on 11 nursing homes and admission bans on 18 (Kansas Department on Aging, 2004). The Department did not terminate any nursing homes from Medicaid or install temporary management in any nursing homes.

Table 2. Simultaneous Survey Protocol

Protocol
<ol style="list-style-type: none"> 1. The RST guided all aspects of the survey process and followed normal policies and procedures. 2. RST assignments (e.g., who would conduct the closed record review) were shared with the SST so that the respective team members would be informed of their responsibilities. 3. All team meetings to discuss findings were held in separate locations and tape recorded for evaluation by the research team. 4. Preliminary off-site preparation was conducted in separate locations. The SST received the same presurvey documents to review as the RST. 5. The RST and SST were matched teams and respective SST members followed respective RST members one on one. 6. Team members were not allowed to discuss assessments or interpretations with members of the other team. 7. If the RST did not raise a concern, the SST was not allowed to pursue that issue. The SST was to document the issue in field notes. 8. Members of the SST followed respective RST members continuously (e.g., into residents' rooms to observe care and into meetings to interview staff). 9. All survey-related information (e.g., policies and procedures) were requested by and directed to the RST. Copies were made for the SST. 10. Teams and facilities were informed that the findings of the SST were not related to the facility's certification and state licensure.

Notes: RST = regular survey team; SST = simultaneous survey team.

The Department also recommended additional federal penalties to CMS.

Methods

Setting and Sample

Kansas has six geographical survey regions. Each region has at least two trained survey teams, a quality improvement coordinator, and a regional manager.

During the summer of 2003, we randomly selected two nursing homes from each region from a list of facilities scheduled for resurvey. We excluded from consideration nursing homes with fewer than 50 beds in order to reduce the burden on small facilities of having two survey teams in their home. Twelve homes comprised the sample for what we labeled "simultaneous surveys."

The simultaneous survey teams consisted of one in-region team (the regular survey team or RST) and one randomly selected out-of-region team (the simultaneous survey team or SST). The regional manager overseeing the annual survey selected the RST. The manager from another randomly selected region selected the SST. In order to ensure that survey differences were not due to their composition, we matched teams in size and expertise. For example, if the RST included their quality improvement coordinator, the SST also sent their quality improvement coordinator.

This design reflected two considerations. First, as we noted above, there were indications that the survey process varied by region. In order to examine this, the SST needed to come from a different survey region than the RST. Second, in order to ensure that the regular survey would be seen as valid by all interested parties, the RST needed to be assigned by the usual practice in that region. Otherwise a simultaneous survey might place a nursing home at a competitive advantage or

disadvantage. Clearly, other designs might be preferable in other circumstances.

Procedures

Table 2 outlines the simultaneous survey protocol. The RST entered facilities following the normal protocol as prescribed by CMS. A member of the research team immediately informed the administrator that the SST would be following them as part of a quality improvement evaluation. A member of the research team also informed the administrator that the SST would not be interviewing staff, looking at or requesting additional records, or evaluating residents on their own. The SST would be shadowing the RST and reviewing its information. The RST directed the survey in accordance with policies and procedures. Members of the SST followed their RST counterparts to observe the same environmental dynamics; however, we did not allow the two team members to discuss interpretations or assessments with each other.

Survey teams usually meet several times during a survey to review what information they have collected to that point. These meetings then guide the remainder of the survey. For example, teams can use these meetings to decide which resident problems should be emphasized or which additional staff interviews are needed. The RST and SST conducted their meetings at the same time in different locations and tape recorded them. We had instructed SST members to document the problem areas and interviews they would follow up on if they were conducting a regular survey; we used the information obtained from both teams in order to evaluate consistency and provide insights into decision-making processes that influenced survey results. A member of the research team was onsite to ensure that the RST and SST

Table 3. Deficiencies Cited by the RST and the SST

Facility	Total Deficiencies		G+ Deficiencies ^a		Same F Tag, Different Scope or Severity	Distinctly Different F Tags
	RST	SST	RST	SST		
1	22	23	2	2	5	14
2	3	3	0	0	1	0
3	30	31	3	5	6	14
4	9	19	0	1	4	11
5	16	24	0	1	9	11
6	17	17	2	1	7	6
7	19	15	0	1	4	5
8	18	23	1	2	6	15
9	8	9	1	1	1	7
10	13	16	0	0	6	7
11	0	1	0	0	0	1
12	6	3	0	0	0	5
Total	161	187	9	14	49	96
Intraclass correlation coefficient	0.87		0.76			
95% confidence interval	0.64–0.96		0.38–0.92			

Notes: RST = regular survey team; SST = simultaneous survey team.

^aG+ deficiencies include G, H, I, J, K, and L, but none higher than H were cited.

members followed the protocol and did not share information with one another.

Protocol Rationale

We took several issues into consideration when designing this protocol. One was to avoid compromising the quality of resident care. Survey teams tend to disrupt normal routines, and we were concerned that repeated inspections would lead to repeated disruptions. In addition, our primary goal was to evaluate the performance of two teams exposed to the same information. Because nursing homes must address violations that teams observe during the course of an inspection, having back-to-back surveys would not have guaranteed that a follow-up survey team would have been exposed to the same problems. Conducting simultaneous surveys minimized disruption and ensured that both teams analyzed the same information.

Data Analysis

Our aim was to describe how and why the conclusions of the RST and SST differed. We used a triangulated design using both quantitative and qualitative methods (Fielding & Fielding, 1986; Jick, 1979). Our analysis of how the conclusions differed was largely quantitative. We designed the qualitative analyses to add depth to the analyses and to help answer why the reports of the teams differed.

Our approach examined the data at two very different levels of aggregation. First, treating each nursing home facility as a random effect, we calcu-

lated the intraclass correlation coefficient (ICC). The ICC equals the between-facility variance divided by the sum of the within-facility variance (from RST and SST) plus the between-facilities variance. Perfect agreement between the two survey teams would result in an ICC of 1.0, and complete randomness would result in an ICC of 0.0. Recognizing that differences in the scope and severity of deficiencies matter as well as the number of deficiencies, we cross-tabulated the deficiencies by the levels of harm cited by the RST and SST and calculated a Kappa statistic. Kappa measures how much the agreement between the teams exceeds the amount expected by chance. Complete agreement would give a Kappa of 1.0, and agreement that is no better than chance would give a Kappa of 0.0.

In order to assess why the conclusions differed, we performed a content analysis (Weber, 1990). Two registered nurse researchers, one with formal training in the survey process, independently reviewed the content of all of the written documentation for each team (researcher field notes, team notes, and meeting transcripts). They then met to resolve any differences in their reviews. In order to ensure confidentiality, we substituted numbers for resident names in these materials, and we restricted access to the materials to the research team.

In order to explore what prompted differences between the teams, the content analysis examined the data that the RST and SST used to reach their conclusions. At issue was whether the teams described different problems or characterized the same problems in different ways. For the same infraction, for example, one team could cite F-tag F221 “no unnecessary physical restraints” and another team could cite F-tag F223 “free from abuse.” If both registered nurse researchers agreed that the RST and SST had cited the facility for separate shortcomings, they categorized the F tag as “distinctly different.”

Results

ICCs

Table 3 shows that the RST and SST cited similar numbers of deficiencies. The ICC for total deficiencies cited by the two teams was 0.87 with a 95% confidence interval of 0.64 to 0.96. Given that values greater than 0.70 indicate good reliability, this is quite high (Kramer & Feinstein, 1981). The RST and SST also cited similar numbers of G+ deficiencies. The ICC was 0.76 with a 95% confidence interval of 0.38 to 0.92. The SST cited more deficiencies than the RST for 8 of the 12 nursing homes, but a paired *t* test failed to reject the hypothesis that the means were the same.

Counts do not fully describe the decisions of the RST and SST. Table 3 also shows that in 49 instances the RST and SST agreed about which regulation was being breached but differed on the

scope and severity. In another 96 instances, the two teams cited distinctly different deficiencies, meaning that they identified different failures to comply with the regulations. The number of distinctly different deficiencies rose with the number of citations. The correlation with RST citations was 0.76 and the correlation with SST citations was 0.89. Both correlations were significantly different from 0 at the 0.01 level.

Kappa Statistics

Table 4 cross-tabulates the findings of the RST and SST, focusing on the levels of harm identified. With 12 facilities and 189 regulations, 2,268 violations were possible. Overall, the level of agreement was moderate, as we estimated a Kappa of 0.57 (Landis & Koch, 1977). Kappa estimates the degree of consensus while controlling for the amount of chance agreement to be expected based on the marginal distributions (Stemler, 2004). Because the RST and SST found no deficiencies most of the time, we needed this control in order to avoid overstating reliability.

In most instances neither team found a violation. The RST found no violations 92.9% of the time, and the SST found no violations 91.8% of the time. The SST agreed with the RST 96.5% of the time.

The teams seldom cited deficiencies entailing no actual harm with potential for minimal harm. The RST gave 11 A, B, or C citations, and the SST gave 9. The similar totals masked considerable disagreement. The SST found no deficiency for 55% of the A–C deficiencies cited by the RST and found a D–F deficiency for 18%. The RST found no deficiency for 11% of the A–C deficiencies cited by the SST and found a D–F deficiency for 56%.

Deficiencies with D–F scope and severity levels, which entail a finding of no actual harm with the potential for more than minimal harm, were the most common citations. Most disagreements also involved these deficiencies. Of the 141 cited by the RST, the SST cited no deficiency for 29%, an A–C deficiency for 4%, a D–F for 63%, and a G–I for 4%. Of the 164 D–F deficiencies cited by the SST, the RST cited no deficiency for 42%, an A–C deficiency for 1%, a D–F deficiency for 54%, and a G–I for 2%. In short, both teams cited no deficiency in a substantial number of the cases in which the other team issued a D–F deficiency.

Deficiencies involving actual harm were uncommon. Even so, the teams differed in their conclusions. The SST cited a D–F deficiency for 4 of the 9 G–I deficiencies cited by the RST and found no breach of the remaining regulation. The RST cited a D–F deficiency for 6 of the 14 G–I deficiencies cited by the SST and found no breach in four instances.

Neither team cited J, K, or L deficiencies, which involve immediate jeopardy for residents.

Table 4. Cross-Tabulations of Deficiencies by Level of Harm

Deficiency	No Deficiency	A–C	D–F	G–I	J–L	RST Totals
No deficiency	2,033	1	69	4	0	2,107
A–C	6	3	2	0	0	11
D–F	41	5	89	6	0	141
G–I	1	0	4	4	0	9
J–L	0	0	0	0	0	0
SST totals	2,081	9	164	14	0	2,268

Notes: RST = regular survey team; SST = simultaneous survey team.

A–C deficiencies find no actual harm with potential for minimal harm. D–F deficiencies find no actual harm with potential for more than minimal harm. G–I deficiencies find actual harm for residents. J–L deficiencies find immediate jeopardy for residents.

Kappa = 0.57.

Content Analysis

As noted above, ICC and Kappa calculations do not fully take into account the differences between the RST and SST. A closer examination of Facility 6 illustrates this. The RST and SST cited the same number of deficiencies, yet there were important differences in their findings. In seven instances the teams disagreed on the scope and severity of the deficiencies, and in six instances the teams cited distinctly different deficiencies. Most of the scope and severity differences were minor, but not all. The RST and SST both identified quality of care deficiencies in the management of pain. The RST assigned an E deficiency, and the SST assigned a G, implying actual harm to residents. The RST and SST both identified deficiencies in the treatment of residents with pressure ulcers. The RST assigned a G deficiency, and the SST assigned a D. In addition, the RST cited three deficiencies that the SST did not: not having an adequate activities program, improperly ordering medications, and not having a backup power supply system. The SST cited four deficiencies that the RST did not: failing to reassess a resident whose condition had changed, not taking adequate care to prevent urinary tract infections, having an overly high medication error rate, and failing to investigate a bruise of unknown origin.

Some disagreements reflected different interpretations of the facts, even though the RST shaped the information that both teams had. For example, in Facility 4 the RST issued a D quality of care citation because the facility failed to follow its own protocol in caring for a resident with a pressure ulcer. The SST identified additional problems with the care provided to this resident and saw similar problems in the care of another resident. The SST issued a G quality of care citation. In another instance, both the RST and SST cited Facility 3 for failures to provide an appropriate accounting of resident funds. The initial citations were both Es, but the SST ultimately assigned an H. The difference appeared to spring

from the conclusion of the SST that at least three items that had been purchased with residents' funds could not be found anywhere in the facility, an issue that the RST did not address. The SST issued an additional H citation for staff treatment of residents and revised its citation for improper accounting of resident funds citation to an H.

Overall, SSTs cited 26 more deficiencies than RSTs, with 18 of these coming from Facilities 4 and 5. For Facility 4, the SST final report identified 10 more deficiencies than the RST final report. The SST issued seven D citations for problems that the RST did not identify or discuss. The SST also issued two citations for problems that the RST combined into one deficiency. After consultation with the regional office, the RST chose not to cite two problems that both teams had identified. In one instance the RST discussed a problem that the SST cited, but decided not to cite the facility. (The RST also cited one deficiency that the SST did not.) For Facility, 5 the SST identified eight more deficiencies than the RST. Five of these deficiencies were due to inconsistencies between the care plan and the care provided that the SST examined and the RST did not. The missing care included activities for one resident, assistance with eating for another resident, protective booties for a resident at risk for pressure ulcers, a contracture boot for another resident, and range of motion therapy for yet another resident.

Our observers noted a striking difference in how the teams tracked medication administration. In Facility 4 the RST focused on one of the medications given to a resident, but the SST made notes on all of the resident's medications. The two teams found similar numbers of errors, but the SST calculated a much lower error rate because the denominator was much larger. The RST gave an E deficiency to Facility 4 for medication administration; the SST did not.

In their discussions, SST members critiqued the RST fairly regularly. For example, the SST notes for Facility 6 included comments that, "I would have followed up more on [the] broken thermostat," and "I would have knocked and checked" to see if a resident scheduled for an interview was in her room with the door closed. The SST notes for Facility 11 noted that there were unasked questions about a "resident being left alone on toilet and orthostatic hypotension" and "fall investigation." Additionally, some teams identified deficiencies by "running through the regulations." Other teams identified deficiencies by running through the leader's concerns.

Yet attributing these differences to the teams obscures the important roles of other staff.

Teams discuss concerns with their regional managers and quality improvement coordinators several times during a survey. Furthermore, teams discuss their findings with these administrative staff following their decision-making meeting. Again, this process has both strengths and weaknesses. On the one hand,

the experience of regional managers and quality improvement coordinators allows them to assist more junior surveyors by providing guidance and putting information into perspective. On the other hand, most regional managers and quality improvement coordinators are not on site and so provide guidance without seeing the evidence firsthand. Analyses of meeting and field notes indicated that the number of changes between the initial and final reports ranged from 0 to 14 per team.

Several comments indicated that regional managers had a significant influence on the survey process. For example, some regional managers did not encourage surveyors to write deficiencies for paperwork violations unless there were concomitant care problems. In addition, some surveyors noted that their regional managers instructed them that hand washing had to be a huge issue before they should cite it. One team commented that their regional manager would never let them go into an extended survey for a particular F tag. Some teams made a point of staying for the first meal after entering the facility, and others did not. Some teams were very methodical in their decision-making style, going in order through the regulations, whereas others discussed concerns according to their priority or in top-of-mind order. In short, different teams used different processes.

An important finding was that teams differed in assessments of scope and severity for the same resident care issue. Our content analysis identified several instances in which there were no clear right or wrong assessments of scope and severity. When teams disagreed on the scope and severity, we could trace these differences to differences in interpretations of the regulations and of the interventions provided by the facility.

An example dealing with pressure ulcer prevention and healing illustrates the difficulty with scope and severity determinations. The *Facility Guide to OBRA Regulations, and Interpretive Guidelines and the LTC Survey Process* offers the following guidance:

A determination that development of a pressure sore was unavoidable may be made only if routine preventive and daily care was provided. Routine preventive care means turning and proper positioning, application of pressure reduction or relief devices, providing good skin care, (i.e., keeping the skin clean, instituting measures to reduce excessive moisture), providing clean and dry bed linens, and maintaining adequate nutrition and hydration as possible. (p. 22)

Their notes indicated that surveyors seldom had difficulty in determining whether the facility identified the resident as being at risk. But surveyors looking at the same evidence disagreed on whether the facility interventions were aggressive enough or

whether the facility tried enough different interventions. Surveyors scrutinized the data collected and took their decisions very seriously but had differing perceptions of when a facility had done enough.

Discussion

Even though the teams examined the same data, they often differed in the number, scope, and severity of deficiencies cited. The teams also routinely assigned different F tags when they cited facilities. In short, the teams generated substantially different surveys from the same facts. Yet abstracting from the details of the surveys, the teams painted very similar pictures of facilities' overall compliance with federal regulations.

These data support two very different interpretations. One stresses the variability of the survey process; the other stresses its global consistency. The variability interpretation notes that the two survey teams often reached different conclusions about whether a deficiency existed, what regulation had been breached, the scope of the deficiency, or the severity of the deficiency. These differences, furthermore, might well have consequences. The penalties imposed by the survey agency, the career prospects of facility managers, and the responses of consumers are likely to be different for a nursing home that gets 7 D deficiencies than for a nursing home that gets 12 D deficiencies and 1 G.

The variability of the survey process reduces its value to nursing home managers, who should be the primary users of its detailed findings. The same process can draw no deficiencies from one survey team and multiple deficiencies from another. As a result, nursing home administrators and directors of nursing cannot be confident that a good survey means that a process works well. Nor can administrators and directors of nursing be confident that genuine improvements in care will result in a better survey if the next team relies on different interpretations of the regulations and what constitutes having done enough. Speaking for a number of her peers, one director of nursing described the survey process as "demoralizing." Improvement efforts are inhibited by a survey process that falls short of systematic, replicable data gathering and analysis (Schnelle, Osterweil, & Simmons, 2005).

The variability of the survey also reduces its value to regulators and policy makers. The inspection is supposed to provide assurance that a nursing home is in substantial compliance with federal and state regulations, either at the time of the inspection or after completion of a plan to correct problems. An unreliable survey process may mean that nursing homes that do not actually meet federal or state standards will be eligible for Medicare and Medicaid payments. The many disagreements of these two teams about whether a regulation had been breached, which regulation had been breached, and

how serious the breach was cannot make federal or state officials comfortable.

The variability perspective would also note that we had designed the structure of this study in order to exclude some forms of variation. Had they not been constrained to look at the data assembled by the RST, the members of the SST might well have gathered different facts and identified different problems. Indeed, comments to this effect by members of the SST were routine. It is likely that this study understates the variability of the survey process.

Yet these data also highlight the overall consistency of the survey results. The total numbers of deficiencies and the number of G+ deficiencies cited by the RST and SST were quite similar. If consumers rely on the total number of deficiencies or the number of high-level deficiencies as measures of quality, our results suggest that consumers should view surveys as highly reliable. We do not know how consumers use nursing home survey results, but their structure suggests that consumers should use them as part of a broader assessment process. Surveys may not reflect current conditions in a nursing home and should be used with care, just like any other measure.

Viewed at a macro level, this study suggests that, given the same data, the two teams reached very similar conclusions. Viewed at a micro level, this study suggests just the opposite. Although state survey agencies and consumers may feel comfortable focusing on macro results, managers must make decisions at the micro level, and their concerns about reliability weaken the credibility of the survey process. In order to reduce the variability of survey results, changes in the survey process and in the training of surveyors warrant consideration. The CMS trial of the Quality Indicator Survey appears to be a promising initiative (CMS, 2004). This five-state experiment enhances training, sampling, and decision support software to make surveys more structured.

This article suggests that surveyors need more specific criteria, in the form of decision-making algorithms, to reduce the influence of individual perceptions. These findings concur with other evaluations of survey consistency (GAO, 2003b; Office of the Inspector General, 2003, 2004). CMS has begun a process of developing and evaluating clearer guidelines for surveyors. Our findings support that effort.

These results also suggest that continued efforts to standardize training and decision rules are important. Especially at the state level, common understandings of what constitutes a breach of the regulations should reduce the angst of the industry and increase the confidence of regulators and the public. In assigning the number, scope, and severity of deficiencies, consistency is of primary importance.

One should not overlook the limitations of this study. It applies to one state with a specific administrative structure. Moreover, the sample used in this study was not large. And, although they were

randomly selected and generated data comparable to statewide averages, we cannot guarantee that the facilities or survey teams were representative of Kansas. The results should not be generalized to other states. Furthermore, this study eliminated differences in the information collected. As a result, the differences reported here were entirely due to differences in interpretation. As we noted above, these results seem likely to understate the variability of survey results in the wild.

It is important to remember that the survey process is designed to measure compliance with federal regulations. It is tempting to infer that a survey with few deficiencies identifies a good facility and a survey with many deficiencies identifies a bad one. Indeed, numerous research studies and consumer guides do exactly that (e.g., Castle, 2000; Castle & Mor, 1998; Harrington, O'Meara, Kitchener, Simon, & Schnelle, 2003). Yet, as one surveyor noted, "The number of deficiencies is not a good quality indicator for whether I would put my mom somewhere or not. You know it relates back to what was the scope and severity of those deficiencies and what were those deficiencies really about" Our results suggest that the survey process is only moderately reliable in describing the scope and severity of nursing home deficiencies. Given that compliance with federal regulations may well have changed since the survey was completed, consumers should use the survey results with care.

Many states and CMS rely on public reporting of survey results as a spur to better nursing home care. Indeed, this appears to represent an important de facto shift from a policy of pure deterrence to a policy of deterrence plus transparency (Chou, 2002). Consumers evidently seek this information. Yahoo! reports that "Nursing Home Compare" is the nation's second most popular nursing home care site and is one of the most frequently visited sections of the Medicare Web site (Office of the Inspector General, 2004; Yahoo! Health Directory, 2005). As a result, the reliability of nursing home surveys becomes an even more visible public policy issue. Survey results will have the greatest impact on nursing home quality if consumers and the industry believe that deficiencies are valid, reliable measures of quality. This belief will be undercut by variations in the number, scope, and severity of deficiencies when the facts are held constant. The appropriate policy response is to acknowledge these variations and address them by clarifying definitions and interpretations, by improving training, and by providing feedback to surveyors. Simultaneous surveys like the ones reported here should become standard features of survey agencies. Using simultaneous surveys as a calibration tool is clearly feasible.

References


Arling, G., Kane, R. L., Lewis, T., & Mueller, C. (2005). Future development of nursing home quality indicators. *The Gerontologist*, 45, 147-156.

- Castle, N. (2000). Nursing homes increasing and decreasing restraint use. *Medical Care*, 38, 1154-1163.
- Castle, N., & Mor, V. (1998). The use of physical restraints in nursing homes: A review of the literature since the Nursing Home Reform Act. *Medical Care Research and Review*, 55, 139-170.
- Centers for Medicare and Medicaid Services. (2004). *Action plan for nursing home quality*. Retrieved August 25, 2005, from <http://www.cms.hhs.gov/quality/nhq/NHActionPlan.pdf>
- Centers for Medicare and Medicaid Services. (2005). *The official U.S. government site for people with Medicare*. Retrieved June 15, 2005, from <http://www.medicare.gov>
- Chou, S. (2002). Asymmetric information, ownership and quality of care: An empirical analysis of nursing homes. *Journal of Health Economics*, 21, 293-311.
- Fielding, N. G., & Fielding, J. L. (1986). *Linking data: The articulation of qualitative and quantitative methods in social research*. Beverly Hills: Sage.
- Forbes-Thompson, S., Dunton, N., Gajewski, B., Wrona, M., Becker, A., Chapin, R., et al. (2003). *Kansas nursing facility project evaluation*. Kansas City: Kansas Department of Aging.
- Government Accountability Office. (2003a). *Nursing home deficiency trends and survey and certification process consistency*. Retrieved October 11, 2006, from <http://oig.hhs.gov/oei/reports/oei-02-01-00600.pdf>
- Government Accountability Office. (2003b). *Nursing home quality: Prevalence of serious problems, while declining, reinforces importance of enhanced oversight* (Report No. GAO-03-561). Retrieved June 15, 2005, from <http://www.gao.gov/new.items/d031016t.pdf>
- Government Accountability Office. (2004). *Prevalence of serious quality problems remains unacceptably high, despite some decline* (Report No. GAO-03-1016T). Retrieved June 15, 2005, from <http://www.gao.gov/new.items/d031016t.pdf>
- Grassley, C. R. & Letter to Mark McClellan. (July 7, 2004). Retrieved October 11, 2006 from <http://www.canhr.org/news/GrassleyLetter.html#LetterToCMS>.
- Harrington, C., O'Meara, J., Kitchener, M., Simon, L. P., & Schnelle, J. F. (2003). Designing a report card for nursing facilities: What information is needed and why. *The Gerontologist*, 43(Special Issue II), 47-57.
- Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24, 602-611.
- Kansas Department on Aging. (2004). *Sunflower connection*. Retrieved May 8, 2006, from http://www.agingkansas.org/kdoa/lce/facts_newsletter/jan_2004_final.pdf
- Kramer, M., & Feinstein, A. (1981). The biostatistics of concordance. *Clinical Pharmacology and Therapeutics*, 29, 111-123.
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Mullan, J. T., & Harrington, C. (2001). Nursing home deficiencies in the United States. *Research on Aging*, 23, 503-531.
- Office of the Inspector General. (2003). *Nursing home deficiency trends and survey and certification process consistency*. Retrieved February 24, 2006, from <http://oig.hhs.gov/oei/reports/oei-02-01-00600.pdf>
- Office of the Inspector General. (2004). *Inspection results on Nursing Home Compare: Completeness and accuracy*. Retrieved August 18, 2005, from <http://www.canhr.org/pdfs/oei-01-03-00130.pdf>
- Office of the Inspector General. (2005). *Nursing home enforcement: The use of civil monetary penalties*. Retrieved February 24, 2006, from <http://oig.hhs.gov/oei/reports/oei-06-02-00720.pdf>
- Ousley, M. K. (2001). *Testimony before the Subcommittee on Health of the House Committee on Ways and Means, March 15, 2001*. Retrieved June 15, 2005, from <http://waysandmeans.house.gov/legacy/health/107cong/3-15-01/3-15ous.htm>
- Pate, T. (2005). *The Facility Guide to OBRA Regulations and the Long-Term Care Survey Process*. Miamisburg, OH: Med-Pass.
- Schnelle, J. F., Osterweil, D., & Simmons, S. F. (2005). Improving the quality of nursing home care and medical-record accuracy with direct observational technologies. *The Gerontologist*, 45, 576-582.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved March 9, 2006, from <http://PAREonline.net/getvn.asp?v=9&n=4>
- Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage.
- Yahoo! Health Directory. (2005). *Site listings by popularity*. Retrieved August 17, 2005, from http://dir.yahoo.com/Health/Long_Term_Care/Nursing_Home_Care/

Received November 11, 2005

Accepted June 13, 2006

Decision Editor: Linda S. Noelker, PhD



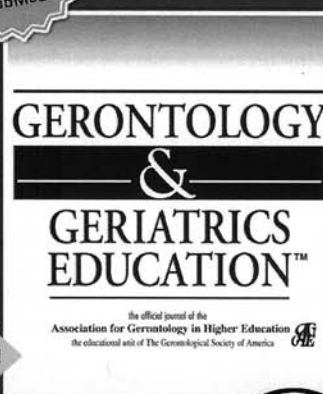
→ **AGHE INSTITUTIONAL MEMBERS:**
YOUR AGHE REPRESENTATIVE RECEIVES THE JOURNAL AS A MEMBER BENEFIT!
Details inside.

A USEFUL RESOURCE for staying informed about critical
EDUCATIONAL ISSUES related for AGING and the aged!

GERONTOLOGY & GERIATRICS EDUCATION™

*the official journal of the Association for Gerontology in Higher Education,
the educational unit of The Gerontological Society of America*

Included in
Index Medicus,
MEDLINE, and
PubMed!



Now the
**official
journal
of AGHE!**

Editor: Pearl M. Mosher-Ashley, PhD
Professor, Department of Psychology, Worcester State College, Worcester, Massachusetts

Managing Editor: Judith Gardner Ainlay, MSW
Coordinator, Consortium Gerontology Studies Program, Colleges of Worcester Consortium, Inc., Worcester, Massachusetts

EDITORIAL ADDRESS: Colleges of Worcester Consortium, 484 Main Street, Suite 500, Worcester, MA 01608;
Tel: (508) 754-6829; Fax: (508) 797-0069; E-mail: jainlay@cowc.org

SITE-WIDE ONLINE ACCESS with every library print subscription!

■ about the journal

Gerontology & Geriatrics Education, now, the official journal of the Association for Gerontology in Higher Education, the educational unit of The Gerontological Society of America, is geared toward the exchange of information related to research, curriculum development, course and program evaluation, classroom and practice innovation, and other topics with educational implications for gerontology and geriatrics. It is designed to appeal to a broad range of students, teachers, practitioners, administrators, and policy makers and is dedicated to improving awareness of best practices and resources for gerontologists and gerontology/geriatrics educators.

Articles in this highly regarded journal cover research results, observations, evaluations, theoretical discussions, and recommendations related to gerontology and geriatric course work, practice placements, and curriculum design and implementation. The journal also:

- reports innovations in the teaching of gerontology and geriatrics at the undergraduate, graduate, and postgraduate levels, and in continuing education, paraprofessional, and public education
- discusses issues, methods, and materials in the training and supervision of gerontology and geriatrics educators, researchers, and practitioners in all levels of health care, including long-term settings
- explores new roles for gerontology and geriatrics educators in public and private programs and in community, medical, and academic institutions, including corporate and industrial settings
- communicates new methods for developing gerontology and geriatrics educational programs in academic, medical, and applied settings—and new approaches for supporting such educational programs
- reports on research and evaluation that has been carried out concerning the development of individual courses, concentrations, majors, or degree and certificate programs in gerontology and geriatrics in any discipline
- includes research and discussion on aging-related issues that have relevance for gerontology and geriatrics educators

■ noteworthy reviews

"INTERESTING AND WELL-WRITTEN. . . Provides insights into what works and new ideas on how to evaluate programs. I look forward to receiving each issue because I find useful articles that help me teach health professions trainees as well as gerontology students."

—Kathryn Hyer, PhD, Associate Professor,
University of South Florida

"A LEADING SOURCE FOR HIGH-QUALITY INTERDISCIPLINARY SCHOLARSHIP in the art and science of aging education. . . THE GOLD STANDARD for open, scholarly discourse on gerontological pedagogy. It is difficult, if not impossible, to identify another journal of equal quality that successfully brings together on a continuing basis such a diverse set of leading thinkers committed to infusing content in aging in both professional and disciplinary curricula."

—Lenard W. Kaye, DSW, Professor, School of Social Work
and Director, UMaine Center on Aging,
University of Maine

"AN INVALUABLE RESOURCE for multidisciplinary educators committed to gerontology and geriatrics."

—Nancy R. Hooyman, PhD, Professor and Dean Emerita,
University of Washington School of Social Work



10 Alice Street • Binghamton, NY 13904-1580
Tel: 1-800-429-6784 • Fax: 1-800-895-0582
Outside US/Canada Tel: (607) 722-5857 • Fax: (607) 771-0012
E-mail: orders@HaworthPress.com • Web: www.HaworthPress.com

Gerontology & Geriatrics Education™, its trademarks and copyrights are owned entirely by and published by The Haworth Press, Inc.