Comparison of Accuracy and Interreader Agreement in Side-by-Side versus Independent Evaluations of MR Imaging of the Medial Collateral Ligament of the Elbow

Kelly H. Zou, PhD, John A. Carrino, MD

Rationale and Objectives. The authors compared independent and side-by-side evaluation of magnetic resonance (MR) images of the medial collateral ligament (MCL) of the elbow, with regard to sensitivity, specificity, and interreader agreement.

Materials and Methods. Six MR imaging sequences were used to image the MCLs in 28 cadaveric specimens, eight with surgically created lesions. Two reading methods were used. For independent evaluation, the images were first evaluated independently and rated on a five-point scale by two musculoskeletal radiologists experienced in interpreting MR images and blinded to the MCL integrity. The images were then reevaluated on the same scale by both readers after at least 2 weeks, with images from all six sequences shown side by side. For each MR sequence and reading method, the sensitivity and specificity were estimated nonparametrically, and differences were tested with the McNemar test. Interreader agreement was assessed with a \( \kappa \) statistic, and differences were tested with Z and \( \chi^2 \) tests after adjustment for the dependence structure between correlated \( \kappa \) statistics.

Results. For all sequences, side-by-side evaluation generally yielded higher specificity than independent evaluation, as well as better agreement between readers.

Conclusion. Observer performance is superior when multiple MR imaging pulse sequences are reviewed simultaneously rather than independently and separately. Side-by-side review of different MR pulse sequences enabled higher accuracy and lower interreader variability for evaluation of the elbow MCL. These findings have implications for the design of studies to optimize MR imaging protocols by using multiple pulse sequences and multiple readers.

Key Words. Diagnostic radiology, observer performance; elbow, MR; ligaments, injuries; ligaments, MR; statistical analysis.

© AUR, 2002

The assessment of diagnostic accuracy and reliability is an important topic for clinical radiologic diagnostic studies. When one is comparing two readings under different conditions, such as diverse image settings (varied window and level or alternate image processing algorithms), competing pulse sequences or protocols, or different modalities, it is efficient to adopt a correlated study design in which the two approaches are administered at the same time to the same subjects (1,2). Such a design allows for better control of subject-to-subject variability. When one is comparing diagnostic accuracy between two correlated sets of diagnostic results, one may evaluate and compare the sensitivity or specificity at a natural common fixed level of specificity at the natural threshold of the tests. More sophisticated receiver operating characteristic curve (ROC)
methods at all possible threshold levels for comparing diagnostic accuracies have also been developed (1,3).

In evaluations of inter- or intraobserver variability based on counts and proportions derived from rating data, the $\kappa$ statistic is commonly used for measuring agreement (4). Arrive et al (5) developed and evaluated a scale to assess the methodologic quality in clinical investigations of radiologic studies. The reliability of this scale was evaluated according to interreader agreement as measured with $\kappa$. In this article, we focus on the $\kappa$ statistic for rating data rather than the intraclass correlation coefficient, commonly used for continuous data (6,7).

Correlated $\kappa$ statistics may be compared in evaluating the interreader reliability results for two or more pulse sequences or image-reading methods in the same set of subjects. Because the same subjects were used for all readings, a subject-specific effect is likely to influence all ratings for that subject, making them more similar. Consequently, the dependence structure between the ratings must be accounted for to enable unbiased inferences when the $\kappa$ statistics for sequences are compared.

Such correlated reliability study designs have been adopted by several investigators and are illustrated in the following two clinical examples found in the literature. In one clinical example, two radiologists and two nonradiologists (hematologists) used the radiographic vertebral index to evaluate 40 radiographs in patients with myeloma. Correlated interreader variability was compared for the two radiologists versus the two nonradiologists (8). In another clinical example, two pathologists each evaluated photomicrographs twice to determine the presence or absence of dysplasia of the urothelium in 27 patients, and correlated intrarreader variability measures were compared for the two (2,9).

In a recent study (10), we also employed a correlated study design to assess six magnetic resonance (MR) techniques used to image 28 elbow ligaments; the images were interpreted by two readers using two reading methods: independent and side-by-side evaluation. In that investigation, we focused on the accuracy and observer variability achieved with independent evaluation and used side-by-side evaluation to determine qualitative subjective reader preferences between pulse sequences. The current study was performed to test the hypothesis that side-by-side evaluation of MR images obtained with multiple pulse sequences would yield both higher diagnostic accuracy and better interreader agreement than independent evaluation of the same images.

### MATERIALS AND METHODS

**Image Acquisition**

Twenty-eight freshly frozen elbow joints from 17 male and 11 female cadavers were imaged. The age at death ranged from 54 to 86 years (mean, 71 years). The specimens were amputated at the midhumeral and midforearm levels. Two orthopedic surgeons performed extraarticular dissections of each specimen and inspected ligaments to determine intactness. Complete tears were then surgically created at the typical location for medial collateral ligament (MCL) tears (mid substance of the anterior bundle) in eight of 28 specimens.

All specimens underwent MR imaging with the following six pulse sequences: (a) T1-weighted spin echo (SE) (500/18 [repetition time msec; echo time msec]; imaging time, 3 minutes 16 seconds); (b) proton-density (PD)–weighted fast SE (2,000/48; echo train length, four; number of signals acquired [NSA], two; imaging time, 3 minutes 16 seconds); (c) fat-suppressed T2-weighted fast SE (3,100/72; echo train length, eight; NSA, two; imaging time, 2 minutes 35 seconds); (d) gradient-recalled echo (GRE) with a high matrix (500/10; flip angle, 30$^\circ$; NSA, three; imaging time, 4 minutes 52 seconds); (e) fat-suppressed T1-weighted SE with intraarticular administration of a dilute gadolinium solution (MR arthrography) (500/18; NSA, one; imaging time, 3 minutes 28 seconds); and (f) PD-weighted fast SE with a high matrix (high-resolution PD-weighted fast SE) (2,000/48; echo train length, four; NSA two; imaging time, 3 minutes 55 seconds). The matrix was $512 \times 256$ for the high-resolution PD-weighted fast SE and GRE sequences and $256 \times 256$ for all others.

For all sequences, the section thickness was 3 mm with an intersection gap of 0.3 mm, the field of view was asymmetric (15 $\times$ 11 cm), and saturation pulses were placed inferior and superior, outside the field of view. For fat-suppressed images, frequency-selective chemical presaturation was performed (Chemsat; GE Medical Systems, Milwaukee, Wis). The MR images were obtained with a GE Signa 5x platform 1.5-T magnet (GE Medical Systems) after the specimens were thawed. Each elbow was placed in a receive-only extremity coil in full extension. The first series obtained was a localizer in the standard coronal plane. The next series was oblique axial with the imaging plane adjusted to be perpendicular to the long axis of the distal humerus (true axial with respect to the humerus). From this series an oblique coronal plane was selected, parallel to a line drawn between the humeral...
epicondyles. This plane, used for all sequences, produced 12 locations covering the entire elbow joint from anterior to posterior. The sequences employed are summarized by Carrino et al (10–12).

Each elbow was studied before and after intraarticular administration of gadopentetate dimeglumine (Magnevist; Berlex, Wayne, NJ). A 25-gauge needle was inserted into the elbow joint with fluoroscopic guidance. In each elbow, 12 mL of 1 mmol/L gadopentetate dimeglumine (1 mL in 250 mL of saline) was injected, and the intraarticular location was confirmed with 1–2 mL of iodinated contrast material. All injections and imaging were performed independently by a single radiologist who was not involved in interpreting the images.

Image Analysis and Reading Methods

The images obtained with the different MR pulse sequences were evaluated independently by two musculoskeletal radiologists (readers A and B) experienced in MR imaging and blinded to the surgical results. An initial training session was provided for identification and evaluation of the MCL.

Independent evaluation.—In the first set of evaluations, the images obtained with the different sequences were interpreted independently at separate sessions at least 2 weeks apart. Each radiologist saw 28 cases in each reading session, reviewing images obtained with only one pulse sequence per session, for a total of six sessions. The cases were randomized and reordered between reading sessions. A standardized score sheet with a five-point scale was used to grade the status of the MCL (1 = definitely normal, 2 = probably normal, 3 = possibly normal, 4 = probably abnormal, 5 = definitely abnormal). No attempt was made to hide the imaging sequence used, because the sequences are readily distinguished by radiologists experienced with MR imaging.

Side-by-side evaluation.—After at least 2 weeks, the images were reinterpreted by each reader with images obtained with all six pulse sequences placed side by side on a view box. The readers were then asked to score images obtained with each of the six pulse sequences again on a standardized score sheet employing the same five-point scale used before. This method of reading has been used by us previously (10–12).

Statistical Methods

The main purpose of the analysis was to compare the effect of interpretation context on the resulting diagnostic accuracy and interreader variability. For each reader, the rating data for the six MR pulse sequences were derived from the same cases and were therefore correlated.

We used the following statistical methods to compare correlated diagnostic accuracy: (a) nonparametric estimation of sensitivity and specificity at clinically meaningful cutoff points, (b) comparison of correlated sensitivity or specificity with the McNemar test, (c) estimation of interreader agreement with the κ statistic, and (d) comparison of correlated κ statistics with the Z and the χ² tests.

Diagnostic accuracy.—First, for combinations of reader and reading method, we used sensitivity and specificity values at a clinically meaningful cutoff point to summarize the discriminatory abilities of the six MR sequences with reference to the gold standard (surgically created tears of the MCL). The natural cutoff point was defined as between ratings 3 (possibly normal) and 4 (probably abnormal).

The more popular methods of ROC curve analysis that use the software programs of Metz et al (13) were initially considered in the analysis. We attempted pairwise comparisons of the summary ROC measures between reading methods by using the CORROC2 program of Metz et al (13), as well as their LABMRMC program, which is based on the Dorfman-Berbaum-Metz multimodality approach for correlated rating data (14). On many occasions, however, our correlated data were degenerative (15), and ROC methods for dealing with correlated degenerative data were not readily available in the literature, which will be the subject of future investigation. Thus, we focused on sensitivity and specificity in our analysis of diagnostic accuracy.

The hypothesis for comparing correlated sensitivities (or specificities) was formally tested with the McNemar test (4,16). For a contingency table consisting of observations cross-classified on the two (row and column) variables with the same levels, the McNemar statistic is often used to test the null hypothesis of symmetry, namely, that the probability of an observation classified into cell (i, j) is the same as the probability of an observation classified into cell (j, i) (for i = 1,2; j = 1,2 in a two-by-two contingency table). Testing that the correlated underlying sensitivity values are equal is the same as testing the symmetry of a two-by-two table consisting of true-positive and false-negative fractions in the diseased sample. Similarly, testing that the correlated underlying sensitivity values are equal is the same as testing the symmetry of a two-by-two table consisting of true-negative and false-positive fractions in the nondiseased sample.
For each method of reading, the interreader variability (reader A vs reader B) was determined by using a \( \kappa \) statistic of the dichotomized data based on the clinically meaningful cutoff point defined earlier (4). To compute \( \kappa \), complete tears were considered present for grades of 4 or 5 and were considered absent for grades of 1, 2, or 3. The values for \( \kappa \) are between \( 1 \) and 1. The strength of agreement was interpreted on the basis of the \( \kappa \), as recommended by statistical and medical literature (17,18): 0.81–1.00 indicated very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and less than 0.20, poor. Finally, the correlated \( \kappa \) statistics for the two methods were compared by means of the Z and \( \chi^2 \) tests (2) after adjustment for the dependence due to reading method.

### RESULTS

#### Diagnostic Accuracy

For all sequences, side-by-side evaluation was generally more specific than independent evaluation (Table 1). For reader A, the specificities for independent evaluation ranged from 0.85 to 1.00, while those for side-by-side evaluation were 1.00 in all cases. For reader B, the specificities for independent evaluation ranged from 0.60 to 1.00, compared with a range of 0.85–0.95 for side-by-side evaluation; the specificity for GRE imaging improved significantly with side-by-side evaluation, from 0.60 to 0.90 (\( P = .04 \) by McNemar test).

The sensitivities were not significantly different. For reader A, the sensitivities ranged from 0.25 to 0.88 for independent evaluation and from 0.50 to 0.75 for side-by-side evaluation. For reader B, they ranged from 0.25 to 1.00 for independent evaluation and from 0.50 to 0.63 for side-by-side evaluation.

#### Interreader Agreement

Better agreement was observed for side-by-side evaluation than for independent evaluation in all sequences (T1-weighted SE, 0.51 vs \( \kappa = -0.17 \), respectively; PD-weighted fast SE, 0.87 vs 0.78; T2-weighted fast SE, 0.87 vs 0.43; GRE, 0.51 vs 0.16; MR arthrography, 0.70 vs 0.60; and high-resolution PD-weighted fast SE, 0.66 vs 0.36) (Table 2). The differences between the \( \kappa \) values were statistically significant for T1-weighted SE imaging (\( P = .03 \) and .05 by Z and \( \chi^2 \) tests, respectively) and T2-weighted fast SE imaging (\( P = .04 \) and .05 by Z and \( \chi^2 \) tests, respectively).

### Table 1

<table>
<thead>
<tr>
<th>MR Sequence*</th>
<th>Reader A</th>
<th>Reader B</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1-weighted SE</td>
<td>0.25, 0.95</td>
<td>0.50, 1.00</td>
</tr>
<tr>
<td>PD-weighted fast SE</td>
<td>0.38, 1.00</td>
<td>0.50, 1.00</td>
</tr>
<tr>
<td>T2-weighted fast SE</td>
<td>0.50, 0.95</td>
<td>0.50, 1.00</td>
</tr>
<tr>
<td>GRE</td>
<td>0.63, 0.85</td>
<td>0.75, 1.00</td>
</tr>
<tr>
<td>MR arthrography</td>
<td>0.88, 1.00</td>
<td>0.63, 1.00</td>
</tr>
<tr>
<td>High-resolution PD-weighted fast SE</td>
<td>0.50, 1.00</td>
<td>0.63, 1.00</td>
</tr>
</tbody>
</table>

Note.—Sensitivity is listed before specificity in each pair.
*See Image Acquisition for detailed information about pulse sequences.
†These specificities were significantly different (\( P = .04 \)).

### Table 2

<table>
<thead>
<tr>
<th>MR Sequence*</th>
<th>Independent Evaluation</th>
<th>Side-by-Side Evaluation</th>
<th>( \kappa )</th>
<th>( P ) Value (one sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1-weighted SE</td>
<td>( -0.17 )</td>
<td>0.51</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>PD-weighted fast SE</td>
<td>0.78</td>
<td>0.87</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>T2-weighted fast SE</td>
<td>0.43</td>
<td>0.87</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>GRE</td>
<td>0.16</td>
<td>0.51</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>MR arthrography</td>
<td>0.60</td>
<td>0.70</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>High-resolution PD-weighted fast SE</td>
<td>0.36</td>
<td>0.66</td>
<td>0.09</td>
<td>0.10</td>
</tr>
</tbody>
</table>

*See Image Acquisition for detailed information about pulse sequences.

**Interreader agreement.**—For each method of reading, the interreader variability (reader A vs reader B) was determined by using a \( \kappa \) statistic of the dichotomized data based on the clinically meaningful cutoff point defined earlier (4). To compute \( \kappa \), complete tears were considered present for grades of 4 or 5 and were considered absent for grades of 1, 2, or 3. The values for \( \kappa \) are between \( -1 \) and 1. The strength of agreement was interpreted on the basis of the \( \kappa \), as recommended by statistical and medical literature (17,18): 0.81–1.00 indicated very good agreement; 0.61–0.80, good; 0.41–0.60, moderate; 0.21–0.40, fair; and less than 0.20, poor. Finally, the correlated \( \kappa \) statistics for the two methods were compared by means of the Z and \( \chi^2 \) tests (2) after adjustment for the dependence due to reading method.
This study found that when six MR sequences are performed to evaluate for MCL tears of the elbow, assessments of multiple sequences displayed simultaneously were more specific and had less interobserver variability than assessments that rated each pulse sequence separately. The effect on sensitivity was variable; results were improved with side-by-side evaluation for some sequences but worsened for others.

In clinical practice it is typical to use multiple MR pulse sequences in one examination. Different pulse sequences in the same plane are performed to exploit different types of tissue contrast. In musculoskeletal imaging, there are often two major goals: anatomic assessment (eg, coronal, sagittal, and axial sections) and tissue characterization. By combining information from different pulse sequences and multiplanar sections, MR imaging helps to determine the pathoanatomic and pathophysiologic status of the articulation or structure of interest. Two aspects of clinical imaging addressed in this work are diagnostic accuracy and observer performance as measured by interreader variability.

A common paradigm for assessing the diagnostic efficacy of MR imaging is to compare a new pulse sequence with an established sequence (11) or to compare findings from each pulse sequence tested with those from a reference standard such as a cadaver (10,12) or a surgical procedure. It is not conventional to compare one sequence with a combination of two or more sequences. While one sequence may not be more accurate than another, if one particular sequence or combination of sequences has less interreader variability, this advantage would make MR imaging more useful in clinical settings without expert interpreters. Therefore, the use of multiple sequences may provide added value even if the overall accuracy of a combination is not statistically significantly better than that of individual sequences. In musculoskeletal imaging, these issues are becoming important, given the rapidly increasing applications of noninvasive diagnostic techniques for bone and joint disorders. The reproducibility of interpretation for musculoskeletal imaging warrants further evaluation.

A radiology report is a composite of all the information obtained from an imaging study (in this instance, different MR pulse sequences), but there may be discrepancies between pulse sequences that are not captured by a single summary diagnosis. Occasionally, radiologists report a discrepancy between sequences and recommend an additional study to arbitrate a decision. This possibility has relevance for future studies designed to identify optimal protocols as aggregates versus optimal pulse sequences as components. In ascertaining the superiority of one pulse sequence over another, independent evaluation is appropriate, as there is a tendency to develop a consensus among pulse sequences with side-by-side evaluations, a tendency avoided when pulse sequences are viewed in isolation.

Our investigation had several limitations. First, we compared separate, independent evaluations of each sequence against evaluations all six at once. It is unlikely that all six sequences would be used in a single clinical protocol. We did not compare the independent evaluation with a single summary diagnosis from the side-by-side evaluation, which is the clinical interpretation paradigm. Therefore, one extension of our study would be to derive and compare combinations of two or more pulse sequences. This method would be laborious but can be constrained by clinical requirements (eg, “anatomic” and “fluid-sensitive” information) and the finite number of MR pulse sequences available.

It has also been shown that radiologists’ diagnoses tend to be strongly influenced by the context of interpretation. Context bias was originally described in relation to the prevalence of disease in other recently observed cases (19), but it may also apply to the situation in our study. The presence of a single sequence (among numerous others) that shows an abnormality may predominate the diagnostic process. Even though our readers viewed the same set of images in the independent and side-by-side evaluations, in side-by-side evaluation it is likely that indeterminate or borderline findings are reconciled with the other sequences, making the findings more concordant (less interobserver variability). This effect underscores the difficulty of obtaining unbiased estimates of diagnostic accuracy for both new and existing technologies when they are evaluated simultaneously.

We must also emphasize that the data in this experiment do not fully support a claim of reduced interreader variability, because both readers were radiologists chosen for their expertise in musculoskeletal MR imaging. We thought it would be useful to determine first how domain experts perform. If certain sequences universally perform poorly with the most experienced observers, then there is no sense in pursuing them with novices or nonspecialists. If we want MR imaging to be more widely useful, we need to evaluate the performance of general radiologists under the same conditions as in this experiment. We also
need to include radiologists of different subspecialties and levels of expertise.

This investigation was an exploratory analysis limited by its small number of cases, lesions, and radiologists. Unfortunately, the number of cases is subject to the availability of cadavers. While determining the integrity of the MCL is a common indication for elbow MR imaging, MCL tears are relatively uncommon. We purposefully chose not to have half the cases abnormal, which would give any reader a 50% chance of being correct (equivalent to a coin flip). Such an even split of the cases could limit the number of sequences to test with other radiologists and facilitate the performance of larger-scale confirmatory studies.

There are limitations and practical difficulties in performing ROC curve studies for every possible diagnosis with every possible modality and every level of observer. Nevertheless, the ROC method for assessing human performance retains a very important role for task-specific questions such as this investigation. ROC studies are time consuming and costly and can be difficult to implement because they require a large number of human observations, especially for a large number of possible conditions. In addition, several practical and important issues for ROC analyses, such as the limited and degenerative nature of our data, are unresolved methodologically, but solutions are being pursued. Therefore, we advocate performing smaller exploratory studies to help determine whether to pursue a question further with a large-scale ROC study.

In summary, we have shown that in the evaluation of observer performance, a study design that uses side-by-side assessments of different pulse sequences will yield different results than an design in which each sequence is graded separately and independently. For studies evaluating diagnostic accuracy, some investigators may use the results of the entire examination, while others may test a specific pulse sequence, but often the context is not clearly delineated or addressed in the analyses. It is important that correlated diagnostic accuracy and reliability be analyzed and compared efficiently when such data are acquired, as illustrated in this investigation.

REFERENCES


ACKNOWLEDGMENTS

We thank William B. Morrison, MD, and William N. Snearly, MD, for reviewing the cases and Peter M. Mur-