Assessment of level of care: Implications of interrater reliability on health policy

David H. Gustafson, Richard Van Koningsveld, and Robert W. Peterson

In Wisconsin, level-of-care assessments are used to set Medicaid reimbursement and determine nursing home eligibility. This study examined three methods of assessing level of care: 1) the Wisconsin quality assurance project (QAP) method, based on observations of patients, patient records, and staff interviews; 2) the Wisconsin standard (STD) method, based primarily on a clinical record review; and, 3) an adaptation of New York's "DMS-I," a checklist with numerical weights used to set level of care. Results address interrater reliability, the agreement between assessments by research teams and actual levels of care set by the State, and the implications that agreement has for reimbursement.

Introduction

Inspection of care (IOC) is a federally mandated annual process of reviewing nursing home Medicaid residents to ensure appropriate placement in a facility that is staffed and equipped to provide that care. In some States, the IOC process is used not only to assess appropriateness of placement and quality of care, but also to review the level of care upon which reimbursement is based. Wisconsin, for example, uses the IOC for all three purposes. At the time of the study reported here, Wisconsin's IOC determined whether a nursing home resident living in an intermediate care facility (ICF) could be appropriately cared for or actually required placements in a skilled nursing facility (SNF). The Wisconsin IOC also was used to identify quality of care problems, which were then considered as part of the annual survey and certification process. Finally, the Wisconsin IOC classified residents into one skilled level of care, SNF, and four intermediate levels, ICF-1, ICF-2, ICF-3, and ICF-4 (ICF-1 the highest level and ICF-4 the lowest level). The individual resident classifications establish the reimbursement rate for each individual and, in turn, the total Medicaid reimbursement to the nursing home. In 1982, average daily reimbursement in Wisconsin ranged from $36.21 for an SNF resident to $19.73 for an ICF-4 resident.

Although States' use of the IOC process varies, the ability to determine the appropriate level of care for a person is important for both the effective and equitable distribution of Medicaid funds, as well as for quality assurance. This article reports the results of a study that tested the interrater reliability of three methods of determining level of care. Three questions arose from the issue of determining level of care:

1) How reliably can level of care be assessed;
2) What impact does the level-of-care process have on reimbursement; and
3) To what extent does level-of-care assessment approximate patient needs?

This study was supported by Grant No. 18-P-97664/5 from the Health Care Financing Administration.

Reprint requests: David H. Gustafson, Center for Health Systems Research and Analysis, University of Wisconsin, 1225 Observatory Drive, Madison, WI 53706.

This study did not directly examine the validity of the level-of-care concept. Several studies, however, have examined the level-of-care concept. Some have examined the characteristics that must be possessed by an effective level-of-care assessment. Willemain (1980) examined three strategies for using level of care in nursing home reimbursement. One strategy involved placing a resident in one of two levels of care (SNF or ICF). A second strategy placed the resident along a continuous level-of-care scale. A third strategy assigned each nursing home a facility-specific reimbursement rate that was not a function of the level of care of individual residents. A computer simulation was used to examine the impact each strategy would have on reimbursement error. In this simulation Willemain varied: 1) the case mix found in the facility; 2) the relative concern for over- and underpayment; and 3) the uncertainty involved in assessing level of care. He found that when assessment uncertainty was low or moderate, placing patients on a continuum was the best use of level of care. When assessment uncertainty was high, the placement of patients along a continuum of care was the least effective alternative in two-thirds of the cases and never the best alternative. An unanswered question from the Willemain study was "how much uncertainty is involved in level-of-care assessment?"

Our study examines assessment uncertainty in terms of the interrater reliability of three currently operational level-of-care assessment processes. Other investigations have examined level-of-care assessment processes. Bishop et al., (1980), argues that an annual level-of-care assessment (currently required by the Federal Government for Medicaid residents of nursing homes) is problematic when assessing care of patients with stable chronic conditions because there are no indicators to clearly predict the future course of the chronic condition. The instability of a resident's condition and a large number of levels of care increases the need for a sensitive and reliable assessment process (Willemain, 1980).

Bishop also pointed out that many level-of-care assessment tools are heavily dependent on data from the patient record. So "when the quality of record keeping is low, the data abstracted by the assessment instrument will not be valid." One of the level-of-care assessment methods compared in our article stresses

direct observation of residents and their interaction with staff as a means of overcoming record inadequacies.

For level of care to be a reasonable concept, different people looking at the same patient must agree on the level of care for that patient. Sager (1979) found that perceptions of level of care vary widely among long-term care professionals. Using data abstracted from medical records and nursing staff interviews for a sample of patients in Baltimore nursing homes, Kane et al., (1980) found that 20 percent of the patients should have been classified SNF while, in fact, 80 percent of the beds were classified SNF. There are many possible explanations for this discrepancy. One is that a patient who enters a nursing home as an SNF patient may improve to ICF status, but not be sent to an ICF home because of the potential "trauma of transfer." Second, the data suggest there may be a wide discrepancy between supply of and need for different types of nursing home beds. Although this discrepancy between supply and demand for beds may be a problem in some States, our study took place in a State where patients can be changed from one level to another without changing homes. So, supply can be removed as a potential explanation of the discrepancy issue if residents are assigned a level of care that is too high.

Some people suggest that global[1] level-of-care judgments should be replaced by decision rules under which basic data are collected and an equation used to calculate level of care. Cavaiola and Young (1980), developed a long-term care decision rule. Using nonlinear multiple regression techniques, they reduced a set of 37 patient descriptors (e.g., level of bladder function) to a set of 12 descriptors that best predicted the level of care assigned to 306 patients. Level-of-care assignments for those patients (SNF, Intermediate A, and Intermediate B) had been determined in a previous study by McKnight (1967). The best model developed by Cavaiola and Young correctly predicted 81 percent of the Intermediate B levels. The reliability of the initial level-of-care assignments was not reported.

Greene and Monahan (1981) interviewed nursing home staff to collect 22 items designed to measure impairment status of 292 nursing home residents in Arizona. A multiple discriminate analysis model was used to predict level-of-care decisions on these patients. The model agreed with the nursing home's level-of-care assignments 70 percent of the time. Predicted level of care was lower than actual in 22 percent of the cases.

Part of our study also investigated the extent to which an independent assessment of level of care agreed with the State's assessment of level of care. Two differences existed. First, our study used multidisciplinary teams instead of a discriminate analysis to make the independent assessment. Second, the setting for this study was a State in which Medicaid reimbursement is directly tied to level-of-care assignment.

Foley and Schneider (1980) examined six other decision rules for assigning patients to a level of care. Profiles were developed for 690 patients in New York. The profiles contained sufficient information to apply each of the decision rules. Patients were classified as SNF or ICF by each decision rule. The agreement rate between pairs of decision rules ranged from 38 percent to 91 percent. Foley and Schneider concluded that the level of care assigned to a patient depended upon the State in which the patient resides and the decision rule applied. This is rather disconcerting in terms of the consistency of Medicaid eligibility determinations across States. Foley and Schneider also point out that most States do not use formal decision rules that disaggregate judgments into components. Rather, the majority of States use a more global approach, where the surveyor makes the level-of-care decision without explicitly following a set of decision scales, e.g., the Wisconsin IOC. A major question is whether these more global approaches lead to improved reliability in assignment determination. Our study examines whether interrater reliability would be higher with explicit level-of-care decision rules than with the general guidelines used in most States. This article addresses these issues.

The purposes of the study reported here were to:

1) Evaluate the reliability of three level-of-care assessment methods, although the evaluation of two of those methods had a higher priority than the third.
2) Evaluate the impact of the two Wisconsin level-of-care assessment methods on reimbursement to nursing homes.

Level-of-care methods

Three methods of assessing level of care were examined in this study: 1) the State of Wisconsin's standard (STD) process; 2) Wisconsin's Quality Assurance Project (QAP) method which is an experimental procedure,[2] and 3) a decision rule which is an adaptation of the DMS-I instrument used in the State of New York. The Wisconsin STD method employs both a nurse and a social worker to assess level of care, but the nurse makes the final level-of-care determination. The process includes a review of the resident's clinical record augmented by an unstructured patient review done at the surveyor's discretion. The QAP was designed to provide a more indepth review of a sample of residents in each facility. The process is carried out by a nurse and social worker who work as a team in gathering information and jointly arriving at a level-of-care determination. The process includes the

[1] We use the term global to mean that the surveyor decides what level of care is appropriate, rather than deciding how the resident scores on a set of criteria that are then aggregated by a formula to arrive at a level-of-care judgment.

[2] More detailed information on the QAP method can be obtained by contacting Ms. Joy Stewart, Division of Health, Wisconsin Department of Health & Social Service, 1 West Wilson, Madison, Wisconsin, 53701.

observation and interview of the patient, an interview with facility staff, and a clinical record review. The QAP method is designed to place more emphasis on the determination of quality of care, while also making determinations of appropriate level of care.

While there are important differences between the Wisconsin STD and the QAP methods, both share the characteristic of being processes in which observation and review are relatively unstructured. Given the research mentioned above, we felt it was important to contrast this unstructured approach with an example of a structured decision rule. Therefore, the third method investigated in this study is an adaptation of the DMS-I process used in New York.

In New York, the hospital or nursing home staff completes the DMS-I and calculates the score for a patient. The State reviews the level-of-care assignment. In contrast, both Wisconsin methods assign responsibility for determining level of care with the State field surveyor.

Assessment teams: Composition and training

Three teams were trained to conduct level-of-care assessments using the QAP, STD, and DMS-I methods. Each team was composed of a registered nurse and a social worker. Each team member had experience in long-term care, although only two (both nurses) had been State surveyors prior to this study.

The nurse in charge of level-of-care training for the State of Wisconsin provided the study teams with training on the STD and QAP methods identical to that given State surveyors. Research personnel were trained to complete the DMS-I. Training was conducted through consultation with New York State personnel and by using training materials provided by New York.

Procedures

All teams visited seven nursing homes for the main study. An eighth home was visited in a "pilot" study (described below). Thirty patients were reviewed in each of the seven (main study) homes. Each team was paired with another team for the purpose of rating the level of care for 10 patients at each home. Assignments of patients to teams were made randomly using a Latin Square design. Two teams assessed the level of care for each patient. A summary of the team assignment design is shown in Table 1. The teams employed the Wisconsin STD or QAP to match the method actually used by the State surveyors in the previous inspection of care (IOC). Thus, the teams used the Wisconsin STD method in three homes where the State had used the STD method, and the QAP method in three homes where the State had used the QAP method. The DMS-I method was used in the seventh home as part of a reduced study intended to obtain a preliminary reliability assessment of a decision rule for assigning level of care.

Table 1
Procedure used to assign study teams to nursing home residents

	Team		
Resident	A	B	C
1	X	X	
2	X		X
3		X	X
...			
...			
...			
28	X	X	
29	X		X
30		X	X

Homes were selected on the following criteria:[3] 1) the home had its levels of care examined by an IOC team within 3 weeks prior to the teams' visit; 2) the home had residents in the SNF level and some residents in ICF levels 1-3; 3) the home had a minimum of 100 beds; and 4) the home did not have a high proportion of mentally retarded or psychiatric residents.

Results[4]

Proportion of agreement

Our first, and most conservative test, was to determine whether agreement rates were significantly better than chance alone. Two-dimensional contingency tables such as Table 2 were prepared for each team combination (i.e., AB, AC, BC) within each method. A Kappa statistic (Tinsley and Weiss, 1975) was used to test the hypothesis of no difference from chance. That hypothesis was rejected at the .001 level for all team pairs except teams AB in the standard (STD) method where the hypothesis can be rejected at the .10 level. Thus, there is evidence of some agreement between assessments.

The data were combined across teams to test the hypothesis that a rating method's (i.e., STD and QAP) level of agreement was no better than chance. Table 2 presents that data for the QAP and STD methods. The hypothesis of "agreement no better than chance" can again be rejected at the .001 level for QAP and .01 level for STD method.

Percent agreement scores were calculated for each of the methods (QAP, STD, DMS-1, and the Pilot test.) The purpose of this calculation was to examine two level-of-care agreement rates (SNF vs. ICF) and compare to five level-of-care agreement rates (SNF vs. ICF-1, ICF-2, ICF-3, and ICF-4.) Table 3 presents the results.

[3]One home had residents assigned only to the SNF and ICF-1 levels. However, in this home it turned out that raters assessed those residents for SNF and ICF-1, 2, and 3 levels.

[4]Tables presenting raw data are available upon request from David H. Gustafson, Center for Health Systems Reserach and Analysis, University of Wisconsin, 1225 Observatory Drive, Madison, WI 53706.

Table 2A

Agreement across any pair of research teams on level-of-care assessments using the Quality Assurance Project (QAP) method

						Total residents
	31	3		1		35
	5	9	3			17
K = .53	2	6	11	3		22
Z = 9.14*			4	8	1	13
P ≤ .001			1	2	0	3
Total residents	38	18	19	13	1	90

Table 2B

Agreement across any pair of research teams on level-of-care assessments using the standard (STD) method

						Total residents
	15	3	2	1		21
	3	9	4	3		19
K = .42		4	6	3	2	15
Z = 2.74*		1	5	14	9	29
P ≤ .01				3	3	6
Total residents	18	17	17	24	14	90

Table 3
Percent agreement of paired level-of-care judgments by method

Assessment	QAP [1]	STD [2]	DMS-1 [3]	Pilot
		Percent		
SNF	84	79	86	58
ICF	83	89	87	87
ICF-1	54	53	NA	50
ICF-2	55	31	NA	27
ICF-3	65	53	NA	41
ICF-4	0	45	NA	52

[1]Wisconsin's Quality Assurance Project method.
[2]The Wisconsin standard method.
[3]An adaptation of New York's DMS-1, a checklist with numerical weights.
NOTE: NA = not applicable.

One would expect more reliable survey processes to generate higher agreement rates across care levels. As Table 3 indicates, the methods tested yield approximately the same amount of agreement. However, agreement rates are substantially less within the four ICF categories than between SNF and ICF levels. Conclusions drawn about the relative method reliabilities assume "equal" capability with the methods by our research teams. The teams received training and experience until they felt "comfortable" in making level-of-care judgments with each method. Whether significant increases in the overall agreement rates could be expected from additional experience remains an open question.

The second comparison is between assessments made by our teams and assignments made by the surveyors. Except in the case of DMS-1, the research teams and State teams in a particular home used the same method (STD or QAP) to set levels of care. An analysis of agreement using the Kappa statistic indicated that the hypothesis of agreement no better than chance could be rejected at the .001 level. Table 4 summarizes the next comparison between the State's level-of-care assignment and the research teams' level-of-care assessments. The left part of the table depicts the frequency with which State assignments differed from research teams' assessments when five levels of care were used.

The table also depicts the frequency of level-of-care differences when only two levels of care (SNF and ICF) are considered. Table 4 summarizes the data in terms of percent agreement, higher assessment, and lower assessment by the State surveyors. The results demonstrate a strong tendency on the research teams' part to rate at a lower level of care than the State surveyors, although that tendency is more pronounced in the STD cases (54 percent of the time) than in the QAP cases (42 percent). In the pilot home where the previous year's level of care was not available to the research nurse, the tendency to assign lower levels of care than the State was much greater, 70 percent. The differences become smaller when only two levels of care are used.

Extent of disagreement in inspection of care

The second measure of reliability is the average absolute difference in reimbursement implied by the care levels assigned to each resident by the two research teams. Perfect agreement between two raters over all residents would yield no difference in reimbursement; and as the reliability between the raters decreases, the mean disparity increases.

For this analysis, the average, absolute difference in reimbursement was computed for each pair of teams over the residents they reviewed in common. The data are presented in Table 5. The results of a one-way analysis of variance (based on the data in Table 4) indicate no significant differences between the teams within the methods (P = .24 with STD; P = .91 with QAP).

Table 4
Variance between assessments of level of care by research teams and State survey assignments

Comparison	5 care levels			Skilled nursing facility and intermediate care facility only			
	QAP [1]	STD [2]	Pilot	QAP [1]	STD [2]	DMS-1 [3]	Pilot
Number of residents	90	90	47	90	90	29	47
			Percent Agreement				
State higher	42	54	70	13	21	25	62
State agrees	53	43	27	81	78	72	38
State lower	5	3	3	6	1	3	0

[1]Wisconsin's Quality Assurance Project method.
[2]The Wisconsin standard method.
[3]An adaptation of New York's DMS-1, a checklist with numerical weights.

Table 5
Absolute disparity in reimbursement between teams

Teams	Number of residents	STD [1] method Mean	STD [1] method Variance	Number of residents	QAP [2] method Mean	QAP [2] method Variance
Overall		$2.29	13.450		$1.53	8.543
A-B	30	1.875	11.750	30	1.398	3.742
A-C	30	1.775	10.135	30	1.708	9.779
B-C	30	3.223	18.463	30	1.481	7.108

[1]The Wisconsin standard method.
[2]Wisconsin's Quality Assurance Project method.

A 95 percent confidence interval for the overall mean disparity in reimbursement rate between the research teams is: STD method: ($1.51 to $3.07) and averages $2.29; QAP method: ($0.91 to $2.15) and averages $1.53.

The overall mean difference between research teams in the STD method is $2.29 and under the QAP method $1.53. The $.76 difference between methods would be significant at alpha = .27.

The variances in Table 5 tend to be larger under the STD method than under QAP. An F-test for homogeneity of variance between methods is F = (13.45)/(8.543) = 1.574. The probability of observing an F-ratio as large or larger by chance alone = 0.018.

So, the average difference in reimbursement rates resulting from assessments by the STD and QAP methods was not significant. But, this result demonstrates that when disagreements between teams do occur, the differences are significantly greater in the STD method.

We then examined the difference between the reimbursement rate resulting from assessments by research teams and the reimbursement rate resulting from assignments actually set by State surveyors. Remember that the level-of-care assessment method employed by the research teams was the same as used by the State surveyors for those residents (STD in three homes and QAP in three homes.) We interpret a small difference between research and State teams to be desirable. The results of this comparison are shown in Table 6.

A 95 percent confidence interval for the overall average bias in reimbursement rate between the research teams and the State assignments is: STD Method: (−$3.87 to −$2.52) and averages −$3.197; QAP Method: (−$1.56 to −$0.49) and averages −$1.026.

Although both of these confidence limits suggest significantly lower reimbursement rates than are current by State assignments, the mean bias is significantly higher when both the research teams and the State teams used the STD method.

The variances in Table 6 also tend to be larger under the STD method than under QAP. The F-test for homogeneity of variance between methods is F = (20.49)/(12.99) = 1.574. The probability of observing an F-ratio as large or larger by chance alone = 0.003 (F(87,87,1.574)). This implies that the research and State teams used the QAP method more reliably than they did the STD method.

Table 6
Disparity in reimbursement between teams and State

Teams	Number of residents	STD [1] method Mean	STD [1] method Variance	Number of residents	QAP [2] method Mean	QAP [2] method Variance
Overall		$−3.197	20.486		$−1.026	12.987
A-S	60	−3.397	16.287	60	−1.614	16.058
B-S	60	−2.026	16.521	60	−0.829	9.049
C-S	60	−4.168	28.649	60	−0.636	13.853

[1]The Wisconsin standard method.
[2]Wisconsin's Quality Assurance Project method.

The results of a one-way analysis of variance for each method based on the data of Table 6 is given in Table 7. These analyses examine the consistency of the bias between teams within the methods.

The F-ratio of 3.45 suggests significant (P = .03) differences in the average bias between the teams under the STD method, but not under the QAP method (P = .29). This suggests a lack of consistency between the research teams using the STD method in how much they disagreed with the State's level-of-care assignments for the residents they reviewed.

Table 7

Analysis of variance on disparity between teams within each method

Source	DF [1]	STD [2] method Mean SQ	F	QAP [3] method Mean SQ	F
Total	179	21.046		13.022	
Between teams	2	70.670	3.45	16.123	1.24
Within teams	177	20.486	A = .03	12.987	A = .29

[1]DF = degrees of freedom.
[2]The Wisconsin standard method.
[3]Wisconsin's Quality Assurance Project method.

Net bias

Tables 8 and 9 compare the frequency distribution of care levels assessments made for residents by each pair of teams. Table 8 summarizes this information for the STD process, and Table 9 for the QAP process. The purpose of this analysis was to see if differences between research teams tend to disappear when level-of-care ratings are examined in the aggregate. A CHI-square analysis testing the equality of these frequency distributions indicates two statistically significant differences among the 30 paired comparisons.

Table 8

Distribution of assessment-of-care levels by teams made for 30 residents: Standard (STD) method

Teams	Total	SNF [1]	ICF-1 [2]	ICF-2 [3]	ICF-3 [4]	ICF-4 [5]
Team A	30	3	8	6	11	2
Team B	30	5	5	10	7	3
P(X2)	.51	.49	.41	.68	.35	.66
Team A	30	7	6	4	12	1
Team C	30	6	4	5	8	7
P(X2)	.21	.77	.53	.74	.37	.03
Team B	30	10	6	4	7	3
Team C	30	8	7	2	8	5
P(X2)	.82	.64	.77	.42	.78	.49

[1]Skilled nursing facility.
[2]Intermediate care facility, level one.
[3]Intermediate care facility, level two.
[4]Intermediate care facility, level three.
[5]Intermediate care facility, level four.

Tables 10 and 11 detail the differences in distribution of care-level assessments between each of the teams, and the distribution of current State levels assigned to the same resident. The CHI-square analyses, testing the equality of these frequencies, reveal some interesting and significant differences.

The principal differences (Table 10) are that State surveyors using the STD method, assigned:

- Consistently more residents to the SNF category than did each of our teams in using the STD method (significantly more than A or C (P = .05) and marginally more than B (P = .10).
- Fewer residents to the lower three ICF categories; significantly fewer ICF-3's than team A (P = .05); fewer ICF-2's and ICF-4's than team B (P = .10); and fewer ICF-4's than team C (P = .05).

Table 9

Distribution of assessment-of-care levels made by teams for 30 residents: Quality Assurance Project (QAP) method

Teams	Total	SNF [1]	ICF-1 [2]	ICF-2 [3]	ICF-3 [4]	ICF-4 [5]
Team A	30	13	4	7	5	1
Team B	30	10	10	5	5	0
P(X2)	.37	.54	.10	.57	1.0	.77
Team A	30	9	4	10	6	1
Team C	30	11	4	9	6	0
P(X2)	.87	.66	1.0	.81	1.0	.77
Team B	30	15	8	5	1	1
Team C	30	17	6	4	3	0
P(X2)	.64	.70	.60	.74	.77	.77

[1]Skilled nursing facility.
[2]Intermediate care facility, level one.
[3]Intermediate care facility, level two.
[4]Intermediate care facility, level three.
[5]Intermediate care facility, level four.

Table 10

Aggregate of distributions of assessment-of-care levels made by each team and the State for 60 residents: Standard (STD) method

Teams	Total	SNF [1]	ICF-1 [2]	ICF-2 [3]	ICF-3 [4]	ICF-4 [5]
Team A	60	10	14	10	23	3
State	60	22	15	12	9	2
P(X2)	.03	.03	.84	.67	.01	.66
Team B	60	15	11	14	14	6
State	60	26	13	6	14	1
P(X2)	.04	.08	.68	.07	1.0	.06
Team C	60	14	12	7	16	11
State	60	28	10	12	9	1
P(X2)	.003	.03	.67	.25	.16	.004

[1]Skilled nursing facility.
[2]Intermediate care facility, level one.
[3]Intermediate care facility, level two.
[4]Intermediate care facility, level three.
[5]Intermediate care facility, level four.

Those differences were not as apparent with the QAP method. The principal differences under the QAP method (Table 11) is that State surveyors (using the QAP method) assigned differing proportions of residents to the ICF-1 and ICF-2 categories than did the teams in using the QAP method (P = .05), while the total number of residents assigned to these categories remained remarkably similar. It is interesting to note that in Wisconsin these two categories (ICF-1 and ICF-2) are reimbursed at the same rate but require different levels of professional nursing staff.

Table 11

Aggregate distribution of assessment-of-care levels made by each team and the State for 60 residents: Quality Assurance Project (QAP) method

Teams	Total	SNF [1]	ICF-1 [2]	ICF-2 [3]	ICF-3 [4]	ICF-4 [5]
Team A	60	22	8	17	11	2
State	60	29	23	1	7	0
P(X2)	.001	.33	.01	.001	.35	.15
Team B	60	25	18	10	61	1
State	60	31	23	1	5	0
P(X2)	.04	.43	.44	.01	.76	.68
Team C	60	28	10	13	9	0
State	60	30	22	2	6	0
P(X2)	.01	.78	.03	.005	.44	0

[1]Skilled nursing facility.
[2]Intermediate care facility, level one.
[3]Intermediate care facility, level two.
[4]Intermediate care facility, level three.
[5]Intermediate care facility, level four.

Impact on reimbursement in Wisconsin

As a final analysis, a comparison was made of the impact of QAP and STD assessment processes on the reimbursement for patients assessed in the study. The idea here is that the same 30 residents were reviewed by two teams using the same assessment methods; the reimbursement was computed using the assessments made by each of these teams separately and the two figures were compared. Table 12 presents the results of that comparison. The first column in Table 12 lists the teams whose level-of-care assessments are being compared. The second column lists the sample size on which these calculations were based. The third and fourth columns list the average reimbursement rate resulting from ratings of assessment of levels of care assigned by the two teams. Column 5 lists the differential in amount and percent between the two reimbursement levels. Columns 3-5 present data for the STD method. Columns 6-8 present data for the QAP method. The reimbursement differential is larger in STD than in QAP. This holds true for differences between research teams and differences between research teams and State. The reimbursement differential between State and research teams for the STD method is statistically significant at the .01 level. The research teams showed an average reimbursement 11 percent lower than the State.

Note that the average reimbursement is higher for patients assessed under QAP than for STD patients. This differential is probably because of the differences in the kinds of patients in the homes reviewed by the two methods. There is no evidence to suggest that QAP increases the average level of care of residents reviewed by that method.

Table 12

A comparison of average Medicaid per diem reimbursement [1] per resident over identical sets of residents

Teams	Teams being compared (1)	Number of residents (2)	STD [2] Average reimbursement (3)	STD [2] Average reimbursement (4)	STD [2] Differential in reimbursement (5)	QAP [3] Average reimbursement (6)	QAP [3] Average reimbursement (7)	QAP [3] Differential in reimbursement (8)
			First team	Second team	Amount and percent	First team	Second team	Amount and percent
Research teams	A-B	30	$27.16	$28.37	$1.21 (4.5)	$31.18	$30.99	$0.19 (.6)
	A-C	30	27.97	26.80	$1.17 (4.4)	30.14	30.87	$0.73 (2.4)
	B-C	30	29.30	27.90	$1.40 (4.5)	32.75	32.89	$0.14 (.4)
Research teams vs. State	A-S	60	27.56	30.96	$3.40 (12.3)	30.66	32.27	$1.61 (5.2)
	B-S	60	28.83	30.76	$1.93 (6.7)	31.87	32.76	$0.89 (2.8)
	C-S	60	27.53	31.70	$4.17 (15.1)	31.88	32.52	$0.64 (2.0)

[1]Data are based on the following average per diem rates:
Skilled nursing facility = $36.21
Intermediate care facility-1,2 = $30.62
Intermediate care facility-3 = $21.63
Intermediate care facility-4 = $19.73
[2]The Wisconsin standard method.
[3]Wisconsin's Quality Assurance Project method.

Discussion

This study has some implications for level of care as part of long-term care regulation and reimbursement, but also raised concern about expectations placed on State surveyors. First, the data indicate there is no significant difference between assessment methods as far as agreement rates between assessors are concerned. This suggests that the DMS-1 decision rule offers no particular reliability advantage over the more global methods examined. There may be other methodological approaches that can improve upon the global judgment process, such as multiattribute utility modeling (Gustafson et al., 1983).

Agreement between assessors differs dramatically when moving from two levels of care (SNF or IFC) to five levels of care (SNF or ICF-1-4). The research teams seemed to agree when using the binary SNF or ICF distinction (approximately 85 percent agreement). But, when the ICF assessment had to be disaggregated into sublevels (ICF-1-4), agreement dropped to about 50 percent or lower. DMS-I is not designed to perform that disaggregation, so comparisons were limited to QAP and STD methods. The results suggest that it would be impractical to place residents on Willemain's (1980) continuum of care for reimbursement purposes because uncertainty involved in assessment would be too high. Unless a more effective assessment process can be developed, it may be best to limit levels of care to SNF or ICF or to employ an alternative reimbursement process.

A second conclusion is that when both State and research teams used the STD method to assess five levels of care, the State tended to assign higher levels of care (54 percent of the time) than did the research teams. The tendency toward higher ratings was even more pronounced when our research teams had no knowledge of the level of care assigned to the patient in the previous year. When both State and research teams used the QAP to assess five levels of care, the State still tended to assign higher levels of care than did the research teams (42 percent of the time), but not as frequently as with the STD method. When only SNF or ICF distinctions were made, the State and research teams came into much closer agreement, although when disagreement occurred it still tended to be with the State teams assigning a higher level of care. The results are in the same direction as the Kane et al. (1980) and the Green and Monahan (1981) studies. Moreover, the results again call into question the appropriateness of disaggregation of the ICF level unless new definitions are developed (even then the appropriateness would be an empirical question). The results also question the appropriateness of considering the previous year's level of care when that knowledge may bias the assessors' performance.

Third, when reliability is examined in terms of reimbursement rates, one finds that there is greater variance between research teams when they employ the STD method. Moreover, when research team distributions are compared to level-of-care distributions actually assigned by State teams, the differences do have substantial reimbursement implications. When both the State and research teams used the STD method, the State assessments resulted in an 11-percent greater reimbursement than the research teams' assessments. The QAP method produced assessments that resulted in a 3-percent reimbursement differential between State and research teams. These results lead us to conclude that if reimbursement is to be determined by level-of-care assignment, methods examined here can employ no more than two levels of care (SNF or ICF) until new assessment methods can be developed and tested for reliability.

In a related study, Tyson and Gustafson (in press, 1984) found that State surveyors using the QAP method tended to reduce reimbursement more than surveyors using the STD metod. A reimbursement savings per day of $.096 for the STD method vs. $.264 for the QAP method was found.

So, the results of the comparison of the QAP and STD methods suggest that five levels of care cannot be reliably assessed using existing methods. They also suggest that although the probability of disagreeing on assessments showed no difference between STD and QAP, the average magnitude of disagreement was greater in the STD method.

Moreover, the State tends to assign higher levels of care than the research teams with the STD method. The differences are smaller with QAP. When these results are coupled with Tyson's and Gustafson's finding that QAP results in a greater reduction in reimbursement per resident review, they suggest that QAP may be a superior method of assessing level of care. The reason may be that the QAP method was intended to provide a more indepth review of a resident's care needs.

One possible cause for what appears to be elevated level-of-care assignments may be pressures applied to surveyors in the field to raise level of care because of its influence on reimbursement. Surveyors are expected to set unbiased levels of care and to identify (as part of the survey and certification process) deficiencies in quality of care delivered by a nursing home. Since level-of-care reductions are tied to reductions in reimbursement, one must expect a nursing home to resist those reductions. That resistance can range from subtle (but powerful) pressures exerted during the surveyor's visit to the nursing home, to appeals that can result in stressful confrontations with lawyers in administrative court. So, the surveyor (who must face these pressures almost daily) encounters powerful incentives to raise (or at least to not lower) a resident's level of care. Is it realistic to expect surveyors operating under these circumstances to consistently set unbiased levels of care? We doubt it. Will a case mix reimbursement system based on Resource Utilization Groups reduce those incentives? We believe that is an empirical question that needs to be addressed, but not in the safety of experimental settings where the Hawthorne effect can mask the effect of those incentives. Moreover, we believe there needs to be more research into the stresses surveyors face and the impact those stresses have on surveyors' ability to perform.

It is our opinion that reimbursement processes should not rely on judgments made by field staff. Rather, we support a mechanism of incentives that puts a health system (not just nursing homes) at financial risk from inefficiencies of that system and from poor care decisions made within that system. Social health maintenance organizations may offer one such mechanism.

So, although QAP may be a superior level-of-care assessment method, it is believed that another mechanism for reimbursement needs to be developed that avoids the need to link reimbursement to onsite assessment of patient needs. Then surveyors could concentrate on using inspection of care to ensure residents receive the care they need.

Acknowledgments

The State of Wisconsin's Bureau of Quality Compliance (specifically, Judy Fryback, Peg Smelser, Joy Stewart, and Dorothy Diedrich) were extremely supportive throughout this research project. Sandra Casper and Ann Macco helped coordinate data collection on the study. Without their assistance and support, this project could not have succeeded.

The nursing home industry, especially the nursing homes participating in this study, the nursing home associations in Wisconsin, and George MacKenzie, Director of Wisconsin Association of Nursing Homes; Sheldon Goldberg, Director of Wisconsin Association of Homes for the Aging, Inc.; and Terry Hottenroth of Wisconsin County Board Association have been very helpful. Repeatedly, nursing homes opened their doors to us under trying circumstances. Their help is appreciated.

Bruce Steinhardt, Project Officer, edited the article in several phases, and Sherry Terrell and James Beebe provided very helpful methodological advice.

References

Bishop, C., Plough, A., and Willemain, T.: Nursing home levels of care: Problems and alternatives. *Health Care Financing Review*. Vol. 2, Issue 2. HCFA Pub. No. 03068. Office of Research, Demonstrations, and Statistics, Health Care Financing Administration. Washington. U.S. Government Printing Office, Fall 1980.

Cavaiola, L., and Young, J.: An integrated system for patient assessment and classification and nurse staff allocation for long term care facilities. *Health Services Research*. Chicago. American Hospital Association, Fall 1980, pp. 281-306.

Foley, W., and Schneider, D.: A comparison of the level of care predictions of six long term care patient assessment systems. *Am J Public Health* 70(11), Nov. 1980.

Greene, V. L., and Monahan, D.: Inconsistency in level of care decisions in skilled nursing facilities. *Am J Public Health* 71(9), Sept. 1981.

Gustafson, D., Fryback, D., Rose, J., et al.: An evaluation of multiple trauma severity indices created by different index development strategies. *Med Care* 21(7), July 1983.

Kane, R., Rubenstein, L., Brook, R., et al.: *Utilization Review in Nursing Homes: Making Implicit LOC Judgments Explicit*. Reprint. The Ramo Corporation. Santa Monica, California, 1980.

McKnight, E.: *Nursing Home Research Study—Phase I: Quantitative Measurement of Nursing Services*. Colorado State Department of Public Health, August 1967.

Sager, A.: *Estimating the Costs of Diverting Patients from Nursing Homes to Home Care*. Leirnson Policy Institute, Brandeis University. Final Report to U.S. Administration on Aging, Grant #90-A-1026. November 1979.

Tinsley, H. and Weiss D.: Interrater reliability and agreement of subjecture judgment. *Job Counseling Psychology* 44(22):358-376, 1975.

Tyson, T. and Gustafson, D.: Nursing home reimbursement and level of care determination: An evaluation of a sampling approach. *Evaluation and the Health Professions*. In press, 1984.

Willemain, T.: Nursing home levels of care: Reimbursement of resident-specific costs. *Health Care Financing Review*. HCFA Pub. No. 03068. Office of Research, Demonstrations, and Statistics, Health Care Financing Administration. Washington. U.S. Government Printing Office, Fall 1980.