Sample Size Tables For Receiver Operating Characteristic Studies

Nancy A. Obuchowski[1]

OBJECTIVE. I provide researchers with tables of sample size for multiobserver receiver operating characteristic (ROC) studies that compare the diagnostic accuracies of two imaging techniques.

MATERIALS AND METHODS. I computed the number of patients and observers needed as a function of five parameters: the measure of diagnostic accuracy (area under the ROC curve, sensitivity at a false-positive rate ≤ 0.10, or specificity at a false-negative rate ≤ 0.10), conjectured level of accuracy, suspected difference in accuracy between the two imaging techniques, observer variability, and ratio of patients without to patients with the condition.

RESULTS. The numbers of patients and observers required vary dramatically with these five parameters, increasing with more refined measures of accuracy, with lower accuracy levels, with smaller suspected differences, with greater observer variability, and with less balanced designs. The number of patients required for a study can be reduced by increasing the number of observers, and vice versa. When the intra- and interobserver variability is large, a study design with just four observers is usually inadequate.

CONCLUSION. Many factors must be considered when determining the appropriate sample sizes for multiobserver ROC studies. My tables serve only as initial ballpark estimates. Investigators should compute sample size using parameters that reflect their clinical application.

Many of us in radiology have been involved in designing and performing diagnostic accuracy studies, and all of us are involved in interpreting diagnostic accuracy studies. An important issue in designing these studies is determining the number of patients needed for the study. For diagnostic tests that require the interpretation of images by trained observers, a common approach is to have multiple observers interpret the images of the patients in the sample. Therefore, an additional issue is the number of observers needed for the study. I provide researchers with practical tables of sample size for such studies. My tables reveal the magnitude of patient and observer sample sizes needed and illustrate how one can trade between fewer patients and more observers or more patients and fewer observers. These tables can be used in planning your own study and can also aid in examining the adequacy of study designs of other articles.

Materials and Methods

We consider the following type of study. Suppose we want to compare the diagnostic accuracies of two imaging techniques (e.g., MR imaging and CT) for detecting a certain condition or disease (e.g., cerebral aneurysms). A sample of c patients undergo both imaging techniques. We plan for r observers to interpret all the images for the sample of c patients. That is, there are $r \times c \times 2$ interpretations. We assume that the true diagnosis is known for all patients and that the interpretations are scored in such a way that receiver operating characteristic (ROC) curves can be constructed for each observer for each imaging technique (r x 2 ROC curves). For sample size determination, we assume that the primary focus of the research is the comparison of the average accuracy of the observers interpreting one imaging technique (e.g., MR imaging) with the average accuracy of the observers interpreting the other imaging technique (e.g., CT) to determine which imaging technique is superior. We want to determine adequate values for r and c.

A mathematic model of the relationship between the number of observers and number of patients for ROC studies has been previously proposed [1]. I used this model to prepare tables of sample size. The sample sizes were determined for a study with 5% type I error rate and 80% power. The type I error rate is the probability of wrongfully rejecting the null hypothesis (i.e., concluding that the two imaging techniques have different accuracies when, in truth, the accuracies are equal). Conversely, the power of a study is the probability of correctly rejecting the null hypothesis (i.e., concluding that the two imaging

Received January 24, 2000; accepted after revision February 25, 2000.

[1] Department of Biostatistics and Epidemiology and Department of Diagnostic Radiology, The Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH 44195. Address correspondence to N. A. Obuchowski.

AJR 2000;175:603–608

0361–803X/00/1753–603

techniques have different accuracies when, in truth, the accuracies are different). Studies with a type I error rate of less than 5% or power of greater than 80% require larger sample sizes (larger than reported in our tables). The Appendix provides further details about the construction of the tables.

To use my tables, you must answer five questions about the study. First, what measure of diagnostic accuracy will be used to characterize the observers' performance interpreting the images? For my tables, I consider three measures of diagnostic accuracy: the area under the ROC curve, the sensitivity at a false-positive rate less than or equal to 0.10, and the specificity at a false-negative rate less than or equal to 0.10. The area under the ROC curve is the most commonly used measure of accuracy. It is a good global measure of accuracy [2]; however, for particular clinical applications, we are usually interested in only a certain portion of the ROC curve [3]. For example, in breast cancer screening, in which false-negative test results have serious consequences, we demand high sensitivity; therefore, we provide sample size estimates for the specificity at a false-negative rate of less than or equal to 0.10 (sensitivity fixed at ≥ 0.90). However, for detecting asymptomatic cerebral aneurysms, when the treatment that follows positive test results (i.e., true-positives and false-positives) is very risky, we demand high specificity. For these studies, we provide sample size estimates for the sensitivity at a false-positive rate less than or equal to 0.10 (specificity ≥ 0.90).

Second, what is the conjectured accuracy of the observers with these imaging techniques? A synthesis of the relevant literature can often provide a reasonable estimate of the level of accuracy to expect. For my tables, I considered two levels of accuracy: high and moderate. For the area under the ROC curve, I set high accuracy at an area of 0.90 and moderate accuracy at 0.75. For the sensitivity at a fixed false-positive rate of less than or equal to 0.10, I define high accuracy as a sensitivity of 0.80 at a false-positive rate of 0.10 and moderate accuracy as a sensitivity of 0.60 at a false-positive rate of 0.10. Similarly, for specificity at a fixed false-negative rate of less than or equal to 0.10, I define high accuracy as a specificity of 0.80 at a false-negative rate of 0.10 and moderate accuracy as a specificity of 0.60 at a false-negative rate of 0.10.

Third, what is the suspected difference in accuracy between the two imaging techniques? Again, a synthesis of the relevant literature or a pilot study can help determine the difference to expect. For my tables, I considered three levels of suspected difference: small, moderate, and large. A small difference is 0.05 (an absolute difference in the areas under the ROC curve of 0.05 or an absolute difference in sensitivities [specificities] of 0.05 at a false-positive rate [false-negative rate] of 0.10), a moderate difference is 0.10, and a large difference is 0.15.

Fourth, what is the relative frequency of patients with and without the condition in the study sample? The relative frequency in the sample may not reflect the prevalence of the condition in the population. Especially for rare conditions, investigators often use sampling approaches that provide a more balanced number of patients with and without the condition. For my tables, I define R as the ratio of the number of patients without the condition to the number of patients with the condition in the study sample. I consider values for R of 1/1 (balanced design), 2/1, and 4/1.

Finally, what is the variability within (intra-) and between (inter-) the observers? Observer variability is probably dependent on many factors, including the condition of interest, the setting of image interpretations, and the heterogeneity in the observers' training and experience. The easiest way to quantify observer variability is in terms of the range of differences in accuracy. For example, if we expect that for a common set of images, observers' accuracies will differ by about 0.10 (using the area under the ROC curve, the lowest area of any observer in our sample

TABLE 1 Levels of Observer Variability

Level	Expected Range of Accuracy	
	Intraobserver[a]	Interobserver[b]
Small	0.005	0.01
Moderate	0.025	0.05
Large	0.05	0.10

[a] The anticipated difference between the accuracies of an observer who interprets the same images using the same imaging technique at two different times.

[b] The anticipated difference between the highest accuracy of any observer in the study and the lowest accuracy of any observer in the study.

TABLE 2 Number of Patients Required for Detecting a Suspected Difference in Accuracy[a]: Measure of Accuracy—Area Under Receiver Operating Characteristic (ROC) Curve

Observer Variability	Four Observers			Six Observers			Ten Observers		
	S	M	L	S	M	L	S	M	L
Moderate accuracy									
Small difference									
Ratio = 1/1	571	—	—	246	3769	—	116	201	—
Ratio = 2/1	679	—	—	293	4479	—	138	239	—
Ratio = 4/1	983	—	—	424	6488	—	200	345	—
Moderate difference									
Ratio = 1/1	133	291	—	60	78	943	29	32	51
Ratio = 2/1	159	345	—	71	92	1120	35	38	60
Ratio = 4/1	229	500	—	103	133	1622	50	55	87
Large difference									
Ratio = 1/1	59	77	2975	27	30	46	20	20	20
Ratio = 2/1	70	92	3536	32	35	54	30	30	30
Ratio = 4/1	101	132	5122	50	51	78	50	50	50
High accuracy									
Small difference									
Ratio = 1/1	287	—	—	124	1896	—	59	101	—
Ratio = 2/1	363	—	—	157	2395	—	74	128	—
Ratio = 4/1	548	—	—	236	3618	—	112	193	—
Moderate difference									
Ratio = 1/1	67	146	—	31	39	474	20	20	26
Ratio = 2/1	85	185	—	38	50	599	30	30	32
Ratio = 4/1	128	279	—	58	75	905	50	50	50
Large difference									
Ratio = 1/1	30	39	1497	20	20	23	20	20	20
Ratio = 2/1	38	49	1890	30	30	30	30	30	30
Ratio = 4/1	57	74	2856	50	50	50	50	50	50

Note.—Dash (—) indicates inadequate observer sample size; therefore, no patient sample size is provided. Table indicates the total number of patients (i.e., with and without the condition) required as a function of the number of observers (four, six, or 10), the variability among observers (S = small, M = moderate, or L = large), the level of accuracy of the tests (moderate accuracy = ROC area of 0.75, high accuracy = ROC area of 0.90), the suspected difference in accuracy between imaging techniques (small = 0.05, moderate = 0.10, or large = 0.15), and the ratio of patients without to patients with the condition (R = 1/1, 2/1, or 4/1). For detecting a large difference with six or 10 observers, the estimated number of patients is small. I believe that a reasonable minimum is 10 each of patients with and without the condition; therefore, I have set the data accordingly.

[a] 80% power, 5% type I error, two-tailed test.

might be 0.85 and the highest area of any observer in our sample might be 0.95), then we can derive an estimate of variance (Appendix). In Table 1, I set up three levels of observer variability for my sample size tables: small, moderate, and large. The interobserver ranges are based on a review article by Rockette et al. [4] in which the observer variability was estimated and reported from 49 studies. Our small range corresponds with the smallest variability reported by Rockette et al.; our large range corresponds with the largest variability reported by Rockette et al.; and our moderate range is a convenient midpoint. Rockette et al. do not report values of the intraobserver variability because of the paucity of available data. On the basis of my experience, I know that the intraobserver variability can be quite large. I arbitrarily set the intraobserver variability at half the interobserver variability.

Results

Tables 2–4 summarize the total number of patients required for studies that use the three measures of accuracy: the area under the ROC curve (Table 2), the sensitivity at a false-positive rate of less than or equal to 0.10 (Table 3), and the specificity at a false-negative rate of less than or equal to 0.10 (Table 4). Tables 2–4 reveal how the sample size requirements are larger when researchers are interested in a focused portion of the ROC curve (sensitivity or specificity for a fixed false-positive range or false-negative range) rather than the entire ROC curve. Between six and nine times more patients are required for these studies. However, we should not allow these increased sample size requirements to detract from the more refined measures of diagnostic accuracy. Also note that the required patient sample size decreases as the number of observers increases, as the observer variability decreases, as the accuracy increases, as the suspected difference increases, or as the ratio of patients without to patients with the condition approaches 1:1 (balanced design).

In the following text, I illustrate the use of my tables with two examples. First, suppose we are designing a study to compare the accuracies of MR imaging and CT for the detection of cerebral aneurysms, and we choose as our measure of accuracy the sensitivity at a false-positive rate of less than or equal to 0.10. We anticipate the two imaging techniques to have high accuracy (sensitivities near 0.80 at a false-positive rate of 0.10). We would like to be able to detect a difference in sensitivities of 0.10 or more at a false-positive rate of 0.10. We anticipate that there will be four times more patients without an aneurysm than with an aneurysm in our sample, and we expect large observer variability (sensitivities of all the observers at a false-positive rate of 0.10 may differ by as much as 0.10; and the sensitivities at a false-positive rate of 0.10 by a single observer on two different interpreting occasions may differ by as much as 0.05). Referring to Table 3, a study design with 353 patients (71 with an aneurysm and 282 without) and six observers or 123 patients (25 with an aneurysm and 98 without) and 10 observers would be adequate. Note that a study design with only four observers would be inadequate for this example because of the expected large observer variability.

For my second example, suppose we use the tables as an aid to evaluate the adequacy of the sample sizes of an article. Consider the study by Powell et al. [5] that compared the diagnostic accuracies of film-screen radiography and digitized mammograms. (Note that Powell et al. were testing if the two display modes were equivalent, not if one was superior; the statistics are different, but we can still use this study as an example.) Seven observers interpreted 60 cases using both display formats. The study included 30 patients with at least one malignant lesion and 30 patients without a malignant lesion. Each breast was divided into five regions, and each region of each breast was interpreted by the observers. One of the measurements of accuracy was the area under the ROC curve. The average area under the ROC curve of the seven observers was 0.86. The range of areas under the ROC curve for the seven observers was 0.107 (0.827–0.934) for film and 0.098 (0.792–0.890) for digitized images. Five observers interpreted the images in two separate

TABLE 3 **Number of Patients Required for Detecting a Suspected Difference in Accuracy[a]: Sensitivity at a False-Positive Rate of Less Than or Equal To 0.10**

Observer Variability	Four Observers			Six Observers			Ten Observers		
	S	M	L	S	M	L	S	M	L
Moderate accuracy									
Small difference									
Ratio = 1/1	3889	—	—	1668	—	—	785	1395	—
Ratio = 2/1	4659	—	—	1998	—	—	941	1671	—
Ratio = 4/1	6787	—	—	2910	—	—	1370	2434	—
Moderate difference									
Ratio = 1/1	901	2047	—	404	528	6364	194	217	339
Ratio = 2/1	1079	2453	—	484	633	7624	232	260	406
Ratio = 4/1	1572	3572	—	705	922	—	338	379	591
Large difference									
Ratio = 1/1	401	534	—	181	203	312	87	92	108
Ratio = 2/1	480	639	—	217	243	374	105	110	130
Ratio = 4/1	699	931	—	316	354	545	152	160	188
High accuracy									
Small difference									
Ratio = 1/1	2488	—	—	1054	—	—	494	850	—
Ratio = 2/1	3082	—	—	1306	—	—	612	1053	—
Ratio = 4/1	4596	—	—	1947	—	—	913	1570	—
Moderate difference									
Ratio = 1/1	568	1230	—	254	328	3995	121	136	213
Ratio = 2/1	704	1524	—	315	406	4950	151	168	264
Ratio = 4/1	1049	2273	—	469	606	7381	225	251	393
Large difference									
Ratio = 1/1	249	325	—	113	125	192	54	57	67
Ratio = 2/1	308	403	—	139	155	237	67	70	83
Ratio = 4/1	459	600	—	208	230	353	100	105	123

Note.—Dash (—) indicates inadequate observer sample size; therefore, no patient sample size is provided. Table indicates the total number of patients (i.e., with and without the condition) required as a function of the number of observers (four, six, or10), the variability among observers (S = small, M = moderate, or L = large), the level of accuracy of the tests (moderate accuracy = sensitivity of 0.60 at a false-positive rate of 0.10, high accuracy = sensitivity of 0.80 at a false-positive rate of 0.10), the suspected difference in sensitivities between the imaging techniques at a false-positive rate of 0.10 (small = 0.05, moderate = 0.10, or large = 0.15), and the ratio (1/1, 2/1, or 4/1) of patients without to patients with the condition.

[a] 80% power, 5% type I error, two-tailed test.

TABLE 4 **Number of Patients Required for Detecting a Suspected Difference in Accuracy[a]: Specificity at a False-Negative Rate of Less Than or Equal To 0.10**

Observer Variability	Four Observers			Six Observers			Ten Observers		
	S	M	L	S	M	L	S	M	L
Moderate accuracy									
Small difference									
Ratio = 1/1	3889	—	—	1668	—	—	785	1395	—
Ratio = 2/1	4091	—	—	1754	—	—	826	1468	—
Ratio = 4/1	5367	—	—	2301	—	—	1083	1925	—
Moderate difference									
Ratio = 1/1	901	2047	—	404	528	6364	194	217	339
Ratio = 2/1	948	2154	—	425	556	6694	204	228	357
Ratio = 4/1	1243	2825	—	558	729	8781	267	300	467
Large difference									
Ratio = 1/1	401	534	—	181	203	312	87	92	108
Ratio = 2/1	422	562	—	191	213	329	92	97	114
Ratio = 4/1	553	736	—	250	280	431	120	127	149
High accuracy									
Small difference									
Ratio = 1/1	2488	—	—	1054	—	—	494	850	—
Ratio = 2/1	2515	—	—	1065	—	—	500	859	—
Ratio = 4/1	3178	—	—	1346	—	—	631	1086	—
Moderate difference									
Ratio = 1/1	568	1230	—	254	328	3995	121	136	213
Ratio = 2/1	574	1244	—	257	332	4039	123	137	215
Ratio = 4/1	725	1571	—	325	419	5103	156	173	272
Large difference									
Ratio = 1/1	249	325	—	113	125	192	54	57	67
Ratio = 2/1	251	329	—	114	126	194	55	57	68
Ratio = 4/1	318	415	—	144	159	245	69	73	85

Note.—Dash (—) indicates inadequate observer sample size; therefore, no patient sample size is provided. Table indicates the total number of patients (i.e., with and without the condition) required as a function of the number of observers (four, six, or 10), the variability among observers (S = small, M = moderate, or L = large), the level of accuracy of the tests (moderate accuracy = specificity of 0.60 at a false-negative rate of 0.10, high accuracy = specificity of 0.80 at a false-negative rate of 0.10), the suspected difference in specificities between the imaging techniques at a false-negative rate of 0.10 (small = 0.05, moderate = 0.10, and large = 0.15), and the ratio (1/1, 2/1, or 4/1) of patients without to patients with the condition.

[a] 80% power, 5% type I error, two-tailed test.

interpreting sessions. The average range of the difference in the areas under the ROC curve from these two interpretations was 0.05 (the absolute values of the differences for the five observers were 0.072, 0.047, 0.089, 0.011, and 0.029). Refer to Table 2 for the area under the ROC curve and look under high accuracy (because the average area under the ROC curve was nearly 0.90) and large observer variability (because the inter- and intraobserver variabilities were 0.10 and 0.05, respectively). Our particular observer sample size of seven is not included in the table; nor is the situation of multiple regions from the same patient considered in the table. However, we can interpolate that a moderate to large difference in the areas under the ROC curve can be detected with this study design.

Discussion

I presented tables of sample size to provide investigators with a sense of the sample sizes needed for ROC studies and to illustrate the trade-off between the number of patients and the number of observers that can be used for these studies. My tables should serve only as initial ballpark estimates of the sample sizes required for any study. It is important that investigators compute sample size using parameters that reflect their clinical application. For example, investigators should consider the values of sensitivity (or specificity) that they expect (not just the two values I considered), and they should use the false-positive rate or false-negative rate ranges appropriate to their clinical application. Furthermore, investigators should assess the applicability of the assumptions I made. For example, I assumed that the test results from the two imaging techniques were correlated, with a correlation of 0.47. This value is the average correlation reported by Rockette et al. [4] from 20 multiobserver ROC studies. However, if different samples of patients underwent the two different types of imaging, then the correlation would be zero. In this case, the sample sizes reported represent a considerable underestimation of the required sample size.

Other study design issues also affect sample size. For example, in many applications, there is more than one unit per patient; for example, five regions from each of two breasts in evaluating screening mammography [5], or situations with the possibility of two or more lesions in the same patient. These units from the same patient are not statistically independent and cannot be analyzed as independent observations. My tables of sample size assume one unit per patient. When there are multiple units per patient, the sample sizes I presented will usually be an overestimation of the number of patients required. However, it would not be appropriate to assume that the number of units needed is equivalent to the number of patients estimated in my tables; this strategy would result in a study with inadequate power.

Another issue to consider in sample size estimation is the plan for subgroup analyses. Often, investigators want to report accuracy and the comparative accuracies of the two imaging techniques for particular subgroups (subgroups categorized by gender, age, or symptoms). For these analyses, sample sizes larger than those reported here are generally required.

My study of sample size for multiobserver ROC studies has several limitations. First, the sample sizes are derived from one mathematic model [1] of the relationship among accuracy, observer variability, patient variability, and the correlation in accuracy imposed by study design. Other more sophisticated models exist [6, 7], but no associated algorithms are available for calculating sample size. A study by Rockette et al. [4] suggested that for sample size estimation, the simpler model used here is adequate. Second, for simplicity, we made several assumptions about the distributions of the unobserved test results (i.e., binormality with equal variances) (Appendix). These assumptions may not be applicable for all studies; therefore, I encourage readers to assess these assumptions for their particular studies. Finally, I have focused on only one issue of study design: sample size. It is important to recognize that an adequate sample size cannot compensate for systematic biases that are common in our literature [8–13]. Therefore, it is critical that we sample patients and observers carefully for our studies, that we identify and apply appropriate procedures for determining the true diagnoses, and that images are interpreted in a blinded fashion.

References

1. Obuchowski NA. Multiobserver, multiimaging technique receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Acad Radiol* **1995**;[suppl2]:S22–S29
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**;143:29–36
3. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* **1989**;9:190–195
4. Rockette HE, Campbell WL, Britton CA, Holbert JM, King JL, Gur D. Empiric assessment of parameters that affect the design of multiobserver receiver operating characteristic studies. *Acad Radiol* **1999**;6:723–729
5. Powell KA, Obuchowski NA, Chilcote WA, Barry MM, Ganobcik SN, Cardenosa G. Film-screen versus digitized mammography: assessment of clinical equivalence. *AJR* **1999**;173:889–894
6. Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med* **1996**;15:1807–1826
7. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of observers and patients with the jackknife method. *Invest Radiol* **1992**;27:723–731
8. Reid MC, Lachs MS, Feinstein AR. Use of methodologic standards in diagnostic test research: getting better but still not good. *JAMA* **1995**;274:645–651
9. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* **1978**;299:926–930
10. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* **1989**;24:234–245
11. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* **1988**;167:565–569
12. Begg CB. Experimental design of medical imaging trials: issues and options. *Invest Radiol* **1989**;24:934–936
13. Black WC. How to evaluate the radiology literature. *AJR* **1990**;154:17–22
14. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple observers and multiple tests: an ANOVA approach with dependent observations. *Comm Stat* **1995**;24:285–308
15. Steen F. *Elements of probability and mathematical statistics*. Boston: Prindle, Weber & Schmidt, **1982**:404
16. Obuchowski NA. Computing sample size for receiver operating characteristic studies. *Invest Radiol* **1994**;29:238–243
17. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Stat Med* **1997**;16:1529–1542

See Appendix on next page

APPENDIX: Details Regarding the Derivation of Tables 2–4

It is assumed that the null and alternative hypotheses of primary interest are

$$H_o: u_1 = u_2 \qquad H_A: u_1 \neq u_2 \qquad (1)$$

where u_i is the mean diagnostic accuracy of imaging technique i for the population of observers. For sample size estimation, under the alternative hypothesis we specify that $u_1 - u_2 = \Delta$, where Δ is the suspected difference. Obuchowski and Rockette [14] proposed a modified F statistic to test the null hypothesis in equation 1. The noncentrality parameter of the F statistic, denoted λ, can be used to derive sample size estimates [1]. The parameter λ is a function of three variances, four correlations, the number of observers (indicated by J), and Δ,

$$\lambda = \frac{J\Delta^2}{2(\sigma_b^2(1 - r_b) + \sigma_w^2 /K + \sigma_c^2[(1 - r_1) + (J - 1)(r_2 - r_3)]} \qquad (2)$$

where σ_b^2 is the variability in observers' accuracies when using the same imaging technique for the same sample of patients (interobserver variability), σ_w^2 is the variability in a single observer's accuracies when using the same imaging technique for the same sample of patients on different interpreting occasions (intraobserver variability), σ_c^2 is the variability between samples of patients and is a function of the number of patients in the sample, r_1 is the correlation between accuracies estimated from the same sample of patients by the same observer using different imaging techniques, r_2 is the correlation between accuracies estimated from the same sample of patients by different observers using the same imaging technique, r_3 is the correlation between accuracies estimated from the same sample of patients by different observers using different imaging techniques, r_b is the correlation between accuracies obtained when a set of observers examines the same sample of patients using different diagnostic tests, and K is the number of times each observer interprets the same images from the same imaging technique.

For the tables in this paper, I used three different levels of σ_b^2 and σ_w^2. I estimated σ_b^2 and σ_w^2 from the ranges in Table 1 by assuming that observers' accuracies follow a normal distribution. For example, for a study design with 10 observers and a large variability, $\hat{\sigma}_b$ = (range) × (constant derived from the normal distribution [15]) = (0.10) × (0.3249) = 0.0325. From two interpreting sessions by the same observer, $\hat{\sigma}_w$ = (range) × (constant derived from the normal distribution [15]) = (0.05) × (0.8862) = 0.0443. I assumed that $K = 1$ because in most multiobserver investigations, the images from each imaging technique are interpreted only once by each observer.

For the correlations r_1, r_2, r_3, and r_b, we used values from the review article by Rockette et al. [4]. Specifically, we set r_1 equal to the average correlation of all the studies reviewed by Rockette et al. (0.47). Note that the range of values of r_1 was 0.35–0.59. The average value of r_2 minus r_3 was –0.0011. Rockette et al. recommend using zero for sample size estimation, and this is the value I used. For r_b, Rockette et al. recommend a value of 0.8 for sample-size estimation on the basis of data from two large studies; I used this value for my tables.

Note that these variances and correlations were estimated by Rockette et al. [4] for the area under the ROC curve and not for the partial area under the ROC curve. For my tables, I assumed that these variances and correlation values are also reasonable when the partial area under the ROC curve is used.

For different values of J, I used the function PROBF in SAS software (SAS Institute, Cary, NC) to determine the value of the noncentrality parameter that would provide 80% power with a 5% type I error rate. These values of λ are 18.12, 12.36, and 9.92, for J equal to 4, 6, and 10, respectively. I substituted the above values of σ_b^2, σ_w^2, r_1, r_2, r_3, and r_b into equation 2. Then for the different combinations of J and Δ, I solved for σ_c^2. The value of σ_c^2, when positive, represents the maximum patient variance permitted for the study design. Because σ_c^2 is a function of the number of patients, I was able to calculate the number of patients needed for each combination of J and Δ. When the estimate of σ_c^2 was negative or almost zero, I concluded that the number of observers was inadequate for the study design.

For the area under the ROC curve summary measure, an estimate of σ_c^2 for sample size estimation is [16]

$$\hat{\sigma}_c^2 = \left\{(0.0099 \times e^{-A^2/2}) \times ([5A^2 + 8] + [A^2 + 8] / R)\right\} / N \qquad (3)$$

where $A = \Phi^{-1}(\theta) \times 1.414$, and θ is the anticipated area under the ROC curve, Φ^{-1} is the inverse of the cumulative normal distribution function, R is the ratio of sample sizes of patients without the condition to with the condition, and N is the number of patients with the condition. Substituting into equation 3 the anticipated value of the area under the ROC curve (0.75 or 0.90), the value of R (1.0, 2.0, or 4.0), and the maximum patient variance calculated from equation 2, the number of patients needed with the condition can be determined. Then the total patient sample size needed is N (1 + R).

For the partial area under the ROC curve in the false-positive rate range from 0.0 to 0.10 (or for the partial area under the ROC curve in the false-negative rate range from 0.0 to 0.10), we first need to convert the values of Δ and the values for the interobserver and intraobserver ranges given in Table 1 into values for the partial area under the ROC curve. (Note that these values have been defined in terms of the sensitivity [specificity] at a false-positive rate [false-negative rate] of 0.10.) To do this we need to specify an ROC curve among all possible ROC curves that pass through the point (false-positive rate = 0.10; sensitivity = 0.60 [or 0.80]). We assume that the test results of patients with and without the condition follow a binormal distribution with equal variance (binormal parameter $b = 1.0$). Therefore, for sensitivities of 0.60 and 0.80 at a false-positive rate of 0.10, the corresponding binormal parameter a's (the standardized difference in the means of the two distributions) are 1.535 and 2.120, and the corresponding partial areas under the curves are 0.0424 and 0.0637, respectively. For a suspected difference between the two imaging techniques of 0.05 (in terms of sensitivity at a false-positive rate of 0.10), we can define two ROC curves with parameter a's of 1.47 and 1.60, which correspond to sensitivities of 0.575 and 0.625 and partial areas of 0.0400 and 0.0447, respectively. Therefore, I translated a difference of 0.05 in terms of sensitivities at a false-positive rate of 0.10 into a difference in terms of the partial areas of 0.0047. I performed this transformation for each of the values of Δ and the values for the interobserver and intraobserver ranges given in Table 1. I used these transformed values in equation 2 to determine the maximum patient variance for each study design.

Then an estimate of σ_c^2 for the partial areas used for sample-size estimation is [17]

$$\hat{\sigma}_c^2 = \left\{f^2[1 + b^2/R + a^2/2] + g^2[b^2(1 + R) / (2R)]\right\} / N \qquad (4)$$

where a and b are parameters from a binormal distribution, R and N are the same as in equation 3, and f and g are functions of the binormal parameters and the range of false-positive rates of interest [17]. I set $b = 1$ and $a = 1.535$ or 2.120, corresponding to sensitivities of 0.6 and 0.8 at a false-positive rate of 0.1. The corresponding values of f and g are 0.0365 and –0.0619 (for moderate accuracy) and 0.0346 and –0.0613 (for high accuracy). By substituting these values into equation 4, I calculated the number of patients needed with the condition. Then the total patient sample size needed is N (1 + R).