Review Article

Testing for Equivalence of Diagnostic Tests

Nancy A. Obuchowski[1]

The technology of diagnostic radiology is changing rapidly. Associated with these changes is a need to compare technologies with one another in an appropriate and valid fashion. To make meaningful comparisons, appropriate questions must be asked and tested. However, in comparing diagnostic technologies, we often ask and test the wrong question. In this article, an example is described in which the objective of the study is to show equivalence between technologies. Equivalence studies are common in diagnostic radiology; yet equivalence is often poorly defined, and the statistical methods for testing equivalence are rarely applied. The appropriate statistical methods for testing equivalence are presented here along with a methodologic approach for defining equivalence. Equivalence is defined here in terms of patient outcomes (e.g., end points such as the quality and quantity of life), rather than intermediate end points, such as image contrast and resolution or sensitivity and specificity [1].

Background

Determining the Appropriate Question

New technology is often more conservative (i.e., less radiation, less invasive, more convenient) than existing technology. When comparing technologies, determination of whether the more conservative technology is clinically comparable to the existing one is often of interest. For example, suppose one wants to know if fast spin-echo T2-weighted MR imaging can replace conventional T2-weighted imaging of the spine. The fast T2 sequence is not expected to be diagnostically superior; rather, one wants to know if the new sequence is accurate enough to replace the old. Similarly, when comparing electronic imaging with conventional plain film, determination of whether the electronic version leads to similar patient outcomes as plain film is important. These examples illustrate the comparison of the same test but with different acquisitions and recordings, respectively. A third example is the comparison of two tests. Suppose one wants to determine if three-dimensional time-of-flight MR angiography (MRA) can replace catheter angiography (CA) as a presurgical tool for carotid endarterectomy. MRA may not be as sensitive or specific as CA. However, the lesser accuracy of MRA must be weighted against the morbidity and mortality associated with CA in determining if MRA is a suitable replacement.

In situations similar to these three examples, the question is not whether the new technology is diagnostically superior to the old. Instead, the question is whether the new technology is clinically equivalent to the old.

Irrespective of the appropriate study question, in diagnostic radiology the former question is almost always tested. Specifically, data are collected on two competing technologies, and then a statistical test is performed to detect differences. If the difference between them is not statistically significant (i.e., conventionally, the p value > .05), then it is often incorrectly concluded that the two technologies are diagnostically equivalent. However, a nonsignificant difference does not imply equivalence. Two interrelated reasons why a nonsignificant difference does not imply equivalence are the statistical error rate (particularly the type II error rate, defined next) is not negligible, and equivalence has not been defined.

Why Nonsignificance Does Not Imply Equivalence

When testing for differences, a type II error is the probability of obtaining a nonsignificant p value (i.e., p value > .05) when a difference really exists. (A related concept is statistical power, defined as 1.0 – type II error rate.) Studies with large samples are less likely to suffer a type II error than studies with smaller samples. For any given study, however, the probability of a type II error is rarely negligible (i.e., rarely ≤5%). In fact, it is often 20% or greater, meaning that when a statistical test has an associated p value of greater than .05, it is not appropriate to conclude that the technologies are equivalent due to the recognized possibility of a type II error. Instead, one must conclude something like there is insufficient evidence to show a dif-

Received March 25, 1996; accepted after revision July 31, 1996.

[1]Departments of Biostatistics and Epidemiology and Radiology, The Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH 44195-5196. Address correspondence to N. A. Obuchowski.

AJR 1997;168:13–17 0361–803X/97/1681–13

ference smaller than a specific amount. What specific amount is small enough to conclude equivalence? The answer to this question depends on the clinically relevant difference.

A clinically relevant difference is the difference between two technologies that has clinical implications. Let (Δ_L, Δ_U) denote an interval, such that the two technologies are considered to be equivalent (i.e., the difference is clinically ignorable) if the difference in their accuracies is between Δ_L and Δ_U (L = lower bound of the interval; U = upper bound of the interval). Differences less than or equal to Δ_L and differences greater than or equal to Δ_U are clinically important. For example, suppose an investigator wants to determine if digitized film is equivalent to conventional plain film for detecting breast cancer on screening mammograms. For illustration, let the sensitivity of the recordings serve as the measure of diagnostic accuracy, and suppose that we have defined these two recordings to be clinically equivalent if their sensitivities are within 10% of each other's (i.e., $\Delta_L = -10\%$ and $\Delta_U = +10\%$). (Later, we present a method for defining Δ_L and Δ_U.) The idea is depicted in Figure 1, where the two recordings are clinically equivalent if the sensitivity of plain film minus the sensitivity of digitized film is between −10% and +10%; the difference in sensitivities is clinically relevant if the difference is less than or equal to −10% or greater than or equal to 10%.

Note that it is not necessary for the interval to be symmetric around zero. In fact, the interval might not even contain zero. This situation might occur when both short- and long-term outcomes of the tests are considered in defining Δ_L and Δ_U. For example, one might compare CA, which is accurate but invasive, to MRA, which might be less accurate but noninvasive. If the risk of morbidity and mortality associated with CA outweighs the consequences of somewhat less accurate diagnoses by MRA, then the accuracy of MRA can be considerably less than the accuracy of CA for the two tests to be clinically equivalent. In this scenario, if the diagnostic accuracy of MRA and CA were identical, then MRA might be preferred over CA (i.e., the two tests would not be clinically equivalent).

In this paper, I first illustrate the statistical methods for testing for equivalence. A statistic specifically for testing equivalence is presented, then I illustrate how a confidence interval for the difference can also be applied to address the question of equivalence. Finally, I describe a method, based on decision analysis, for defining Δ_L and Δ_U.

Methods

A fictitious example from diagnostic radiology is used to illustrate the methods of testing for equivalence and defining equivalence. In the example, an investigator wants to determine if digitized film is clinically equivalent to conventional plain film for detecting breast cancer on screening mammography. A study is conducted whereby a radiologist interprets the images of 200 patients displayed on plain and digitized film. Biopsy results or diagnoses of cancer from long-term follow-up serve as the gold standard. Receiver operating characteristic curves are constructed for plain film and digitized film. From the receiver operating characteristic curve, various measures of diagnostic accuracy could be considered, including the area under the full curve [2], the area under a particular portion of the curve [3], and the sensitivity (or specificity) at various fixed specificities (or fixed sensitivities). For illustration purposes, we use sensitivity at a constant specificity of 90%; however, other measures of accuracy can also be used. We denote the sensitivity of plain film as θ_{PF} and the sensitivity of digitized film as θ_{DF}.

Results

Testing for Diagnostic Equivalence

The statistical methods for testing equivalence are not new. They are used frequently in pharmacology to test equivalence of bioavailability of drug formulations [4, 5]. In the paragraphs that follow, I will describe the two one-sided hypothesis tests procedure proposed by Schuirmann [5].

In the mammography example, suppose that the estimated sensitivities are 90% for plain film and 85% for digitized film at a constant specificity of 90%. We denote the estimated sensitivities of plain film and digitized film by $\hat{\theta}_{PF}$ and $\hat{\theta}_{DF}$, respectively. The estimated difference between the sensitivities, denoted by $d = \hat{\theta}_{PF} - \hat{\theta}_{DF}$, is 5%. (Ideally, one should transform the estimated sensitivities to normal deviate space and perform these calculations on the transformed data [6]. However, here I am trying simply to clarify the necessary calculations.)

The first step to testing for equivalence is defining Δ_L and Δ_U. The definition should be specified before the study begins or at least before exploring the data. In the next section, I present a formal method for determining Δ_L and Δ_U. For now, however, let us assume arbitrarily that the two recordings have been defined to be equivalent if their sensitivities are within ± 10% of each other. In other words, if the difference in sensitivities is greater than −10% and less than +10%, then the two tests are considered to be clinically equivalent. Thus, $\Delta_L = -10\%$ and $\Delta_U = +10\%$.

Next, refer to Table 1, where the statistical test for equivalence is summarized alongside the statistical test for differences. For the test for equivalence, the null hypothesis is that the difference in diagnostic accuracy between plain film and digitized film is either less than or equal to Δ_L or greater than or equal to Δ_U. The alternative hypothesis is that the tests are equivalent. This set of hypotheses is reversed for the test of differences. In particular, for the test of differences, the null hypothesis is that of equivalence and the alternative hypothesis is that of a difference.

Similarly, the type I and II errors of a test for equivalence are analogous to the type II and I errors, respectively, of a test for differences. This distinction is important. In particular, in studies in which the objective is to assess the equivalency of two technologies, a false claim of equivalence can have enormous conse-

TABLE 1 Testing for Differences Versus Testing for Equivalence

Concept	Test for Differences[a]	Test for Equivalence
Hypotheses	H_0: $(\theta_1 - \theta_2) = 0$ H_A: $(\theta_1 - \theta_2) \neq 0$	H_0: $(\theta_1 - \theta_2) \leq \Delta_L$ or $(\theta_1 - \theta_2) \geq \Delta_U$ H_A: $\Delta_L < (\theta_1 - \theta_2) < \Delta_U$
Statistical test	$z = \lvert d \rvert / SE(d)$	$z_1 = \{d - \Delta_L\}/SE(d)$ and $z_2 = \{\Delta_U - d\}/SE(d)$
Reject H_0 if	$z > CR_{\alpha/2}$	$z_1 > CR_\alpha$ and $z_2 > CR_\alpha$
Type I error	False claim that two tests differ, when, in truth, they are equivalent	False claim that two tests are equivalent, when, in truth, they differ
Type II error	False claim that two tests are equivalent, when, in truth, they differ	False claim that two tests differ, when, in truth, they are equivalent

Note.—H_0 = null hypothesis; H_A = alternative hypothesis; θ_i = diagnostic accuracy of test i in the population of patients; Δ_L and Δ_U = lower and upper values of the equivalence interval, respectively; d = estimated difference in diagnostic accuracy between the two tests; $SE(d)$ = standard error of d; CR_α = critical value for the upper α% of the standard distribution.

[a]The test for differences should be followed by construction of the confidence interval for the difference.

quences: a standard test might be replaced with an inferior one that puts the public at risk. Thus, this type of error must be kept to a minimum (usually ≤5%). In designing a study to test equivalence, the type I error rate of the test for equivalence should be set to less than or equal to 5%; if using the test for differences, the type II error rate should be set to less than or equal to 5%.

Using the test for equivalence from the second column of Table 1, we have two test statistics: $z_1 = \{d - \Delta_L\}/SE(d)$ and $z_2 = \{\Delta_U - d\}/SE(d)$, where $SE(d)$ is the estimated standard error of d, given by

$$SE(d) = \sqrt{Var(\hat{\theta}_{PF}) + Var(\hat{\theta}_{DF}) - 2 \times Cov(\hat{\theta}_{PF}, \hat{\theta}_{DF})}$$

where $Var(\hat{\theta}_{PF})$ is the variance of the estimated sensitivity of plain film and $Cov(\hat{\theta}_{PF}, \hat{\theta}_{DF})$ is the covariance between the estimated sensitivities of the two recordings. Note that the covariance between the estimated sensitivities is important here because in this fictitious study the same images were displayed on both plain and digitized film; thus, the estimates of sensitivity are likely to vary together (i.e., covary). Methods for estimating the variance and covariance are given by Wieand et al. [7]; these methods reflect the uncertainty in estimating the specificity (as well as the uncertainty in estimating the sensitivity). Let us suppose that the $SE(d)$ was estimated to be 0.036. Then, $z_1 = \{0.05 + 0.10\}/0.036 = 4.17$ and $z_2 = \{0.10 - 0.05\}/0.036 = 1.39$.

Compare z_1 and z_2 to the appropriate standard distribution (often, the standard normal distribution). If z_1 and z_2 are both greater than the critical value, CR, for the upper α% of the standard distribution, then reject the null hypothesis and conclude that the two recordings are equivalent. α is the significance level of the test. We set α equal to the maximum type I error rate that we are willing to allow. Usually α is set to 0.05, corresponding to a 5% risk of incorrectly concluding that the two recordings are equivalent when, in truth, they differ. For a significance level of .05, the critical value from the standard normal distribution is $CR_{0.05} = 1.645$. We compare each one-sided test to 1.645. Although z_1 is greater than 1.645, z_2 is not greater than 1.645. Thus, we do not conclude that plain and digitized film have equivalent diagnostic accuracies.

Note that the observed difference in sensitivities of plain and digitized film (i.e., +5%) lies between Δ_L and Δ_U. However, the data are insufficient for ruling out the possibility that the real difference (i.e., the difference for the population of patients, not just for this sample of patients) is greater than or equal to +10% (i.e., that plain film is more sensitive than digitized film).

Now, let us consider the test for differences summarized in the first column of Table 1. The test statistic for differences is $z = |d|/SE(d) = 0.05/0.036 = 1.39$. For a significance level of .05, the critical value from the standard normal distribution is $CR_{\alpha/2} = 1.96$. Because 1.39 is not greater than 1.96, we do not conclude that there is a difference in sensitivities between the two recordings.

Although we cannot conclude that the sensitivities differ, we cannot automatically conclude that they are equivalent. To make a statement about the possible equivalence between the two recordings, based on this method, we need to examine the confidence interval (CI) of the difference in sensitivities. A $(1 - \alpha)$% CI is the interval in which we are $(1 - \alpha)$% confident that the unknown value of the difference lies. The $(1 - \alpha)$% CI is given by $d \pm CR_{\alpha/2} \times SE(d)$. The 95% CI for the difference in sensitivities is (–0.02, 0.12). This CI contains values that are not in the equivalence interval (Δ_L, Δ_U). Specifically, values greater than or equal to +0.10 are not in the equivalence interval. Since we cannot rule these values out, we do not conclude that the two recordings are equivalent.

Note that even when a significant difference is found between two technologies, a CI for the difference should still be constructed because it is possible that in a very accurate study, one might detect a statistically significant difference when, in truth, the tests are equivalent as defined by the interval (Δ_L, Δ_U). Thus, regardless of the result of the test for differences, it is critical to examine the CI for the difference to determine if it is entirely contained in the interval (Δ_L, Δ_U).

In summary, both methods (i.e., the test for equivalence and the CI approach) can be used to address the question of equivalence. In fact, Munk [8] presents a proof that the test for equivalence at a significance level of α (i.e., each one-sided test is performed at α) is analogous to a $(1 - 2\alpha)$% CI procedure. For either approach, the interval (Δ_L, Δ_U) must be defined before examining the data to avoid biases.

Defining the Interval (Δ_L, Δ_U)

We propose a strategy for defining the interval (Δ_L, Δ_U) that uses decision analysis to estimate the short- and long-term patient outcomes associated with diagnostic testing. This strategy is to define a range of patient outcomes in which the differences are clinically ignorable. Then, convert this range of patient outcomes into a range of diagnostic accuracies, thus defining the interval (Δ_L, Δ_U).

In defining the interval (Δ_L, Δ_U) for the mammography example, we use patient survival at 10 years as the measure of patient outcome. This patient outcome measure was chosen because it is the primary measure of efficacy in many studies of the management of breast carcinomas; to my knowledge, other measures of patient outcome, such as health-related quality of life, have not been studied [9]. Later, I will discuss why this measure is a crude measure of patient outcome and describe alternative approaches.

The first step in this strategy is to delineate all possible downstream events associated with diagnostic testing. To do this, construction of a decision tree is helpful (Fig. 2). Starting at the far left on such a tree, the first entry indicates screening by mammography. The split that follows indicates the result is either positive or negative for malignancy. If the test result is negative, the patient is not treated. The negative result is either a true negative or a false negative, depending on the accuracy of the test and the prevalence of disease. If the test is positive, the lesion is biopsied. We assume here that biopsy is perfectly accurate at distinguishing benign from malignant lesions. A positive biopsy is followed by surgery.

The second step is to compute the difference in average patient outcome for patients studied with plain film versus patients studied with digitized film. We first describe this difference in terms of symbols because many of the parameters will cancel out and, thus, we will not need to estimate them. We denote the prevalence of a malignant lesion by P, and the sensitivity and specificity of a test by Se and Sp, respectively. The expected 10-year survival rates are denoted as follows: a true negative is $U_{ND,NT}$ (ND = no disease, NT = no follow-up or treatment), a false negative is $U_{D,NT}$, a true positive is $U_{D,T}$, and a false positive is $U_{ND,T}$.

Using these symbols, the probability of the four possible outcomes is a true negative = $(1.0 - P) \times Sp$, a false negative = $(P) \times (1.0 - Se)$, a true positive = $(P) \times (Se)$, and a false positive = $(1.0 - P) \times (1.0 - Sp)$. The expected survival rate is the sum of the survival rates of these four outcomes, weighted by the probability of the outcome: survival = $(1.0 - P) \times (Sp) \times U_{ND,NT} + (P) \times (1 - Se) \times U_{D,NT} + (P) \times (Se) \times U_{D,T} + (1 - P) \times (1 - Sp) \times U_{ND,T}$.

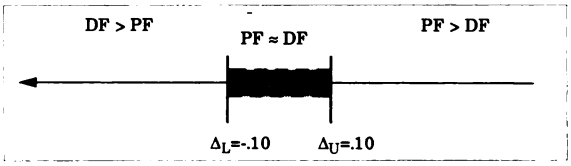

Fig. 1.—Comparison of digitized film (DF) and plain film (PF). Shaded area shows interval in which PF and DF are clinically equivalent; beyond this region, two recording techniques are not clinically equivalent. DF > PF means that DF provides clinically defined better patient outcomes than PF; PF > DF means that PF provides clinically defined better patient outcomes than DF. The lower and upper values of accuracy defining interval where techniques are clinically equivalent are given by Δ_L and Δ_U, respectively.

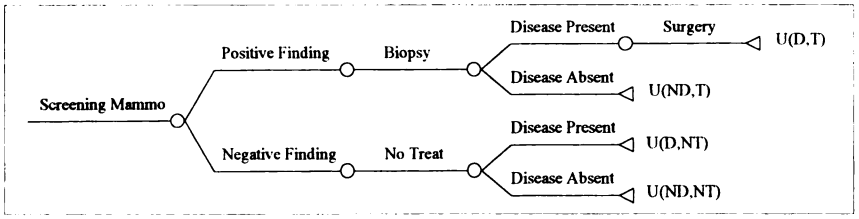

Fig. 2.—Decision tree for screening mammography shows downstream events associated with diagnostic testing. Starting at far left, first entry indicates screening by mammography. Split that follows indicates result is either positive or negative for malignancy. If test result is negative, patient is not treated. Negative result is either true negative or false negative. If test is positive, lesion is biopsied, and positive biopsy is followed by surgery. Expected outcome for true negative is denoted U(ND,NT) (ND = no disease, NT = no follow-up or treatment), for false negative is U(D,NT), for true positive is U(D,T), and for false positive is U(ND,T).

Because the expected 10-year survival rate for nondiseased patients is the same whether they are untreated or undergo biopsy, we set $U_{ND,NT} = U_{ND,T} = U_{ND}$.

We are interested in the difference in expected survival between plain and digitized film. With some algebra it can be shown that this difference is given by equation 1.

$$\text{Survival}_{PF} - \text{Survival}_{DF} = (P) \times (U_{D,T} - U_{D,NT}) \times (Se_{PF} - Se_{DF}). \quad (1)$$

Thus, we do not need to estimate U_{ND} or the specificity of the two recordings.

The 10-year survival rate from breast cancer after surgery (i.e., $U_{D,T}$) is estimated at 70% [10]. The 10-year survival rate from a malignant lesion undetected by screening (i.e., $U_{D,NT}$) is approximately 47% [11]. Note that this estimate may be too low because some lesions may remain clinically occult for several years and may be detected at a later screening. For illustration purposes, we use the 47% estimate; however, one might consider a range of possible values and assess the impact on the interval (Δ_L, Δ_U). Finally, the estimated annual incidence rate of breast cancer in 55-year-old women is 0.0023 [11]. For women screened regularly, we assume that the prevalence rate of breast cancer, P, is roughly 0.0023, although it may be slightly higher. Plugging these estimates into equation 1, we obtain

$$\text{Survival}_{PF} - \text{Survival}_{DF} = 0.053 \times (Se_{PF} - Se_{DF}).$$

The third step in our strategy is determining the minimally clinically relevant difference in terms of the patient outcome measure we have chosen. For the mammography example, we must define the minimally clinically relevant difference in terms of 10-year survival rates. This is a critical step in the process and requires careful thought because this step has a large impact on the resulting values of Δ_L and Δ_U. For illustration purposes, we define the minimally clinically relevant difference as a difference in 10-year survival of 0.001% (i.e., one additional death per 100,000 patients). Differences less 0.001% will be considered clinically equivalent.

The last step in our strategy is to convert the interval of equivalent patient outcomes, that is, (–0.001%, 0.001%), to a range of diagnostic accuracies. From equation 1, we set

$$\text{Survival}_{PF} - \text{Survival}_{DF} = 0.053 \times (Se_{PF} - Se_{DF}) = -0.001\%$$

and

$$\text{Survival}_{PF} - \text{Survival}_{DF} = 0.053 \times (Se_{PF} - Se_{DF}) = +0.001\%.$$

Solving for $(Se_{PF} - Se_{DF})$, we get –0.019 and +0.019, respectively. Thus, Δ_L = –0.019 and Δ_U = +0.019. In words, plain film and digitized film have equivalent clinical outcomes, defined as a difference of less than one additional death in 100,000 patients over a 10-year period, if the difference in their sensitivities is in the interval (–0.019, 0.019).

Recall that in the initial part of this example, we used Δ_L = –0.10 and Δ_U = +0.10 and set the specificities of both recordings to 90%. However, on the basis of our new definition of equivalence, the interval is given by Δ_L = –0.019 and Δ_U = +0.019, and the specificities of the two recordings do not need to be equivalent.

In other examples, the specificities of the two competing technologies will not cancel out as they did in equation 1. When both sensitivity and specificity are functions of the difference in patient outcomes between the competing techniques, we prefer to equate the specificities of the two tests, so that the interval (Δ_L, Δ_U) is defined in terms of the difference in the two tests' sensitivities. We then can trace out a receiver operating characteristic region where the height of the reigon corresponds to the required sensitivity of the challenging test. Phelps and Mushlin [12] refer to this region as the challenge region.

In other applications, the decision tree and its analysis may be more complicated than in this example. For example, in comparing MRA to CA, one should consider the complications associated with CA as well as the downstream effects of testing. Regardless of the form of the decision tree, our strategy is the same: delineate the possible outcomes associated with diagnostic testing, compute the difference in expected patient outcome between the two competing technologies in terms of the differences in their accuracies, define the minimally clinically relevant difference, and convert the range of clinical equivalence to a range of diagnostic accuracies.

Defining patient outcome in terms of survival rates is rather crude for several reasons. First, this approach ignores patient morbidity. Morbidity is associated with biopsy and surgery and may also be associated with a false-negative result (i.e., false reassurance and later regret [13]). If we ignore these outcomes, then we undervalue the impact of the test [13]. However, we can expand the decision tree in Figure 2 to incorporate health states other than being alive or dead. Then, we must quantify the quality of life in the different health states. Methods exist for doing this [14]. (Note that in the mammography example, if we had accounted for the morbidity associated with biopsy, then the

specificities of plain film and digitized film would not have canceled out of equation 1.) Second, our approach ignores the timing of events. Patients with mammographically detected lesions who undergo surgery may have earlier staged lesions and thus may live longer than patients whose disease was undetected by screening. Thus, the timing of events is important. To incorporate the timing of events into a measure of patient outcome, we must estimate the length of time a patient spends in each health state. Then, we can define patient outcome in terms of units called quality-adjusted life-years [14], which is a measure of both the quality and the quantity of life.

Discussion

The objective of this paper was to emphasize the role of equivalence studies in diagnostic radiology. I have tried to achieve this by pointing out the unique aspects of an equivalence study and by presenting the appropriate methods for testing for equivalence.

The methods presented here have dealt strictly with testing whether two technologies are equivalent. However, situations may occur where two technologies are not equivalent, but one technology is a suitable replacement for the other. For example, a new technology could be more accurate and have fewer risks than an existing technology. In this situation, the two technologies are not equivalent (because the new one is better than the existing one), but the new technology is a suitable replacement for the existing technology. If the objective of a study is to test the ability of the new technology to replace the existing technology, then a one-sided hypothesis is appropriate, rather than the two-sided equivalence hypothesis. Thus, the relevant null hypothesis is that the new technology is not as accurate as the existing technology. The alternative hypothesis is simply that the new technology is at least as accurate as the existing technology.

Probably the most important part of an equivalence study (and a replacement study) is the definition of equivalence. This is a complicated issue because the results of diagnostic imaging have both short- and long-term effects that should be accounted for in the definition of equivalence. Our strategy for defining equivalence relies on the concepts and methods of decision analysis, which uses current literature for estimates of the risks and benefits of existing therapies. However, these estimates are subject to modification whenever new studies are completed. Furthermore, current therapies may be replaced by improved therapies. Thus, equivalence defined by our method is conditional on what is currently known about patient outcomes and needs to be revised as we increase our knowledge about the natural history of disease and the efficacy of treatments.

Acknowledgments

I thank two reviewers whose comments improved the presentation of this manuscript and Dr. Michael T. Modic for supporting this effort.

References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* **1991**;11: 88–94
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. Anderson S, Hauck WW. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun Stat: Theor Methods* **1983**;12:2663–2692
5. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm* **1987**;15:657–680
6. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making* **1984**;4:137–150
7. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **1989**;76:585–592
8. Munk A. On a method of combining double t-test and Anderson-Hauck test (letter). *Biometrics* **1994**; 50:885–886
9. Kattlove H, Liberati A, Keeler E, Brook RH. Benefits and costs of screening and treatment for early breast cancer. *JAMA* **1995**;273:142–148
10. Sacks NPM, Baum M. Primary management of carcinoma of the breast. *Lancet* **1993**;342:1402–1408
11. Eddy DM. Screening for breast cancer. *Ann Intern Med* **1989**;111:389–399
12. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making* **1988**;8:279–289
13. The "Value" Working Group. Defining and measuring the "value" of diagnostic imaging. *J Magn Reson Imaging* **1996**;1:7–9
14. Yin D, Forman HP, Langlotz CP. Evaluating health services: the importance of patients' preferences and quality of life. *AJR* **1995**;165:1323–1328