Inter-Rater Reliability Testing For Utilization Management Staff

SUE MCQUILLIAN, FSA
Consulting Actuary, Milliman USA

INTRODUCTION

Recent regulatory pressures and certification requirements have heightened the need for payer organizations to abide by specific standards regarding medical management operations. Payer organizations that are making medical-necessity determinations regarding reimbursement for health care services, as well as risk-bearing provider groups to whom some of these functions have been delegated, must insure consistency and appropriateness in their determinations. The most powerful accreditation body for health care payers, the National Committee for Quality Assurance (NCQA), requires that payer organizations carry out periodic inter-rater reliability assessments to insure consistency in medical management decision making.

An inter-rater reliability assessment or study is a performance-measurement tool involving a comparison of responses for a control group (i.e., the "raters") with a standard. Inter-rater reliability (also called inter-observer reliability) traditionally refers to how well two or more raters agree and is derived from the correlation of different raters' judgments. For the purposes of this paper, inter-rater reliability is a measurement of how well raters agree with a standard, which is more of an assessment of the validity of the responses. The purpose of the study is to determine whether the raters have been consistently trained and are applying that training in a consistent fashion. The analysis is intended to gauge the raters' observations and reactions resulting from a specific situation. The principles discussed herein would apply to any set of utilization management guidelines.

Author correspondence:
Sue McQuillian, FSA
1099 18th Street
Suite 3100
Denver, CO 80202
Phone: (303) 672-9070
E-mail: sue.mcquillian@milliman.com

This paper has undergone peer review by appropriate members of MANAGED CARE'S Editorial Advisory Board.

An inter-rater reliability assessment can be used to measure the level of consistency among a plan or provider group's utilization management staff and adherence to organizational medical management criteria or standards. Reasons for conducting an inter-rater reliability study within an organization include:

- Minimizing variation in the application of clinical guidelines;
- Evaluating staff's ability to identify potentially avoidable utilization;
- Evaluating staff's ability to identify quality-of-care issues;
- Targeting specific areas most in need of improvement;
- Targeting staff needing additional training; and
- Avoiding litigation due to inconsistently applied guidelines.

NCQA requires that health plans develop and implement an inter-rater reliability process for Health Plan Employer Data and Information Set (HEDIS) compliance. NCQA is an independent review organization dedicated to evaluating and reporting on the quality of managed care organizations. HEDIS is a set of standardized performance measures developed by NCQA with assistance from managed care organizations and employers concerned with quality health care. The performance indicators in HEDIS are continually developing, but most involve measuring access to care, health plan service, provider qualifications, activities that assist people to recover from illness, and management of chronic illness. The combination of these measures is intended to provide a tool for performance comparison of different health plans.

An inter-rater reliability study can provide measurement of many of these quality indicators. It can most readily be used to measure access, as NCQA looks for "fair and consistent health plan decisions about medical treatments and services provided to plan members." It also can be used to measure service, as NCQA looks for "actual improvements that the plan has made in care and service" (*"What NCQA Looks for in a Health Plan,"* «http://www.ncqa.org»). This latter indicator can be measured by reviewing the results of the inter-rater reliability study from year to year.

An inter-rater reliability assessment can be a useful tool for a health plan or provider organization. However, as with any benchmarking exercise, it cannot in and of itself enhance performance. To improve outcomes, the assessment must be followed up by analysis of the results and, most importantly, by action.

Medical management clinical guidelines, whether developed internally or purchased and then adjusted to meet specified objectives and local practice standards, are an extension of the organization's overall philosophies and goals. It is essential to the future viability of the plan or provider organization, the welfare of members and corporate partners, and the organization's standing in the community that they be applied appropriately and in a uniform fashion. Periodic benchmarking of clinical guideline application via an inter-rater reliability study is one way for an organization to insure its intentions for utilization management are met.

This paper will examine several elements that health plans or provider groups with utilization management responsibilities should consider when designing and implementing an inter-rater reliability study. The example presented on page 5 illustrates some aspects of the process. The example, although fairly simple, demonstrates how easily an inter-rater reliability study can be performed. However, inter-rater reliability is a complex concept, and a much more detailed analysis is possible. End users of any inter-rater reliability analysis should be advised of the method and depth of the analysis to avoid confusion or misunderstandings.

Design, implementation of an inter-rater reliability study

The primary considerations in the design and implementation of an inter-rater reliability study are whom and what to test. The assistance of an experienced clinical consultant can be helpful in designing a relevant, unbiased study and in determining the appropriate target focus group.

In most payer organizations, utilization management (UM) staff are divided by function. For example, some organizations have separate individuals performing preauthorization and concurrent review. Other organizations, particularly smaller entities, follow a UM strategy requiring that personnel perform both functions.

We recommend testing all personnel involved in making medical-necessity determinations and in identifying quality issues, including the health plan or provider group's medical directors. The medical director is often responsible for medical policy and quality management for the organization, making his or her participation in the inter-rater reliability process essential. Excluding any UM personnel from the study can bias the assessment and limit its ability to measure overall consistency and performance.

The illustrative inter-rater reliability study described below presents raters (i.e., UM personnel) with a set of case studies and questions that involve medical-necessity determinations and quality issue identification. The respondents' answers are compared with a standard, which should be an experienced person within the organization who truly understands the guidelines. That person may be a medical director, a nurse, or perhaps a head of utilization-management personnel. Relevance of a particular case study is affected by the respondent's function in the UM organization, so case studies should be designed with the job function of the UM staff in mind. The key is to choose cases that include events that trigger a decision normally made by the rater. For example, in a study of preauthorization personnel, a greater number of cases should involve requests for surgery to elicit situations requiring a decision by the preauthorization staff member.

When staff perform all UM functions, a variety of case study types should be presented with results for each function reported separately. A sample case study is presented in Appendix A on page 59.

Questions relating to the case studies should be chosen to gauge consistency of responses with the organization's UM standards. Included should be questions pertaining to:

- The medical appropriateness of the care provided in the case study;
- Identification of care delays;
- The appropriateness of the setting and provider type (for example, inpatient versus outpatient versus home);
- Which of the organization's guidelines is applicable; and
- The intensity of services provided.

Simply designed questionnaires, requiring clear-cut responses (for example, yes-or-no answers or reference to a specific guideline), yield more valid results. The survey questions should be constructed to garner responses that will enable the plan or provider group to identify and target specific problem areas for further testing or more intensive training.

Another approach to the inter-rater reliability study is to perform a retrospective review of randomly selected charts from actual cases handled by the individuals being rated. An advantage to this method is that the performance of the rater on the job is measured directly. However, a disadvantage is the lack of comparability, if some raters perceive their cases as more difficult. This method is also less streamlined and more expensive, since a standard must be created for each case for every rater.

The required frequency of inter-rater reliability testing is partly dependent on the experience level of UM staff. Employees who have a long history with the plan and are familiar with the guidelines should be tested at least twice annually. Newer employees who are not fully acquainted with the organization's clinical guidelines initially will need to be tested more frequently. Interactive commercial software is available for continuous electronic tracking of outcomes and variances from medical management guidelines, enabling performance monitoring of individual staff. This is another option for a plan or provider group to monitor consistency and accuracy of guideline application.

With respect to HEDIS compliance, NCQA requires that individual health plans be accountable for designing and implementing a suitable inter-rater reliability process: "Health plans are responsible for developing inter-rater or rater-to-standard reliability process. Certified auditors will review the plan's process for validating their medical record data during a HEDIS Compliance Audit." This open-ended response allows for some leeway in creating and implementing an inter-rater reliability process. Various designs are possible to accomplish both NCQA's and the plan or provider group's objectives. The example below is an illustration of a valid method.

Example

The following example was extracted from an inter-rater reliability assignment completed for a health plan. For this project, we prepared a survey consisting of multiple case studies followed by a series of questions pertaining to the studies. The questions were the same for every case study.

The surveys were distributed to plan preauthorization and concurrent review nurses. They were split into two sets of case studies to target each group of nurses with questions most relevant to their functions within the plan.

An experienced Milliman & Robertson clinical consultant completed the survey to determine the standard for the inter-rater comparison. The *M&R Care Guidelines*[1] *Inpatient and Surgical Care,* which contains Optimal Recovery Guidelines (ORGs) for the care of patients without clinical complications, are the guidelines currently being used by the plan for medical management. The ORGs describe efficient care processes and expected recovery (i.e., best practices) for specific conditions, based on the patient's diagnosis and, for surgery, procedure codes. They include instructions for case management and day-by-day clinical goals for the patient. Our clinical consultant used the ORGs in preparing the standard template.

Preauthorization nurses

We reviewed the preauthorization nurses' responses versus the standard for each question by case and developed percentages to indicate the number of responses consistent with the standard. We grouped our results two ways: first, by case (Table 1), and then by respondent (Table 2). We also grouped results by medical and surgical cases (Table 1), to see if we could discern a pattern of consistent or inconsistent responses.

As an illustration of how to interpret our results, Table 1 indicates that respondents agreed with the standard for the applicable ORG 64 percent of the time for all cases (Total). For medical cases, they agreed with the standard 67 percent of the time, and for surgical cases, 61 percent of the time. On a case-by-case basis, the respondents agreed with the standard between 22 percent (case 14) and 100 percent (case 11) of the time.

Table 2 again shows that respondents agreed with the standard 64 percent of the time for the applicable ORG, with the percentage varying from 25 percent for respondent P9 to 88 percent for respondent P7 for all cases combined.

In reviewing the results of our analysis for preauthorization nurses, we made some observations:

- In general, responses for medical admissions were more consistent with the standard than for surgical admissions.
- Identification of quality issues is an area that may require some attention. The average percentage of consistent responses for quality issues over all cases was only 27 percent. Only one respondent had more than 50 percent consistency in responses versus the standard for this category. This is an open-ended question, requiring more than just a yes-or-no answer or numerical response.
- Identification of potentially avoidable services is also an area that lacks consistency.
- Consistency was better, at 78 percent overall, with respect to identification of medically appropriate admissions.
- Consistency varied widely by respondent. One respondent in particular, P2, had responses much more consistent than the average. Others had far less.

There was considerable variation in responses to the questions regarding the applicable ORG. As mentioned above, the *M&R Care Guidelines™* are the guidelines presently being used by the plan for medical management. The variance is of concern because it indicates that some respondents may not understand how to properly apply the ORGs.

Although survey respondents' answers do not truly constitute a random variable, as they should congregate around the standard, we have included a measure of the deviation of responses from the mean for each question in Table 2 ("standard deviation"). The higher the standard deviation, the more variance there is in the consistency of rater responses with the standard.

For example, the standard deviation for the question relating to quality issues is only 14 percent. Responses to this question showed poor correlation with the standard for all respondents. This appears to indicate that identification of quality issues is a concern for the plan's preauthorization nurses; however, it may also

[1] *M&R Care Guidelines* is a trademark of Milliman USA, formerly Milliman & Robertson Inc.

indicate that the question was generally misunderstood and needs to be reworded. The standard deviation for the question relating to potentially avoidable services was 35 percent. The higher number is due to the broad range in consistency among respondents — from 0 percent to 80 percent agreement with the standard for this question.

Concurrent review nurses

Tables 3 and 4 contain results for concurrent review nurses by case study and respondent, respectively. All case studies used for this portion of our analysis were for medical admissions.

We developed the following observations from our analysis of the study results for concurrent review nurses:

Consistency of responses with the standard for identification of the appropriate ORG was low at 42 percent. This can be compared with a 64 percent consistency rate for plan preauthorization nurses. Only four respondents (27 percent of all respondents) averaged greater than 50 percent consistency with the standard for all cases. The variance from the standard is of concern to the plan because it indicates that respondents may not understand how to apply the plan's medical management guidelines appropriately.

Identification of medically appropriate admissions was also a problem. Consistency for this category for concurrent review nurses was only 59 percent versus 78 percent for preauthorization nurses.

Consistency in identification of an appropriate goal length of stay was

TABLE 1 Example
Inter-rater reliability study — preauthorization nurses
Percentage of responses consistent with M&R standard, by case

	Questions						
	Optimal recovery guideline (ORG)	Medically appropriate admission	Appropriate setting	Goal length of stay	Number of potentially avoidable days	List potentially avoidable services	List quality issues
Total	64%	78%	74%	61%	54%	29%	27%
Total medical only (Cases 3, 8, 11)	67%	78%	78%	70%	52%	28%	6%
Total surgical only (cases 14, 15, 16, 17)	61%	75%	67%	47%	50%	33%	41%
Case Number							
3	56%	44%	56%	56%	56%	22%	N/A
8	44%	100%	100%	67%	44%	N/A	0%
11	100%	89%	78%	89%	56%	33%	11%
14	22%	44%	44%	33%	33%	N/A	11%
15	89%	100%	89%	33%	33%	N/A	100%
16	78%	78%	67%	56%	67%	33%	11%
17	56%	78%	67%	67%	67%	33%	N/A
19	67%	89%	89%	89%	78%	22%	N/A

only 43 percent, as opposed to 61 percent for preauthorization nurses.

Identification of potentially avoidable services and quality issues was a problem in the preauthorization nurse study, and it appears to be an even greater problem with the concurrent review nurses. Consistency in identifying potentially avoidable services and quality issues is 17 percent and 12 percent, respectively, for concurrent review nurses. No respondents had more than 50 percent consistency in their answers versus the standard for either category.

Overall consistency varied widely by respondent. One respondent in particular, C13, had responses much more consistent than the average. This respondent's consistency with the standard in identifying the appropriate ORG was low (29 percent); however, this person's consistency for the other categories was relatively high. Others exhibited a much greater variation from the standard.

The standard deviation of the consistency of responses with the standard ranges from 15 percent to 19 percent. As discussed above, correlation of responses with the standard was poor for concurrent review nurses for all questions. The low standard deviation indicates that the problem is not confined to a few individuals, but is more widespread within this group.

Recommendations

Consistency in the application of organizational guidelines, as well as within the utilization management function, is essential for compliance with NCQA standards. It is also crucial to the continuation of good

TABLE 2 Example
Inter-rater reliability study — preauthorization nurses
Percentage of responses consistent with M&R standard, by respondent

	Questions						
	Optimal recovery guideline (ORG)	Medically appropriate admission	Appropriate setting	Goal length of stay	Number of potentially avoidable days	List potentially avoidable services	List quality issues
Total	64%	78%	74%	61%	54%	29%	27%
P1	75%	38%	50%	25%	25%	0%	20%
P2	63%	100%	63%	75%	63%	80%	60%
P3	63%	75%	75%	75%	63%	60%	20%
P4	75%	88%	88%	75%	75%	20%	20%
P5	50%	63%	50%	63%	38%	80%	20%
P6	75%	88%	88%	75%	75%	0%	40%
P7	88%	88%	88%	75%	63%	0%	20%
P8	63%	75%	75%	38%	38%	20%	20%
P9	25%	88%	88%	50%	50%	0%	20%
Standard deviation	18%	19%	16%	19%	18%	35%	14%

provider relations and member satisfaction with the plan. The low level of consistency in the application of plan guidelines for all UM nurses, and in particular, the concurrent review nurses, was a subject of great concern for the plan.

For each rater, we recommend setting a goal of at least 80 percent consistency with the plan standard for each measure in the inter-rater reliability assessment. We expect higher consistency for questions requiring a yes-or-no response and for questions pertaining to the appropriate company guideline than for more open-ended questions.

Our recommendation for this plan was to conduct follow-up training for both sets of nurses, with emphasis on areas exhibiting the lowest degrees of consistency — quality issues and identification of the appropriate guideline to use for both groups, and additional training for concurrent review nurses on identification of medically appropriate admissions and potentially avoidable days. Reviews of either case studies or actual inpatient admissions on a periodic basis (e.g., quarterly) or implementation of electronic tracking software would also be useful for the plan to continually monitor performance compared to standards.

Other options an organization can explore to improve consistency in application of clinical guidelines include:

Frequency of training and follow-up. UM nurses may face scheduling conflicts that can interfere with continuing education. The plan may need to schedule regular retraining at specified intervals both to refresh knowledge of items covered in previous training and to explain revisions and additions to company guidelines.

Training format. There are many different training methods available for UM guidelines. Some will be more successful than others for specific individuals. An interactive discussion with the nurses could help to reveal more effective training methods.

Guidelines content. UM guidelines should reflect an organization's medical management policy and goals. This is typically the case if they are internally developed, although the guidelines should be reviewed on a regular basis to ensure that they continue to reflect best practices. When guidelines are purchased, they must be modified to reflect company policy. A conflict may create confusion among UM staff and result in poor inter-rater reliability assessment results. The organization may want to review its guidelines, concentrating first on areas with the poorest performance.

Designated focal points for questions. It can be helpful, particularly for an organization with inexperi-

TABLE 3 Example
Inter-rater reliability study — concurrent review nurses
Percentage of responses consistent with M&R standard, by case

	Questions						
	Optimal recovery guideline (ORG)	Medically appropriate admission	Appropriate setting	Goal length of stay	Number of potentially avoidable days	List potentially avoidable services	List quality issues
Total	42%	59%	57%	43%	50%	17%	12%
Case Number							
1	27%	40%	40%	33%	60%	20%	0%
4	7%	20%	20%	7%	27%	13%	N/A
5	53%	87%	87%	60%	27%	N/A	20%
7	33%	100%	87%	47%	33%	13%	N/A
10	60%	47%	53%	60%	53%	0%	N/A
12	40%	27%	27%	13%	53%	27%	0%
13	73%	93%	87%	80%	93%	27%	27%

TABLE 4 Example
Inter-rater reliability study — concurrent review nurses
Percentage of responses consistent with M&R standard, by respondent

	Questions						
	Optimal recovery guideline (ORG)	Medically appropriate admission	Appropriate setting	Goal length of stay	Number of potentially avoidable days	List potentially avoidable services	List quality issues
Total	42%	59%	57%	43%	50%	17%	12%
Respondents							
C1	43%	43%	43%	29%	57%	17%	0%
C2	57%	57%	43%	43%	57%	17%	25%
C3	43%	43%	57%	43%	71%	50%	0%
C4	29%	29%	29%	14%	14%	0%	0%
C5	29%	57%	57%	29%	71%	17%	0%
C6	43%	57%	43%	29%	43%	0%	0%
C7	71%	71%	71%	71%	43%	17%	0%
C8	71%	71%	71%	57%	43%	0%	0%
C9	43%	86%	86%	57%	57%	17%	0%
C10	29%	71%	57%	29%	57%	0%	0%
C11	71%	57%	57%	57%	43%	0%	0%
C12	29%	43%	43%	29%	29%	0%	25%
C13	29%	71%	71%	71%	57%	50%	50%
C14	14%	57%	57%	29%	57%	33%	25%
C15	29%	71%	71%	57%	43%	33%	50%
Standard deviation	18%	15%	15%	18%	15%	18%	19%

APPENDIX A

Sample case study

Day 1

This 52-year-old male with history of hypertension presented to the emergency department after experiencing chest pain for 20 minutes accompanied by shortness of breath. The pain was relieved with one nitroglycerin.
Physical exam: WNL, BP 170/90
EKG: sinus tachycardia
CXR: negative
Cardiac enzymes: negative
WBC 13,000
Admitted to telemetry
Cardiac consult: Cardiac catheterization ordered to rule out coronary disease

Day 2

Cardiac catheterization showed a 20% lesion in the left anterior descending. No further episodes of chest pain reported. The patient was discharged with follow-up cardiology appointment.

enced personnel, to have designated experts within the UM area to answer nurses' questions regarding application of the guidelines to specific cases. If such a program already exists, the organization needs to make certain that there are no communication barriers between UM staff and the designated experts.

Medical director acting as a mentor and trainer. As the individual often having ultimate responsibility for medical policy and quality management, an effective medical director can be a valuable resource in ensuring that clinical guidelines are met. The medical director should be an active participant in UM training efforts, from planning the agenda to performing portions of the training. The medical director should be available to staff to answer clinical questions.

Other uses for inter-rater reliability testing

The inter-rater reliability concept has relevance in other areas within a health plan. An example is the underwriting function, which involves adhering to a plan's published guidelines while allowing considerable discretion on the part of the underwriter. Underwriters can be tested for consistency on many procedures, including the following:

- Exceptions to plans and rates offered to groups and individuals;
- Small-group debit point underwriting systems, through assignment of points to specific medical conditions;
- Alternate funding methods offered to groups;
- Assignment of rating relationships by rating tier;
- Compliance with state and federal laws (e.g., the Health Insurance Portability and Accountability Act of 1996);
- Handling of large claims in experience; and
- Triggers requiring the request of medical records or a physical exam for an applicant.

Another example is the claims processing function. Plans generally publish guidelines for claims handling to make certain claims are processed consistently, correctly, and in a manner that does not run afoul of state laws relating to unfair claims settlement practices. Some possible areas for testing include:

- Application of the contract's pre-existing exclusion;
- Eligibility determination;
- Procedures for claims denial;
- Procedures for high dollar claims (e.g., required signoffs);
- Knowledge of various precertification requirements;
- Appropriate application of contractual member cost-sharing provisions;
- Compliance with state and federal laws (e.g, HIPAA); and
- Procedures for identification of possible cases of upcoding, unbundling, and fraud.

An inter-rater reliability study for either the underwriting or claims disciplines can help determine whether company guidelines are being followed properly and promote appropriate risk management.

Conclusion

An inter-rater reliability study is of great value to an organization in identifying and targeting problem areas within its utilization management operations. A lack of consistent application of clinical guidelines among UM personnel in making medical-necessity determinations and identifying quality issues can be costly and create dissatisfaction among members and providers. It can also have a negative effect on the quality of patient care.

An inter-rater reliability study requires a carefully planned design that ensures the data gathered is relevant and valid. A poorly designed case study or question may be misunderstood by the raters and result in poor outcomes. It requires an experienced reviewer to interpret and apply clinical guidelines and analyze results compared with the plan's standard criteria. A survey format, as used in the example above, necessitates some downtime on the part of the raters to review the case studies and complete the surveys. Although the design and implementation of an inter-rater reliability study can be time intensive, it can be a roadmap to identify problems involving compliance with clinical guidelines, leading to substantial improvements in organizational performance and, most importantly, quality of care. MC