Published in final edited form as:

Acad Radiol. 2006 May ; 13(5): 566–572.

Statistical Combination Schemes of Repeated Diagnostic Test Data¹

Kelly H. Zou, PhD, Jui G. Bhagwat, MBBS, MPH, and John A. Carrino, MD, MPH

Abstract

Rationale and Objectives—When diagnostic tests are repeated and combined, a number of schemes may be adopted. Guidelines for their interpretations are required.

Materials and Methods—Three combination schemes, “and” (A), “or” (O), and “majority” (M), are considered. To evaluate these schemes, dependency by specifying κ values quantifying repeated test agreement was structured. In a pilot study, the combined accuracies of magnetic resonance imaging using six different pulse sequences of medial collateral ligaments of the elbows of 28 cadavers, with eight having lesions artificially created surgically, were examined. Images were evaluated simultaneously by using a five-point ordinal scale. For each pulse sequence, individuals for whom the diagnosis varied from once to three repetitions were considered.

Results—Scheme M improves diagnostic accuracy when sensitivity and specificity of a single test exceed 0.5, with maximal improvement at 0.79. Under scheme A, sensitivity decreases to 0.38–0.59. Under scheme O, sensitivity increases to 0.53–0.79. Scheme M yields a small improvement, reaching 0.50–0.71. Under scheme A, specificity increases to 0.95–0.98. Under scheme O, specificity decreases to 0.91–0.98. Scheme M also yields a small improvement, reaching 0.94–0.98.

Conclusion—Scheme A is recommended for ruling in diagnoses, scheme O is recommended for ruling out diagnoses, and scheme M is neutral. Consequently, different schemes may be used to optimize the target diagnostic accuracy.

Keywords

Elbow ligament; magnetic resonance (MR) imaging; κ statistic; sensitivity; specificity; repeated diagnostic test

When comparing two readings under a number of competing pulse sequences, protocols, or different modalities, we often encounter repeated image readings by multiple tests to evaluate reproducibility, reliability, and accuracy on the same set of images (1). This repeated study design better evaluates the variability caused by these factors in different individuals. Comparing diagnostic accuracy between two correlated sets of diagnostic results, one may attempt to improve either sensitivity or specificity at a natural common fixed level of specificity by administering several diagnostic tests repeatedly.

¹From the Department of Radiology (L-2), Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115 (K.H.Z., J.G.B., J.A.C.); and Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115 (K.H.Z.).

Address correspondence to: K.H.Z. e-mail: zou@bwh.harvard.edu.

This publication was made possible by grant no. NIH R01LM007861-01A1 GM074068-01A1 and U41RR019703-01A2 from the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. J.A.C. was supported by the Radiology Research Academic Fellowship (GERRAF) Program of the Association of University Radiologists.

To evaluate such interobserver or intraobserver variability based on counts and proportions derived from rating data, the κ statistic frequently is used for measuring agreement (1–3). Correlated κ statistics may be compared when evaluating reliability results based on two or more pulse sequences or imaging reading methods on the same set of subjects. Because the same subjects were used for all readings, a subject-specific effect is likely to influence all ratings on a given subject, making them more similar. Consequently, dependency between ratings must be accounted for to make unbiased inferences when comparing κ statistics obtained for the same individuals (4).

Previously, we used a correlated study design in which six magnetic resonance (MR) imaging techniques were assessed simultaneously by two radiologists on 28 cadavers, with a subset of cadavers having medial collateral ligament (MCL) damage (1,5,6). We focus on accuracy and observer variability by using a simultaneous evaluation, called side-by-side reading, for qualitative subjective preferences between the different pulse sequences.

The purpose of the present study is to develop statistical methods to assess the impact of repeated image reading on the resulting sensitivity and specificity by using three different schemes. A decision model based on data from clinical pilot analysis is established to examine the accuracy of combining multiple repeated MR imaging for hypothetical future patients with MCL damage. We consider several schemes and determine conditions leading to significant effects on combined test sensitivity and specificity. To evaluate and compare the combination schemes with repeated tests, we specify κ values. The κ statistic quantifies repeated test agreement as a way of specifying the dependence of the repeated tests on the same set of individuals. Effects of selected parameters on the improvement of the accuracy also are studied.

MATERIALS AND METHODS

Imaging Acquisition

Details of specimen preparation and image acquisition are available elsewhere (5,6). Test pulse sequences are shown in Figure 1, and the following six MR imaging parameters were used on all specimens: (1) spin echo T1 weighted (T1SE): repetition time (TR), 500 ms; echo time (TE), 18 ms; scan time, 3 minutes 16 seconds; (2) fast spin echo proton density weighted (PDFSE): TR, 2000 ms; TE, 48 ms; echo train length, 4; number of signals averaged, 2; scan time, 3 minutes 16 seconds; (3) fat suppressed fast spin echo T2 weighted (T2FSE): TR, 3100 ms; TE, 72 ms; echo train length, 8; number of signals averaged, 2; scan time, 2 minutes 35 seconds; (4) gradient recalled echo with a high matrix (GRE): TR, 500 ms; TE, 10 ms; flip angle, 30°; number of signals averaged, 3; scan time, 4 minutes 52 seconds; (5) fat suppressed T1 weighted spin echo with intra-articular administration of a dilute gadolinium solution (MR arthrography [MRAr]): TR, 500 ms; TE, 18 ms; number of signals averaged, 1; scan time, 3 minutes 28 seconds; and (6) fast spin echo proton density with a high matrix (high resolution proton density [HRPD]): TR, 2000 ms TE, 48 ms; echo train length, 4; number of signals averaged, 2; scan time, 3 minutes 55 seconds. The matrix was 256×256 for all sequences, except that HRPD and GRE used 512×256 .

Image Analysis

Images using these different MR pulse sequences were evaluated independently by two musculoskeletal radiologists experienced in MR imaging and blinded to surgical results. An initial training session was provided for identification and evaluation of the MCL ligament for each cadaver (specimen) as the unit of analysis. After at least 2 weeks, images were reinterpreted by each radiologist with all six pulse sequences placed side by side on a view box. Radiologists then were asked to re-score each of the six pulse sequences by means a

standardized score sheet using a five-point scale (definitely normal, probably normal, indeterminate, probably abnormal, and definitely abnormal).

Gold Standard

Discriminatory abilities of the six MR sequences indicate diseased (D) when a complete tear of the MCL ligament was created artificially by the orthopedic surgeon or healthy (H) when the tear was not created on each cadaver. Thus, true disease status was known in the pilot data. It depends on the patient population in practice when designing a future repeated diagnostic test study.

Single Test Accuracy

Accuracy can be summarized by a pair of sensitivity and specificity values at the clinically meaningful cutoff point. In our example, the cutoff point is defined between ratings 3 (possibly normal) and 4 (probably abnormal). Sensitivity is the true positive rate (TPR) in the D population and specificity is the true negative rate (TNR) in the H individuals. We approximated single test sensitivity and specificity as described previously (1), using mean rating averaged between the two radiologists for each pulse sequence.

Dependence Structure Between Repeated Tests

The κ value measures reproducibility for categorical data on the same set of individuals. We use κ to measure reproducibility between the two correlated diagnostic tests (4). To compute κ , complete tears were considered present for grades of 4 or 5 and complete tears were considered absent for grades of 1, 2, or 3. κ Values are between -1 and 1 .

Guidelines for evaluation of κ suggest the following guidelines for interpreting κ values: very good, $0.81-1.00$; good, $0.61-0.80$; moderate, $0.41-0.60$; fair, $0.21-0.40$; and poor, less than 0.20 (7,8). Alternatively and coarsely, $0.75-1.00$ shows excellent reproducibility, $0.40-0.75$ shows good reproducibility, and $0-0.40$ shows marginal reproducibility (2,3).

Dependence Structure Between Repeated Readings

We modeled the decision processes to imply that we have conditional independence for two tests T_i 's, where $i = 1, \dots, I$. In our application, I was considered up to three repetitions. Given the underlying gold standard of either H or D, we modeled the output of T_i as independent diagnoses, conditioning on the gold standard. Hence, these test results on the same set of individuals were correlated.

The same underlying threshold γ was assumed to dichotomize results of the i th diagnostic test T_i . Sensitivity (TPR) and specificity (TNR) of a single test were denoted here by using the symbols $\alpha(T_i) = P(T_i^+/D) = \text{TPR}$ and $\beta(T_i) = P(T_i^-/H) = \text{TNR}$, respectively. We examined results of any two consecutive tests for the D population by assuming that repeated tests have the same sensitivity and specificity. The dependence structure was characterized by k_I and k_0 for individuals D and H, respectively.

Within D, proportions of individuals, U_I , V_I , W_I , and X_I , were derived from four possible combinations of the dichotomized test results, listed in Table 1. By definition, $k_I = (U_I \times X_I - V_I \times W_I) / \{\alpha(I - \alpha)\}$, and under the marginal constraints, we solve for these respective cell probabilities as listed in Table 1. Thus, using the Bayes rule, transition probabilities of a pair of consecutive test results for D are as follows:

$$\begin{aligned}
P(T_{i+1}^+ | T_i^+, D) &= U_1 / a = \pi_1, \\
P(T_{i+1}^- | T_i^+, D) &= 1 - \pi_1, \\
P(T_{i+1}^+ | T_i^-, D) &= a(1 - \pi_1) / (1 - a), \\
P(T_{i+1}^- | T_i^-, D) &= 1 - a(1 - \pi_1) / (1 - a)
\end{aligned}$$

Within H, similarly, proportions of the individuals U_0 , V_0 , W_0 , and X_0 also are listed in Table 1. By definition, of $k_0 = (U_0 \times X_0 - V_0 \times W_0) / \{\beta(1 - \beta)\}$ and under the marginal constraints, these cell probabilities may be computed. Again, using the Bayes rule, the transition probabilities of a pair of consecutive test results for H:

$$\begin{aligned}
P(T_{i+1}^+ | T_i^+, H) &= U_0 / (1 - \beta) = \pi_0, \\
P(T_{i+1}^- | T_i^+, H) &= 1 - \pi_0, \\
P(T_{i+1}^+ | T_i^-, H) &= (1 - \beta)(1 - \pi_0) / \beta, \\
P(T_{i+1}^- | T_i^-, H) &= 1 - (1 - \beta)(1 - \pi_0) / \beta
\end{aligned}$$

Three Combination Schemes

It is optional to perform the test based on whether the test should be ordered and whether it should be repeated. Several diagnostic decisions faced with are as follows: for example, perform the test once under the “single” (S) scheme, perform the test twice under the “and” (A) scheme or “or” (O) scheme, and perform the test three times under the “majority” (M) scheme. Table 2 lists designs of schemes S, A, O, and M. The combined test acts as if it were a new test, although test accuracy varies among different schemes. Combined test accuracy may be computed mathematically by using the described transition probabilities.

RESULTS

Theoretical Test Accuracy Under Three Combination Schemes

Conditions leading to significant effects on combined sensitivity and specificities are as follows: Under scheme A, $\alpha_A = \alpha - \alpha(I - \alpha)(I - k_1)$ and $\beta_A = \beta + \beta(I - \beta)(I - \kappa_0)$. Thus, compared with a single diagnostic test with sensitivity of α and specificity of β , scheme A always increases specificity, particularly when single test specificity is close to 0.5 and κ is less than 1, but always decreases sensitivity.

Under scheme O, $\alpha_O = \alpha + \alpha(I - \alpha)(I - k_1)$, and $\beta_O = \beta - \beta(I - \beta)(I - \kappa_0)$. Thus, compared with a single diagnostic test, scheme O always increases sensitivity, particularly when single-test sensitivity is close to 0.5 and κ is less than 1, but always decreases specificity.

Under scheme M, $\alpha_M = \alpha + \alpha(I - \alpha)(2\alpha - 1)(I - k_1)^2$ and $\beta_M = \beta + \beta(I - \beta)(2\beta - 1)(I - \kappa_0)^2$. Thus, scheme M increases both sensitivity and specificity if their values under a single test are greater than 0.5. Furthermore, for scheme M, the incremental improvement is proportional to the squared “ $I - \kappa$ ” values. Maximal improvement in accuracies with either single sensitivity or specificity is approximately 0.79.

Combined Test Accuracy Derived From the Pilot MCL Study

When applying these results to the MCL study, which motivated this research, Table 3 lists κ for consecutive reading sessions in the pilot analysis by using the mean rating of the two radiologists. Mean accuracy of the two radiologists ranged from 0.50 to 0.69 for sensitivity

and 0.93 to 0.98 for specificity. Of six pulse sequences, κ varied between 0.51 and 0.87, indicating moderate to high reproducibility between the two readers included in this pilot study. For simplicity, we assume that $\kappa = \kappa_0 = \kappa_1$ for the two gold standard categories for designing a future repeated imaging study for hypothetical patients with possible MCL damages. In Table 4, combined test accuracy also is reported under various schemes, including A, O, and M, for such hypothetical patients in a future repeated diagnostic test study.

In Figure 2, we show that sensitivity is within 0.50–0.69 in a single test. Under scheme A, it decreases to 0.38–0.59. Under scheme O, it increases to 0.53–0.79. Scheme M yielded a small improvement, reaching 0.50–0.71 among the six pulse sequences.

In Figure 3, specificity is within 0.93–0.98 in a single test. Under scheme A, it increases to 0.95–0.98. Under scheme O, it decreases to 0.91–0.98. Scheme M also yields a small improvement, reaching 0.94–0.98. In this particular MCL example, because of the small improvement based on three repeated tests and the high baseline specificity, we suggest adopting scheme O in a future validation study to improve sensitivity.

DISCUSSION

Using the methods for computing combined test accuracy, we show mathematically at the beginning of the previous section that scheme A increases specificity, whereas scheme O increases sensitivity.

We conclude that scheme A is recommended to rule in a disease, whereas scheme O is recommended to rule out a disease. Scheme M generally increases both accuracy measures, and maximal improvement occurs when the single accuracy value is 0.79, showing the value of repeating a diagnostic test.

Some clinical examples that highlight when different combination schemes may be needed depending on the clinical scenario include the following. For internal derangement of joints, surgical treatment decisions often are facilitated based on results of MR imaging. There is a risk for overtreatment because complex reconstructions and substantial rehabilitation time lose effectiveness if inappropriately applied. Thus, in this context, one wants to prioritize ruling in a specific diagnosis, and scheme A would be favored to improve specificity. In relation to the underlying test in question, high-contrast resolution imaging (MR arthrography) is advocated and used in our clinical practice in addition to high spatial resolution imaging to determine ligament status, particularly in athletes. In the setting of posttreatment imaging for malignant neoplasm, detecting recurrence and providing early treatment are the priority. There is a risk for undertreatment and missing an opportunity to provide local tumor control. Thus, in this context, one wants to prioritize ruling out recurrence, and scheme O would be favored to improve sensitivity. Current oncology practice for tumor surveillance imaging often includes morphological assessment (eg, MR imaging) and physiological assessment (eg, positron emission tomography).

Several limitations exist in our approaches. Radiologists may wish to wait until quality of the data is improved, rather than administer repeated tests on real individuals compared with cadavers, as in our pilot study. Our model may be sensitive to several assumptions, such as conditional independence and equal marginal accuracy per repetition, but some of these baseline data may be biased estimates of their true underlying parameter values. In our pilot, sensitivity and specificity of MR imaging may be assessed subjectively and may be caused by different ways the images are displayed (1,5,6). We did not consider the presence of established clinical criteria as an additional diagnostic tool. The prior probability of MCL damage was created artificially in our pilot data, which would be different in real practice (the clinical

incidence of MCL tears are much less than the approximately 30% present in the cadaveric model).

In summary, the model established in this article is an attempt to use a decision approach to evaluate repeated diagnostic tests for musculoskeletal imaging, which are frequent in practice. Our combination methods are motivated by pilot clinical data. The appropriateness of conducting repeated diagnostic tests is evaluated. Other methods available for combining classifiers or biomarkers also may be compared in the future (9–12).

Acknowledgements

The authors thank William B. Morrison, MD, and William N. Snearly, MD, for reviewing the cases and Peter M. Murray, MD, and Richard T. Steffen, MD, for performing dissections and creating the ligament lesions.

References

1. Zou KH, Carrino JA. Comparison of accuracy and interreader agreement in side-by-side versus independent evaluations of MR imaging of the medial collateral ligament of the elbow. *Acad Radiol* 2002;9:520–525. [PubMed: 12458877]
2. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley, 1981; 212–255.
3. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: Wiley, 1986;1–45.
4. Donner A, Shoukri MM, Klar N, Bartfay E. Testing the equality of two dependent kappa statistics. *Stat Med* 2000;19:373–387. [PubMed: 10649303]
5. Carrino JA, Morrison WB, Zou KH, Steffen RT, Snearly WN, Murray PM. Lateral ulnar collateral ligament of the elbow: optimization of evaluation with two-dimensional MR imaging. *Radiology* 2010;218:118–125. [PubMed: 11152789]
6. Carrino JA, Morrison WB, Zou KH, Steffen RT, Snearly WN, Murray PM. Noncontrast MR imaging and MR arthrography of the ulnar collateral ligament of the elbow: prospective evaluation of two-dimensional pulse sequences for detection of complete tears. *Skel Radiol* 2001;30:625–632.
7. Altman DG. *Practical Statistics for Medical Research*. New York: Chapman & Hall, 1991.
8. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174. [PubMed: 843571]
9. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Trans Pattern Anal Machine Intell* 1998;20:226–239.
10. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000;1:123–140. [PubMed: 12933515]
11. Thompson ML. Assessing the diagnostic accuracy of a sequence of tests. *Biostatistics* 2003;4:341–351. [PubMed: 12925502]
12. Liu AY, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Stat Med* 2005;24:37–47. [PubMed: 15515132]

Figure 1.
MR pulse sequences with arrow showing artificially created MCL tear: **(a)** T1SE, **(b)** PDFSE, **(c)** GRE, **(d)** T2FSE, **(e)** MRAr, and **(f)** HRPD.

Figure 2.

Combined test accuracy in sensitivity under different combination schemes by MR pulse sequence. The greatest improvement in sensitivity was under scheme O for GRE (0.69–0.79). Scheme M yielded slight increases for all sequences.

Figure 3.

Combined test accuracy in specificity under different combination schemes by MR pulse sequence. Specificity was improved slightly under scheme A for T1SE, GRE, and HRPD (0.95–0.97 under all three sequences) and MrAr (0.93–0.95).

Table 1
Frequency Distributions Stratified by the Gold Standard

Proportions of D Individuals				
Diagnostic Test 1 Outcomes Within D				
	Frequency Distribution	Test 1 Positive	Test 1 Negative	Marginal
Diagnostic test 2 outcomes within D	Test 2 positive	$U_I = \alpha(\alpha + k_I(1 - \alpha))$	$V_I = \alpha - A$	α
	Test 2 negative	$W_I = \alpha - A$	$X_I = 1 - A - 2B$	$1 - \alpha$
	Marginal	α	$1 - \alpha$	1
Proportions of H Individuals				
Diagnostic Test 1 Outcomes Within H				
	Frequency Distribution	Test 1 Positive	Test 1 Negative	Marginal
Diagnostic test 2 outcomes within H	Test 2 positive	$U_0 = (1 - \beta)(1 - k_0\beta)$	$V_0 = 1 - \beta - E$	$1 - \beta$
	Test 2 negative	$W_0 = 1 - \beta - E$	$X_0 = 1 - E - 2F$	β
	Marginal	$1 - \beta$	β	1

α = Sensitivity = true positive rate of each test for the D population.

β = Specificity = true negative rate of each test for the H population.

$k_I = (U_I \times X_I - V_I \times W_I) / (\alpha(1 - \alpha)) = \kappa$ statistic between the two tests for the D (gold standard = 1) population.

$k_0 = (U_0 \times X_0 - V_0 \times W_0) / (\beta(1 - \beta)) = \kappa$ statistic between the two tests for the H (gold standard = 0) population.

Table 2
Combination Scheme Based on Repeated Diagnostic Test

Combination Scheme		Diagnostic Test Results			Composite Diagnostic
		Test 1	Test 2	Test 3	
Single (S)	1 test only	+	NA	NA	+
		–	NA	NA	–
And (A)	2 of 2 positive tests	+	+	NA	+
		+	–	NA	–
		–	+	NA	–
		–	–	NA	–
Or (O)	2 of 2 negative tests	+	+	NA	+
		+	–	NA	+
		–	+	NA	+
		–	–	NA	–
Majority (M)	2 of 3 tests	+	+	+	+
		+	–	+	+
		+	–	–	–
		–	+	+	+
		–	+	–	–
		–	–	–	–

NA, not applicable.

Table 3

Intersession Reproducibility and Mean Sensitivity and Specificity Computed Using Pilot Data From 28 Cadaveric Specimens With MCL Tears Surgically Created in Eight Specimens

Pulse Sequence	κ Statistic	Mean Sensitivity (α)	Mean Specificity (β)
T1SE	0.51	0.50	0.95
PDFSE	0.87	0.50	0.98
T2FSE	0.87	0.50	0.98
GRE	0.51	0.69	0.95
MRAr	0.70	0.63	0.93
HRPD	0.66	0.57	0.95

Table 4
Sensitivity and Specificity Pairs Under Different Combination Schemes Based on Repeated Image Readings When Designing a Hypothetic Study of Repeated MR Imaging of Patients With Possible MCL Damage

Pulse Sequence	κ Statistic	Combination Scheme and (Sensitivity, Specificity) Pair		
		Single (S) (α, β)	And (A) (α_A, β_A)	Or (O) (α_O, β_O)
T1SE	0.51	(0.50, 0.95)	(0.38, 0.97)	(0.62, 0.93)
PDESE	0.87	(0.50, 0.98)	(0.47, 0.98)	(0.53, 0.98)
T2FSE	0.87	(0.50, 0.98)	(0.47, 0.98)	(0.53, 0.98)
GRE	0.51	(0.69, 0.95)	(0.59, 0.97)	(0.79, 0.93)
MRAr	0.70	(0.63, 0.93)	(0.56, 0.95)	(0.70, 0.91)
HRPD	0.66	(0.57, 0.95)	(0.49, 0.97)	(0.65, 0.93)
				(0.50, 0.96)
				(0.50, 0.98)
				(0.50, 0.98)
				(0.71, 0.96)
				(0.64, 0.94)
				(0.57, 0.95)