Andrew J. Dwyer, MD

Matchmaking and McNemar in the Comparison of Diagnostic Modalities[1]

THE Letter to the Editor in this issue of *Radiology* by Drs Mann and Hildebolt (1) on the inappropriate use of a χ^2 test in modality comparison of related samples in the article by Chandnani et al (2) and the response of the authors are notable in that the sole issue of the exchange is statistical technique.

Such dialogue in radiology is important and should be encouraged. Routine criticism and discussion of the presentation and analysis of radiologic data would serve two major functions. First, as in this case, educational: the existing elements of statistical analysis would be reiterated and distinguished and their proper application promoted. Second, investigational: it would facilitate the development and formalization of methods of study design and formats of data presentation, summarization, and analysis—the "statistics" of radiology. The "statistification" of radiology must be an iterative process of inventing descriptive, graphic, and analytic tools and tailoring them to radiologic tasks; the testing of these tools on real data and radiologic problems is essential for validation and efficient development. The Letters to the Editor section of *Radiology* should provide an excellent forum for elucidating, discussing, and remedying the problems uncovered by such exploration and evaluation.

The criticism of Drs Mann and Hildebolt and the brief review by Shaffer et al (3) of the statistical methods of articles recently published in *Radiology* suggest a need for greater familiarity by radiologic investigators with the analysis of "matched data," specifically, the McNemar test. Since most radiologic research yields matched data—the result of the direct comparison of different diagnostic techniques performed on the same patients—the McNemar test and related analysis should be in the radiologic investigator's statistical repertoire. The purpose of this editorial is to illustrate and explain the application and principles of these statistical techniques and to provide a list of references.

	CT		
MR	Positive	Negative	Total
Positive	A	B	
Negative	C	D	
Total			N

	CT		
MR	Positive	Negative	Total
Positive	20	10	30
Negative	1	1	2
Total	21	11	32

Example of data array for samples comprised of matched pairs comparing CT and MR imaging (left), and array with data from the study of Chandnani et al (right).

Matched Data

Comparative radiologic studies commonly produce matched data (matched samples) since, frequently, all the examinations under comparison are performed on each subject. For example, in the study of Chandnani et al (2), infections were experimentally created in the legs of New Zealand white rabbits and every animal was imaged with both magnetic resonance (MR) imaging and computed tomography (CT). Matched data consist of matched pairs of results—in this case, the MR imaging and CT results for each experimental animal. Since the same lesion is imaged by means of both techniques, the data are matched by the characteristics that may influence the imaging results (ie, lesion size, extent, and location).

McNemar Test

The McNemar test (4–7) pertains to matched pairs of binary (dichotomous) test results. The results of each diagnostic test fall into two categories, positive and negative. The data are succinctly presented in a two-by-two array with the rows corresponding to the results of one diagnostic test and the columns to the results of the other; each element of the array is the number of cases observed with the particular combination of test results.

For illustration, consider the data of Chandnani et al for the animals with osteomyelitis (since their report contained only the total number of animals that were CT positive and negative and MR positive and negative, the numbers may be slightly inaccurate; however, no entry in the array can be off by more than 1).

The sample (or observed) "sensitivity" or true-positive rate of each technique to the presence of the osteomyelitis is the number of cases with "positive" findings divided by the total number in the array; hence, the sample sensitivity of MR imaging = $(A + B)/(A + B + C + D) = 30/32 = .94$ and the sample sensitivity of CT = $(A + C)/(A + B + C + D) = 21/32 = .66$ (Figure). By subtraction, the sample (or observed) difference in the sensitivities is the sensitivity of MR imaging minus the sensitivity of CT ($[B - C]/N = 9/32 = .28$). An analogous array can be generated for the cases free of disease (the animals free of infection). This analysis will be concerned with only the data array for the diseased cases (the infected animals). An identical analysis can be applied to the other array; but since the data in that array represent "normals" (animals free of infection), that analysis pertains to the assessment of the false-positive rates and the "specificities" of the diagnostic tests under comparison.

The objective of the McNemar test is to assess the statistical significance of the observed differences in sensitivities between MR imaging and CT. *B* and *C* in the data array are "discordant" data (ie, the MR imaging results differ from the CT results). These discordant data are the basis of the McNemar analysis. The logic behind the test is intuitive; it is based on the fact that if MR imaging and CT have equal sensitivities in the population, the odds should be 50-50 that a discordant case selected at random from the population is either MR positive and CT negative or MR negative and CT positive. Hence, if a number of cases are sampled and the results placed in an array as shown earlier, it is

Index terms: Editorials • Statistical analysis

Radiology 1991; 178:328–330

[1] From the Department of Radiology, Room 1C660 Clinical Center, National Institutes of Health, Bethesda, MD 20814. Received and accepted October 3, 1990. **Address reprint requests to** the author.

See also the Letter to the Editor by Mann and Hildebolt (pp 582–583) in this issue.

to be expected that the discordant cases will be evenly divided between the discordant cells MR positive and CT negative and CT negative and MR positive. The more uneven the distribution of the discordant cases between the discordant cells—the more B differs from C—the greater the evidence that the sensitivity of CT in the population differs from that of MR imaging. The McNemar test assesses the likelihood that the observed distribution of discordant cases between the two discordant cells could occur by chance if the diagnostic techniques being compared had equal sensitivities in the population. The significance level of the result of the McNemar test is the probability of the observed distribution or a more unequal distribution of the discordant data occurring if the diagnostic tests being compared had equal sensitivities.

Binomial Distribution

The statistics governing the distribution of discordant cases with equally sensitive techniques are identical to those of the flip of a coin; the chance of a randomly chosen discordant case being MR positive and CT negative is .5, the same as the chance of "heads" occurring on the flip of a coin. The chance of 10 of 11 discordant cases being MR positive and CT negative and one of 11 being MR negative and CT positive is the same as the chance of one heads occurring in 11 flips of the coin (ie, $11[.5]^{11} = .0005$). Similarly, the chance of two of 11 discordant cases being MR negative and CT positive is the same as the chance of two of 11 coin tosses being heads. These probabilities are given by means of a binomial distribution. The binomial distribution pertains to a series of independent binary events (eg, 11 flips of a coin) in which each event has the same probability of occurrence, p (eg, probability of heads on each flip = .5). The binomial probability distribution gives the probability that a specific number of these events, x, will be positive as a function of n and p (eg, the probability of two of 11 flips of the coin being heads is given by the binomial distribution with $p = .5$, $n = 11$, and $x = 2$). Therefore, if the diagnostic tests compared are equally sensitive, the probability that B cases will be MR positive and CT negative and C cases will be MR negative and CT positive is determined by means of binomial distribution with $p = .5$, $n = B + C$, and $x = B$.

Exact Form of the McNemar Test

The binomial distribution is the basis of the exact form of the McNemar test. For illustration of the exact form of the McNemar test, consider its application to the MR imaging and CT data presented earlier, in which one of 11 discordant cases were MR negative and CT positive and 10 were MR positive and CT negative. The objective of the test is to calculate the probability that the observed discrepancy between the numbers of discordant cases (one and 10) or an equally or more discrepant combination (ie, 10 and one, 11 and 0, 0 and 11) could occur by chance if MR imaging and CT had equal sensitivities in the population. This probability is calculated by means of the binomial distribution and is the significance level, or P value, of the test. The significance level forms the basis for rejection of the null hypothesis that the diagnostic tests being compared are equally sensitive; the smaller the P value, the stronger the evidence for rejection. The P value indicates the likelihood that the null hypothesis is being falsely rejected.

With use of the binomial distribution (assuming equal sensitivities), the probability of one of 11 discordant cases being MR negative and CT positive is .0054, the same as that of the equally disparate combination 10 of 11 being MR negative and CT positive. The probabilities of the more disparate combinations of 0 of 11 and 11 of 11 being MR negative and CT positive are .0005; hence, by the exact form of the McNemar test, the hypothesis that MR imaging and CT have equal sensitivities can be rejected at the $P = (.0054 + .0054 + .0005 + .0005) = .012$ level of significance.

The McNemar Statistic

Approximation rather than an exact method is the basis of an alternative form of the McNemar test. This form is rooted in the fact that if the diagnostic tests being compared are equally sensitive, then $[(B - C) - 1]^2/[B + C]$ approximates a χ^2 distribution with 1 df. The expression $[(B - C) - 1]^2/[B + C]$ is termed the McNemar statistic with continuity correction (drop the -1 within the brackets and it is the McNemar statistic). The greater the difference between B and C, the greater the value of the statistic and the stronger the evidence for rejecting the hypothesis that MR imaging and CT have equal sensitivities. Since this form of the McNemar test is based on a "limit theorem" (ie, the approximation becomes exact only in the "limit" as the size of the sample becomes infinite), it should be applied only to large samples. There are guidelines for determining what constitutes a sufficient sample size such as $B + C > 20$.

To illustrate the use of the McNemar statistic, suppose there were 40 discordant cases, 10 MR negative and CT positive and 30 MR positive and CT negative. The value of the McNemar statistic with correction for continuity would be $[(30 - 10) - 1]^2/[30 + 10] = 9.025$; reference to a table of χ^2 distribution with 1 df reveals 9.025 to exceed the 99.5th percentile of the χ^2 distribution with 1 df (ie, if MR imaging and CT have equal sensitivities in the population, the chances of the McNemar statistic having a value greater than the one observed is <0.5%); hence, the hypothesis that MR imaging and CT have equal sensitivities in the population can be rejected at the $P < .005$ level of significance.

Some methods extend the notion of the McNemar test to matched data tables with more than two rows and columns suitable to the analysis of diagnostic tests with more than two outcomes (4).

Statistical versus Diagnostic Significance

The diagnostic advantage of one modality over another depends on the magnitude of the differences in their sensitivities. One problem with the McNemar test—a trait shared with tests of significance in general—is that trivial differences in sensitivity, no matter how small, become statistically significant as the sample size increases. Hence, it is important to distinguish between statistical significance and diagnostic significance; assessment of the magnitude of the differences in sensitivities of the diagnostic tests is a necessary part of the comparative analysis.

The McNemar analysis provides for estimation of the magnitude of the difference in the sensitivities. Specifically, the estimate of the sensitivities is $(B - C)/N = (10 - 1)/32 = 0.28$ for the matched MR imaging–CT data array.

Confidence Intervals

$(B - C)/N$ provides what is known as a "point" estimate; the estimate of the difference in sensitivities is given as a single value. The "confidence interval" is another type of estimate; it is given as a range of values rather than just a single value. With the confidence interval comes a "confidence level," which indicates the likelihood that the confidence interval derived from the data will include the true value of the parameter that is being estimated for the population from which the data are a sample. The greater the confidence level, the wider the interval (ie, the more certain the conclusion, the less precise the statement).

The confidence interval of the difference in the sensitivities of two diagnostic tests may be approximated from the matched data array as follows (5,8):

$$\left\{\frac{B - C}{N} - [K\ (\text{SE } \Delta\text{sensitivities})] - \frac{1}{N},\ \frac{B - C}{N} + [K\ (\text{SE } \Delta\text{sensitivities})] + \frac{1}{N}\right\},$$

where K is a coefficient dependent on

the desired confidence level of the confidence interval and where the standard error (SE) of the difference in the sensitivities (SE [Δ sensitivities]) is given by

$$\text{SE } \Delta\text{sensitivities} = \frac{\sqrt{[N(B+C)]-[(B-C)^2]}}{N\sqrt{N}}.$$

For an 80% confidence interval, K = 1.28; for a 90% confidence interval, K = 1.645; for a 95% confidence interval, K = 1.96; and for a 99% confidence interval, K = 2.58. (For a confidence level of $[100-\alpha]$%, K equals the $[100-\alpha/2]$ percentile of the standard gaussian distribution with a mean of 0 and a standard deviation of 1; for a $[100-5]$% = 95% confidence interval, K = 1.96, since the area of the standard gaussian distribution with a mean of 0 and a standard deviation of 1 for values ≤1.96 incorporates $[100-\frac{5}{2}]$% = 97.5% of the total area under the gaussian distribution curve.)

From the matched data array described earlier, the SE of the difference between the sensitivities of MR imaging and CT is

$$\text{SE } \Delta\text{sensitivities} = \frac{\sqrt{[32(10+1)]-[(10-1)^2]}}{32\sqrt{32}} = 0.09.$$

The 95% confidence interval of the differences in the sensitivities (ie, the sensitivity of MR imaging minus the sensitivity of CT) is approximately

$$\left\{0.28-[1.96(0.09)]-\frac{1}{32},\ 0.28+[1.96(0.09)]+\frac{1}{32}\right\} = (0.07, 0.49).$$

With a confidence of 95%, it can be inferred that the sensitivity of MR imaging in the population is between 7% and 49% greater than the sensitivity of CT in the population; in other words, MR imaging would detect between 7% and 49% more of the infections than CT would.

The meaning of confidence level may be better appreciated by considering the following: Suppose that instead of using just one sample of 32 infected rabbits, Chandnani et al performed the same experiment numerous times, each time with a new sample of 32 rabbits. Suppose that they calculated a 95% confidence interval for the difference in sensitivities between MR imaging and CT for each sample. Since the results would differ from sample to sample, the calculated confidence intervals would also differ from sample to sample; the confidence level of these confidence intervals corresponds to the fraction that contained the true value of what is to be estimated—the difference between the sensitivities of MR imaging and CT in the population. Since 95% confidence intervals were calculated, 95% of the intervals calculated from the samples should contain the true difference between the sensitivities of MR imaging and CT in the population; hence, the chances are 95% that a randomly selected confidence interval (eg, the confidence interval derived from the random sample of 32 cases in the data array above) will contain the true difference in sensitivities.

Given the assumptions that commonly underlie their calculations, confidence intervals should be used as rough approximations. However, the use of confidence intervals to summarize results should be encouraged for several reasons. They serve to diminish confusion between statistical significance and diagnostic significance by making explicit the magnitude of the differences in sensitivities. They reinforce the reality that the estimates are projections subject to uncertainty. The confidence interval provides a range of plausible values of what is to be estimated; the width of this range indicates the degree of this uncertainty. The extremities of the confidence interval give some indication of the limits of what is inferable from the data at hand.

Complementarity of Results

The comparison of two diagnostic modalities should include consideration of the complementarity of their results. How often does one modality depict an abnormality that was missed with the other? How many additional infections would be detected if CT were performed in addition to MR imaging? The matched data array provides answers to such questions. $B/(B+D)$ gives the fraction of cases negative at CT that were positive at MR imaging (10/11 in the Figure). Similarly, $C/(C+D)$ gives the fraction of cases negative at MR imaging that were positive at CT (1/2 in the Figure).

These fractions may be used to estimate the complementarity of MR imaging and CT in the population. $B/(B+D)$ is an estimate of the probability that a randomly selected infected case negative at CT would be positive at MR imaging; $C/(C+D)$ is an estimate that a randomly selected infected case negative at MR imaging would be positive at CT. Confidence intervals for these probabilities can be estimated exactly from published tables and graphs of the confidence intervals for proportions (derived from the binomial distribution) (2,4,6); they can also be approximated by the following (adapted from reference 6):

$$\left[\frac{2fn+K^2-K\sqrt{K^2+4f(1-f)n}}{2n+2K^2},\ \frac{2fn+K^2+K\sqrt{K^2+4f(1-f)n}}{2n+2K^2}\right]$$

where f is the sample fraction that corresponds to the probability to be estimated (ie, $B/[B+C]$ or $C/[C+D]$), K is a coefficient dependent on the desired confidence level of the confidence interval (identical to the K discussed earlier with the confidence interval for the difference in sensitivities), and n is the total number of cases involved in the estimate (ie, $n=[B+D]$ to estimate the probability of a positive MR image given a negative CT; $n=[C+D]$ to estimate the probability of a positive CT given a negative MR image).

Suppose, for example, that there were 40 CT-negative cases in a matched sample of diseased cases; 25 were MR positive and 15 were MR negative (ie, B = 25 and D = 15). Hence, $f = 25/(15+25) = 0.62$ and $n = (15+25) = 40$. For a 95% level of confidence, K = 1.96. On the basis of the above formulas, the 95% confidence interval of the probability that a CT-negative case randomly selected from the population will be MR positive (0.47, 0.75).

Summary

Comparative studies of radiologic techniques commonly yield matched data due to the ease and desirability of performing all of the techniques on each of the patients. The two-by-two matched data array and the McNemar analysis provide a succinct format for the presentation and proper analysis of matched comparisons of binary (positive and negative) test results. When comparing tests, it is essential not to rely on just the statistical significance of the differences in sensitivities (or specificities); the magnitude of the differences must also be assessed. Confidence intervals provide a useful form of estimation by providing a range of plausible values and an indication of the precision of the estimate. The matched data array also indicates the complementarity of the diagnostic tests being compared. ■

References

1. Mann FA, Hildebolt CF. Inappropriate use of χ^2 test. Radiology 1991; 178:582.
2. Chandnani VP, Beltran J, Morris CS, et al. Acute experimental osteomyelitis and abscesses: detection with MR imaging versus CT. Radiology 1990; 174:233–236.
3. Shaffer PB, Chandnani VP, Beltran J. Reply. Radiology 1991; 178:582–583.
4. Conover WJ. Practical nonparametric statistics. 2nd ed. New York: Wiley, 1980; 100, 130–133.
5. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley, 1981; 112–125.
6. Rosner B. Fundamentals of biostatistics. 2nd ed. Boston: Duxberry, 1982; 168–169, 334–338.
7. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill, 1988; 75–80.
8. Berry CC. A tutorial on confidence intervals for proportions in diagnostic radiology. AJR 1990; 154:477–480.