Radiology

Kimberly E. Applegate, MD, MS
Richard Tello, MD, MSME, MPH
Jun Ying, PhD

Index term:
Statistical analysis

Published online before print
10.1148/radiol.2283021330
Radiology 2003; 228:603–608

[1] From the Department of Radiology, Riley Hospital for Children, Indianapolis, Ind (K.E.A.); Department of Radiology, Boston University, 88 E Newton St, Atrium 2, Boston, MA 02118 (R.T.); and Departments of Radiology and Biostatistics, Indiana University School of Medicine, Indianapolis (J.Y.). Received October 16, 2002; revision requested February 18, 2003; final revision received March 13; accepted April 1. **Address correspondence to** R.T. (e-mail: *tello@alum.mit.edu*).

Hypothesis Testing III: Counts and Medians[1]

Radiology research involves comparisons that deal with the presence or absence of various imaging signs and the accuracy of a diagnosis. In this article, the authors describe the statistical tests that should be used when the data are not distributed normally or when they are categorical variables. These nonparametric tests are used to analyze a 2 × 2 contingency table of categorical data. The tests include the χ^2 test, Fisher exact test, and McNemar test. When the data are continuous, different nonparametric tests are used to compare paired samples, such as the Mann-Whitney *U* test (equivalent to the Wilcoxon rank sum test), the Wilcoxon signed rank test, and the sign test. These nonparametric tests are considered alternatives to the parametric *t* tests, especially in circumstances in which the assumptions of *t* tests are not valid. For radiologists to properly weigh the evidence in the literature, they must have a basic understanding of the purpose, assumptions, and limitations of each of these statistical tests.

The purpose of hypothesis testing is to allow conclusions to be reached about groups of people by examining samples from the groups. The data collected are analyzed by using statistical tests, which may be parametric or nonparametric, depending on the nature of the data to be analyzed. Statistical methods that require specific distributional assumptions are called parametric methods, whereas those that require no assumptions about how the data are distributed are nonparametric methods. Nonparametric tests are often more conservative tests compared with parametric ones. This means that the test has less power to reject the null hypothesis (1). Nonparametric tests can be used with discrete variables or data based on weak measurement scales, consisting of rankings (ordinal scale) or classifications (nominal scale).

The purpose of this article is to discuss different nonparametric or distribution-free tests and their applications with continuous and categorical data. For the analysis of continuous data, many radiologists are familiar with the *t* test, a parametric test that is used to compare two means. However, misuse of the *t* test is common in the medical literature (2). To perform *t* tests properly, we need to make sure the data meet the following two critical conditions: *(a)* The data are continuous, and *(b)* the populations are distributed normally. In this article, we introduce the application of nonparametric statistical methods when these two assumptions are not met. These methods require less stringent assumptions of the population distributions than those for the *t* tests. When two populations are independent, the Mann-Whitney *U* test can be used to compare the two population distributions (3). An additional advantage of the Mann-Whitney *U* test is that it can be used to compare ordinal data, as well as continuous data. When the observations are in pairs from the same subject, we can use either the Wilcoxon signed rank test or the sign test to replace the paired *t* test.

For categorical data, the χ^2 test is often used. The χ^2 test for goodness of fit is used to study whether two or more mutually independent populations are similar (or homogeneous) with respect to some characteristic (4–12). Another application of the χ^2 test is a test of independence. Such a test is used to determine whether two or more characteristics are associated (or independent). In our discussion, we will also introduce some extensions of the χ^2 test, such as the Fisher exact test (13,14) for small samples and the McNemar test for paired data (15).

CATEGORICAL DATA

In many cases, investigators in radiology are interested in comparing two groups of count data in a 2 × 2 contingency table. One of the most commonly used statistical tests to analyze categorical data is the χ^2 test (16). If two groups of subjects are sampled from two independent populations and a binary outcome is used for classification (eg, positive or negative imaging result), then we use the χ^2 test of homogeneity. Sometimes radiologists are interested in analyzing the association between two criteria of classification. This results in the test of independence by using a similar 2 × 2 contingency table and χ^2 statistic. When sample sizes are small, we prefer to use the Fisher exact test. If we have paired measurements from the same subject, we use the McNemar test to compare the proportions of the same outcome between these two measurements in the 2 × 2 contingency table.

χ^2 Test

The χ^2 test allows comparison of the observed frequency with its corresponding expected frequency, which is calculated according to the null hypothesis in each cell of the 2 × 2 contingency table (Eq [A1], Appendix A). If the expected frequencies are close to the observed frequencies, the model according to the null hypothesis fits the data well; thus, the null hypothesis should not be rejected. We start with the analysis of a 2 × 2 contingency table by considering the following two examples. The same χ^2 formula is used in both examples, but they are different in the sense that the data are sampled in different ways.

TABLE 1
χ^2 Test of Homogeneity

Participants	Positive MR Imaging Result	Negative MR Imaging Result	Total
Patients	50	50	100
Control subjects	28	72	100
Total	78	122	200

Note.—Data are the number of participants. χ^2 statistic, 10.17; $P = .001$. The null hypothesis is that the two populations are homogeneous. We reject the null hypothesis and conclude that the two populations are different.

TABLE 2
χ^2 Test of Independence

Presence of Contrast Enhancement	MR Imaging Finding: Malignant Mass	MR Imaging Finding: Benign Mass	Total
Enhancement	14	3	17
No enhancement	3	45	48
Total	17	48	65

Note.—Data are the number of masses. χ^2 statistic, 34.65; $P < .001$. The null hypothesis is that contrast enhancement of a renal mass at MR imaging is not associated with the presence of a malignant tumor, and the alternative hypothesis is that enhancement and malignancy are associated. We reject the null hypothesis and conclude that the presence of contrast enhancement is associated with renal malignancy.

Example 1: test of homogeneity between two groups.—One hundred patients and 100 healthy control subjects are enrolled in a magnetic resonance (MR) imaging study. The MR imaging result can be classified as either "positive" or "negative" (Table 1). The radiologist is interested in finding out if the proportion of positive findings in the patient group is the same as that in the control group. In other words, the null hypothesis is that the proportion of positive findings is the same in the two groups. The alternative hypothesis is that they are different. We call this a test of homogeneity. In this first example, the two groups (patients and subjects) are in the rows, and the two outcomes of positive and negative test results are in the columns. In the statistical analysis, only one variable, the imaging result (classified as positive or negative), is considered.

The results in Table 1 show that 50 patients and 28 control subjects are categorized as having positive findings. The χ^2 statistic is calculated and yields a P value of .001 (17). Typically, we reject the null hypothesis if the P value is less than .05 (the significance level). In this example, we conclude that there is no homogeneity between the two groups, since the proportions of positive imaging results are different.

Example 2: test of independence between two variables in one group.—A radiologist studies gadolinium-based contrast material enhancement of renal masses at MR imaging in 65 patients (18). Table 2 shows that there are 17 patients with enhancing renal masses, with 14 malignant masses and three benign masses at pathologic examination. Among the 48 patients with nonenhancing renal masses, three masses are malignant and 45 are benign at pathologic examination. In this example, the presence or absence of contrast enhancement is indicated in the rows, and the malignant and benign pathologic findings are in the columns. In this second example, only the total number of 65 patients is fixed; the presence or absence of contrast enhancement is compared with the pathologic result (malignant or benign). The question of interest is whether these two variables are associated. In other words, the null hypothesis is that contrast enhancement of a renal mass is not associated with the presence of a malignant tumor, and the alternative hypothesis is that enhancement and malignancy are associated. In this example, the χ^2 statistic yields a P value less than .001. We reject the null hypothesis and conclude that the presence of contrast enhancement at MR imaging is associated with renal malignancy.

One potential issue with the χ^2 test is that the χ^2 statistic is discrete, since the observed frequencies in the 2 × 2 contingency table are counts. However, the χ^2 distribution itself is continuous. In 1934, Yates (12) proposed a procedure to correct for this possible bias. Although there is controversy about whether to apply this correction, it is sometimes used when the sample size is small. In the first example discussed earlier, the χ^2 statistic was 10.17, and the P value was .001. The Yates corrected χ^2 statistic is 9.27 with a P value of .002. This corrected χ^2 statistic yields a smaller χ^2 statistic, and the P value is larger after Yates correction. This indicates that the Yates corrected χ^2 test is less powerful in rejecting the null hypothesis. Some applications of Yates correction in medicine are discussed in the statistical textbook by Altman (19).

Radiology

TABLE 3
Fisher Exact Test for Small Samples

Participants	Positive CT Result	Negative CT Result	Total
Patients	10	10	20
Control subjects	4	16	20
Total	14	26	40

Note.—Data are the number of participants. (Two-sided) Fisher exact test result, $P = .10$. For small samples, the Fisher exact test is used. The null hypothesis is that the two populations of patients and control subjects are homogeneous at CT—that is, they have the same number of positive results. We retain the null hypothesis because the P value does not indicate a significant difference, and we conclude that these two groups are homogeneous. If we incorrectly used the χ^2 test for this comparison, the conclusion would have been the opposite: $\chi^2 = 3.96$, $P = .04$.

TABLE 4
McNemar Test for Paired Comparisons: Angiography versus CT Results in the Diagnosis of Coronary Bypass Graft Thrombosis

CT Result	Positive Angiography Result	Negative Angiography Result	Total
Positive	71	30	101
Negative	13	86	99
Total	84	116	200

Note.—Data are the number of CT results. McNemar χ^2 result, 5.95; $P = .02$. The P value indicates a significant difference, and therefore, we reject the null hypothesis and conclude that there is a difference between these two modalities.

TABLE 5
Incorrect Use of the χ^2 Test for Paired Data for the Evaluation of Angiography versus CT (when paired data are incorrectly treated as independent)

Modality	Positive Result	Negative Result	Total
CT	101	99	200
Angiography	84	116	200

Note.—Data are the number of results. For the χ^2 test with the assumption of two independent samples, $P = .09$. We would incorrectly conclude that there is no significant difference between these two modalities.

Fisher Exact Test

When sample sizes are small, the χ^2 test yields poor results, and the Fisher exact test is preferred. A general rule of thumb for its use is when either the sample size is less than 30 or the expected number of observations in any one cell of a 2 × 2 contingency table is fewer than five (20). The test is called an "exact" test because it allows calculation of the exact probability (rather than an approximation) of obtaining the observed results or results that are more extreme. Although radiologists may be more familiar with the traditional χ^2 test, there is no reason not to use the Fisher exact test in its place, given the ease of use and availability of computer software today.

In example 1, the P value resulting from use of the χ^2 test was .001, whereas the P value for the same data tested by using the Fisher exact test was .002. Both tests lead to the same conclusion of lack of homogeneity between the patient and control groups. Intuitively, the P value derived by using the Fisher exact test is the probability of positive results becoming more and more discrepant between the two groups. Most statistical software packages provide computation of the Fisher exact test (Appendix B).

Example 3: Fisher exact test.—A radiologist enrolls 20 patients and 20 healthy subjects in a computed tomographic (CT) study. The CT result is classified as either "positive" or "negative." Table 3 shows that 10 patients and four healthy subjects have positive findings at CT. The null hypothesis is that the two populations are homogeneous in the number of positive findings seen at CT.

In this example, the sample sizes in both the patient and control groups are small. The Fisher exact test yields a P value of .10. We retain the null hypothesis because the P value does not indicate a significant difference, and we conclude that these two groups are homogeneous. If we use the χ^2 test incorrectly, the P value is .05, which suggests the opposite conclusion—that the proportions of positive CT results are different in these two groups.

McNemar Test for Paired Data

A test for assessment of paired count data is the McNemar test (15). This test is used to compare two paired measurements from the same subject. When the sample size is large, the McNemar test follows the same χ^2 distribution but uses a slightly different formula. Radiology research often involves the comparison of two paired imaging results from the same subject. In a 2 × 2 table, the results of one imaging test are labeled "positive" and "negative" in rows, and the results of another imaging test are labeled similarly in columns. An interesting property of this table is that there are two concordant cells in which the paired results are the same (both positive or both negative) and two discordant cells in which the paired results are different for the same subject (positive-negative or negative-positive).

We are interested in analyzing whether these two imaging tests show equivalent results. The McNemar test uses only the information in the discordant cells and ignores the concordant cell data. In particular, the null hypothesis is that the proportions of positive results are the same for these two imaging tests, versus the alternative hypothesis that they are not the same. Intuitively, the null hypothesis is retained if the discordant pairs are distributed evenly in the two discordant cells. The following example illustrates the problem in more detail.

Example 4: McNemar test for paired data.—There are 200 patients enrolled in a study to compare CT and conventional angiography of coronary bypass grafts for the diagnosis of graft patency (Table 4). Seventy-one patients have positive results with both conventional angiography and CT angiography, 86 have negative results with both, 30 have positive CT results but negative conventional angiographic results, and 13 have negative CT results but positive conventional angiographic results. The McNemar test compares the proportions of the discordant pairs (13 of 200 vs 30 of 200). The P value of the McNemar statistic is .02, which suggests that the proportion of positive results is significantly different for the two modalities. Therefore, we conclude that the ability of these two modalities to demonstrate graft patency is different.

Some radiologists may incorrectly summarize the data in a way shown in Table 5 and perform a χ^2 test, as discussed in example 1 (21). This is a common mistake in the medical literature. In example 1, the proportions compared are 101 of 200 versus 84 of 200. The problem is the assumption that CT angiography and conventional angiography results are independent, and thus, the paired relationship between these two imaging tests is ignored (2,21). The χ^2 test has less power to reject the null hypothesis than does the McNemar test in this situation and results in a P value of .09. We would incorrectly conclude that there is no significant difference in the ability of these two modalities to demonstrate graft patency.

HYPOTHESIS TESTING BY USING MEDIANS

The unpaired and paired *t* tests require that the population distributions be normal or approximately so. In medicine, however, we often do not know whether a distribution is normal, or we know that the distribution departs substantially from normality.

Nonparametric tests were developed to deal with situations where the population distributions are either not normal or unknown, especially when the sample size is small (<30 samples). These tests are relatively easy to understand and simple to apply and require minimal assumptions about the population distributions. However, this does not mean that they are always preferred to parametric tests. When the assumptions are met, parametric tests have higher testing power than their nonparametric counterparts; that is, it is more likely that a false null hypothesis will be rejected.

Three commonly encountered nonparametric tests include the Mann-Whitney *U* test (equivalent to the Wilcoxon rank sum test), the Wilcoxon signed rank test, and the sign test.

Comparison of Two Independent Samples: Mann-Whitney *U* Test

The Mann-Whitney *U* test is used to compare the difference between two population distributions and assumes the two samples are independent (22). It does not require normal population distributions, and the measurement scale can be ordinal.

The Mann-Whitney *U* test is used to test the null hypothesis that there is no location difference between two population distributions versus the alternative hypothesis that the location of one population distribution differs from the other. With the null hypothesis, the same location implies the same median for the two populations. For simplicity, we can restate the null hypothesis: The medians of the two populations are the same. Three alternative hypotheses are available: *(a)* The population medians are not equal, *(b)* the population median of the first group is larger than that of the second, or *(c)* the population median of the second group is larger than that of the first. If we put the two random samples together and rank them, then, according to the null hypothesis, which holds that there is no difference between the two populations medians, the total rank of one sample would be close to the total rank of the other. On the other hand, if all the ranks of one sample are smaller than the ranks of the other, then we know almost surely that the location of one population is shifted relative to that of the other.

TABLE 6
Mann-Whitney *U* Test for Ordinal Data

Fat Saturation	Score <25	Score 25–75	Score >75	Total
No	8	14	2	24
Yes	3	12	7	22
Total	11	26	9	46

Note.—Data are scores from T2-weighted fast spin-echo MR images obtained with or without fat saturation. Wilcoxon rank sum test, $P = .03$. The null hypothesis is that the median scores for the two types of MR images are the same. We reject the null hypothesis and conclude that they are not the same. If we incorrectly used the χ^2 test, we would conclude the opposite: $\chi^2 = 5.13$, $P = .08$.

We give two examples of the application of the Mann-Whitney *U* test, one involving continuous data and the other involving ordinal data.

Example 5: Mann-Whitney U test for continuous data.—The uptake of fluorine 18 choline (hereafter, "fluorocholine") by the kidney can be considered approximately distributed normally (23). Let us say that some results of hypothetical research suggest that fluorocholine uptake above 5.5 (percentage dose per organ) is more common in men than in women. If we are only interested in the patients whose uptake is over 5.5, the distribution is no longer normal but becomes skewed. The Figure shows the uptake over 5.5 in 10 men and seven women sampled from populations imaged with fluorocholine for tumor surveillance. We are interested in finding out if there are any differences in these populations on the basis of patient sex.

We can quickly exclude use of the *t* test in this example, since the fluorocholine uptakes we have selected are no longer distributed normally. The null hypothesis is that the medians for the men and women are the same. By using the Mann-Whitney *U* test, the *P* value is .06, so we retain the null hypothesis. We conclude that the medians of these two populations are the same at the .05 significance level, and therefore, men and women have similar renal uptake of fluorocholine. If we had incorrectly used the *t* test, the *P* value would be .02, and we would conclude the opposite.

Male Patients Uptake	Male Patients Rank	Female Patients Uptake	Female Patients Rank
5.50	1	5.63	2
5.65	3	6.39	7
5.71	4	7.67	12
5.74	5	9.21	14
5.75	6	9.86	15
6.58	8	10.18	16
6.59	9	13.89	17
7.33	10		
7.40	11		
8.41	13		

Mann-Whitney *U* test result for continuous data (fluorocholine uptake over 5.5 [percentage dose per organ] in human kidneys), $P = .06$. We retain the null hypothesis that there is no difference in the medians, and we conclude that the fluorocholine renal uptake in men and women is similar at the .05 significance level (the marginal *P* value suggests a trend toward a significant difference). If we had incorrectly used the *t* test, we would have concluded the opposite: $P = .02$.

Example 6: Mann-Whitney U test for ordinal data.—A radiologist wishes to know which of two different MR imaging sequences provides better image quality. Twenty-four patients undergo MR imaging with a T2-weighted fast spin-echo sequence, and 22 other patients are imaged with the same sequence but with the addition of fat saturation (Table 6). The image quality is measured by using a standardized scoring system, ranging from 1 to 100, where 100 is the best image quality. The null hypothesis is that the median scores are the same for the two populations. In the group imaged with the first MR sequence, the images of eight subjects are scored under 25, those of 14 subjects are scored between 25 and 75, and those of two subjects are scored above 75. In the group imaged with the fat-saturated MR sequence, there are three, 12, and seven subjects in these three score categories, respectively.

In this example, each patient's image score is classified into one of three ordinal categories. Since the observations are discrete rather than continuous, the *t* test cannot be used. Some researchers might consider the data in Table 6 to be a 2×3 contingency table and use a χ^2 statistic to compare the two groups. The *P* value corresponding to the χ^2 statistic is .08, and we would conclude that the two groups have similar image quality. The problem

Radiology

TABLE 7
Wilcoxon Signed Rank Test for Paired Comparisons of Ring Enhancement between Two Spin-Echo MR Sequences

Case No.	MR Sequence*: T1-weighted	MR Sequence*: Fat-saturated T1-weighted	Difference†	Absolute Value of Difference‡	Rank†
1	45	43	−2	2	5
2	45	42	−3	3	8
3	49	47	−2	2	5
4	50	47	−3	3	8
5	49	48	−1	1	3
6	44	50	6	6	13.5
7	42	98	56	56	20
8	49	47	−2	2	5
9	39	44	5	5	11
10	42	42	0	0	1.5
11	44	54	10	10	18
12	47	53	6	6	13.5
13	42	53	11	11	19
14	45	54	9	9	16.5
15	44	48	4	4	10
16	41	47	6	6	13.5
17	45	54	9	9	16.5
18	50	47	−3	3	8
19	51	51	0	0	1.5
20	42	48	6	6	13.5

Note.—Wilcoxon signed rank test, $P = .02$. The null hypothesis is that the median of the paired population differences is zero. The Wilcoxon signed rank test result indicates a significant difference, and therefore, we conclude that the enhanced fat-saturated T1-weighted MR sequence showed ring enhancement better than did the conventional enhanced T1-weighted MR sequence. If we had incorrectly used the paired *t* test, the *P* value would be .07, and we would have had the opposite conclusion. Like the *t* test, the sign test produced a *P* value of .50, and the conclusion would be that the two sequences are the same.

* Data are image quality scores on MR images after contrast material administration.

† Difference in enhancement values between MR sequences.

‡ Rank of the absolute values of the differences; when there is a tie in the ranking, an average ranking is assigned—for example, rank 16.5 rather than ranks 16 and 17 for the tied case numbers 14 and 17; and rank 13.5 for the four cases (6, 12, 16, and 20) that compose ranks 12, 13, 14, and 15.

is that the three image quality score categories are only treated as nominal variables, and their ordinal relationship is not accounted for in the χ^2 test. An alternative test that allows us to use this information is the Mann-Whitney *U* test. The Mann-Whitney *U* test yields a *P* value of .03. We reject the null hypothesis and conclude that the median image scores are different.

Comparison of Paired Samples: Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is an alternative to the paired *t* test. Each paired sample is dependent, and the data are continuous. The assumption needed to use the Wilcoxon signed rank test is less stringent than the assumptions needed for the paired *t* test. It requires only that the paired population be distributed symmetrically about its median (24).

The Wilcoxon signed rank test is used to test the null hypothesis that the median of the paired population differences is zero versus the alternative hypothesis that the median is not zero. Since the distribution of the differences is symmetric about the mean, it is equivalent to using the mean for the purpose of hypothesis testing, as long as the sample size is large enough (at least 10 rankings).

We rank the absolute values of the paired differences from the sample. With the null hypothesis, we would expect the total rank of the pairs whose differences are negative to be comparable to the total rank of the pairs whose differences are positive. The following example shows the application of the Wilcoxon signed rank test.

Example 7: paired data.—A sample of 20 patients is used to compare ring enhancement between T1-weighted spin-echo MR images and fat-saturated T1-weighted spin-echo MR images obtained after contrast material administration (Table 7). We notice that the image quality scores on fat-saturated T1-weighted spin-echo MR images in case 7 is 98, which is much higher than the others. As a result, the difference in values between the two sequences is also much higher than that for the other paired differences. It would be unwise to use a paired *t* test in this case, since the *t* test is sensitive to extreme values in a sample and tends to incorrectly retain a false null hypothesis as a consequence. The nonparametric tests are more robust to data extremes, and thus, the Wilcoxon signed rank test is preferred in this case. The null hypothesis states that the median of the paired MR sequence differences is zero. The Wilcoxon signed rank test provides a *P* value of .02, so we reject the null hypothesis. We conclude that the fat-saturated MR sequence showed ring enhancement better than did the MR sequence without fat saturation. If we had incorrectly used the paired *t* test, the *P* value would be .07, and we would have arrived at the opposite conclusion.

Comparing Paired Samples: Sign Test

The sign test is another nonparametric test that can be used to analyze paired data. Unlike the Wilcoxon signed rank test or the paired *t* test, this test requires neither a symmetric distribution nor a normal distribution of the variable of interest. The only assumption underlying this test is that the data are continuous. Since the distribution is arbitrary with the sign test, the hypothesis of interest focuses on the median rather than the mean as a measure of central tendency. In particular, the null hypothesis for comparing paired data is that the median difference is zero. The alternative hypothesis is that the median difference is not, is greater than, or is less than zero. This simplistic test considers only the signs of the differences between two measurements and ignores the magnitudes of the differences. As a result, it is less powerful than the Wilcoxon signed rank test, and a false null hypothesis is often not rejected (25).

SUMMARY

Hypothesis testing is a method for developing conclusions about data. Radiology research often produces data that require nonparametric statistical analyses. Nonparametric tests are used for hypothesis testing when the assumptions about the data distributions are not valid or when the data are categorical. We have discussed the most common of these statistical tests and provided examples to demonstrate how to perform them. For radiologists to properly weigh the evidence in our literature, we need a basic understanding of the purpose, assumptions, and limitations of each of these statistical tests. Understanding how and when these methods are used will strengthen our ability to evaluate the medical literature.

APPENDIX A

The χ^2 formula is based on the following equation:

$$\chi^2 = \Sigma\left[\frac{(\mathrm{Fo} - \mathrm{Fe})^2}{\mathrm{Fe}}\right], \qquad (A1)$$

where Fo is the frequency observed in each cell, and Fe is the frequency expected in each cell, which is calculated by multiplying the row frequency by the quotient of the column frequency divided by total sample size.

APPENDIX B

Examples of statistical software that is easily capable of calculating χ^2 and McNemar statistics include SPSS, SAS, StatXact 5, and EpiInfo (EpiInfo allows calculation of the Fisher exact test and may be downloaded at no cost from the Centers for Disease Control and Prevention Web site at *www.cdc.gov*). Other statistical Web sites include *fonsg3.let.uva.nl/Service/Statistics.html, department.obg.cuhk.edu.hk/ResearchSupport/WhatsNew.asp,* and *www.graphpad.com/quickcalcs/Contingency1.cfm* (all Web sites accessed January 30, 2003).

Acknowledgments: The authors thank Miriam Bowdre for secretarial support in preparation of the manuscript and George Parker, MD, Phil Crewson, PhD, and an anonymous statistical reviewer for comments on earlier drafts of the manuscript.

References

1. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 526.
2. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. BMJ 1977; 1:85–87.
3. Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945; 1:80–83.
4. Fisher LD, Belle GV. Biostatistics: a methodology for the health sciences. New York, NY: Wiley, 1993.
5. Ott RL. An introduction to statistical methods and data analysis. 4th ed. Belmont, Calif: Wadsworth, 1993.
6. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995.
7. Altman DG. Practical statistics for medical research. London, England: Chapman & Hall, 1991.
8. Chase W, Bown F. General statistics. 4th ed. New York, NY: Wiley, 2000.
9. Conover WJ. Practical nonparametric statistics. 3rd ed. New York, NY: Wiley, 1999.
10. Rosner B. Fundamentals of biostatistics. 4th ed. Boston, Mass: Duxbury, 1995; 370–565.
11. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 112–125.
12. Yates F. Contingency tables involving small numbers and the chi-square test. J Royal Stat Soc Ser B 1934; (suppl 1):2179–2235.
13. Fisher RA. Statistical methods for research workers. 5th ed. Edinburgh, Scotland: Oliver & Boyd, 1934.
14. Fisher RA. The logic of inductive inference. J Royal Stat Soc Ser A 1935; 98:39–54.
15. Cochran WG. Some methods for strengthening the common chi-square tests. Biometrics 1954; 10:417–451.
16. Agresti A. Categorical data analysis. New York, NY: Wiley, 1990.
17. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 691.
18. Tello R, Davidson BD, O'Malley M, et al. MR imaging of renal masses interpreted on CT to be suspicious. AJR Am J Roentgenol 2000; 174:1017–1022.
19. Altman D. Practical statistics for medical research. London, England: Chapman & Hall, 1991; 252.
20. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 537.
21. Dwyer AJ. Matchmaking and McNemar in the comparison of diagnostic modalities. Radiology 1991; 178:328–330.
22. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947; 18:50–60.
23. DeGrado TR, Reiman R, Price DT, Wang S, Coleman RE. Pharmacokinetics and radiation dosimetry of F-fluorocholine. J Nucl Med 2002; 43:92–96.
24. Altman D. Practical statistics for medical research. London, England: Chapman & Hall, 1991; 194–197.
25. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 6th ed. New York, NY: Wiley, 1995; 569–578.