Radiology

How to Demonstrate Similarity by Using Noninferiority and Equivalence Statistical Testing in Radiology Research[1]

Soyeon Ahn, PhD
Seong Ho Park, MD
Kyoung Ho Lee, MD

Demonstrating similarity between compared groups—that is, equivalence or noninferiority of the outcome of one group to the outcome of another group—requires a different analytic approach than determining the difference between groups—that is, superiority of one group over another. Neither a statistically significant difference between groups ($P < .05$) nor a lack of significant difference ($P \geq .05$) from conventional statistical tests provides answers about equivalence/noninferiority. Statistical testing of equivalence/noninferiority generally uses a confidence interval, where equivalence/noninferiority is claimed when the confidence interval of the difference in outcome between compared groups is within a predetermined equivalence/noninferiority margin that represents a clinically or scientifically acceptable range of differences and is typically described by Δ. The equivalence/noninferiority margin should be justified both clinically and statistically, considering the loss in the main outcome and the compensatory gain, and be chosen conservatively to avoid making a false claim of equivalence/noninferiority for an inferior outcome. Sample size estimation needs to be specified for equivalence/noninferiority design, considering Δ in addition to other general factors. The need for equivalence/noninferiority research studies is expected to increase in radiology, and a good understanding of the fundamental principles of the methodology will be helpful for conducting as well as for interpreting such studies.

[1] From the Medical Research Collaborating Center (S.A.) and Department of Radiology (K.H.L.), Seoul National University College of Medicine, Seoul National University Bundang Hospital, Gyeonggi-do, Korea; and Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 138-736, Korea (S.H.P.). Received April 5, 2012; revision requested May 22; revision received June 1; accepted June 15; final version accepted June 29. Supported by a grant from the Korea Healthcare Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (A102065). **Address correspondence to** S.H.P. (e-mail: *parksh.radiology@gmail.com*).

In most research studies, researchers want to prove a significant difference between compared groups, for example, a different treatment effect in patients treated with therapy A versus therapy B or a different diagnostic performance between test A and test B. Thus, a significant difference implies that one treatment or diagnostic test is better than the other. This type of study is colloquially referred to as a superiority (though "inequality" may be a more statistically accurate description) study, in which a statistically significant difference is typically determined with P values calculated by using well-known conventional statistical tests such as the t test and Fisher exact test (the null hypothesis [H_0]: outcome of group A = outcome of group B versus the alternative hypothesis [H_1]: outcomes of group A ≠ outcome of group B) with $P < .05$ being the traditional criterion for a significant difference by rejecting H_0 at a two-sided 5% significance level. Conversely, in some studies, researchers want to prove similarity between the compared entities: for example, if the diagnostic performance of a new test is similar to (or at least not worse than) the existing standard test. In such cases, it may sound reasonable at a glance to apply conventional statistical tests used for superiority testing, and then conclude a similarity when the P value is greater than .05. However, this approach is invalid. Although $P \geq .05$ from superiority testing may provide some "circumstantial evidence" of similarity between compared groups, particularly when the testing is well powered with a large sample size, superiority testing in itself cannot definitively answer the question about similarity. Superiority testing is intended to confirm the presence of significant difference between compared groups and only concerns whether the difference is likely or unlikely to be zero. $P \geq .05$ from superiority testing indicates that the range of statistically possible differences between compared groups includes zero, whereas $P < .05$ means that the difference is unlikely to be zero and thus the groups ought to differ. As P value of superiority testing does not specifically explain how small or large the group difference could statistically be, $P \geq .05$ can only ascertain not enough evidence of difference but cannot conclude a similarity. Moreover, $P \geq .05$ from superiority testing can simply be due to the lack of statistical power to demonstrate the difference, that is, small sample size (1–3).

Definitive demonstration of similarity between compared groups requires a separate statistical logic known as equivalence/noninferiority testing. Studies to prove similarity can be divided into two categories. An equivalence study is a study to prove "equality" between compared groups, whereas a noninferiority study determines if one group is not worse than (ie, not inferior to) the other group. As most clinical studies of this type are noninferiority studies, the term equivalence is often used loosely synonymously with noninferiority in the literature (4). Equivalence/noninferiority testing requires a different statistical inference than the statistical testing for superiority. Both a lack of statistically significant difference ($P \geq .05$) and a statistically significant difference ($P < .05$) between groups from superiority testing can indicate either equivalence/noninferiority or a lack of them (this will be further discussed in the next section) depending on the case. Nevertheless, misinterpreting a failure to reject H_0 of two-sided superiority testing (ie, $P \geq .05$ from conventional statistical tests for superiority testing) as evidence of a similarity has been a widespread problem in therapeutic medical research studies (1–3,5,6). A similar problem may also exist in radiologic or diagnostic imaging research studies, although, to our knowledge, no discrete data are yet available.

Equivalence/noninferiority research studies are uncommon in radiology literature (7). One probable reason is that radiology has been a highly technology-driven field and has always been on the forefront of new medical technology with many new technologies continuously emerging. Therefore, typical radiology research studies have sought to prove superiority of new technologies, which are expectedly better, to previous technologies. However, the need for equivalence/noninferiority radiologic research studies is expected to increase. From the diagnostic imaging viewpoint, some imaging techniques are now quite mature in diagnostic performance and

Essentials

- Research studies to demonstrate similarity between compared groups—that is, equivalence or noninferiority of the outcome of one group to the outcome of another group—are expected to increase in radiology.
- Demonstrating similarity between compared groups requires a different analytic approach than determining a difference between groups—that is, superiority of one group to another—and conventional statistical tests cannot provide answers about equivalence/noninferiority.
- Statistical equivalence/noninferiority is claimed when the confidence interval of the difference in outcome between compared groups is within a predetermined equivalence/noninferiority margin that is typically described by Δ.
- The equivalence/noninferiority margin should be justified both clinically and statistically, considering the loss in the main outcome and the compensatory gain, and be chosen conservatively to avoid making a false claim of equivalence/noninferiority for an inferior outcome.
- Sample size estimation needs to be specified for equivalence/noninferiority design, considering Δ in addition to other general factors.

Published online
10.1148/radiol.12120725 **Content code:** RS

Radiology 2013; 267:328–338

Abbreviations:
CI = confidence interval
H_0 = null hypothesis
H_1 = alternative hypothesis
SD = standard deviation

Conflicts of interest are listed at the end of this article.

efforts are being made to make the imaging techniques safer, more convenient, and less costly while maintaining the diagnostic performance. One example is emerging techniques to reduce the computed tomography (CT) radiation dose. The purpose of any research study to compare reduced-dose CT with conventional standard-dose CT would be proving that reduced-dose CT would work diagnostically as accurately as standard-dose CT rather than proving that one is better than the other. Therefore, an equivalence/noninferiority design would be more appropriate than a conventional superiority study. Additionally, as interventional radiologic treatments improve, an increasing number of interventional procedures could potentially replace more invasive traditional standard treatments. Thus, an evidence-based approach would require proving the equivalence/noninferiority of the efficacy of the interventional procedures to the standard treatments.

To this end, the purpose of this article is to provide a conceptual review of the principles of equivalence/noninferiority testing, particularly from the radiology research perspective. The article will focus on issues related to equivalence/noninferiority statistical testing and interpretation, including the statistical concept of equivalence/noninferiority, the determination of equivalence/noninferiority margin, and sample size calculation and appropriate statistical testing for an equivalence/noninferiority study. However, this review does not intend to cover all the issues related to the design, conduct, and reporting of an equivalence/noninferiority study. Design, conduct, and reporting of an equivalence/noninferiority trial require more comprehensive consideration throughout every step of the research study and are beyond the scope of the present article. However, related information can be found elsewhere (4,8–11).

Statistical Concept of Equivalence/Noninferiority

Suppose a hypothetical study in which one wants to determine if the sensitivity of a new test (P_{New}) is significantly different from the sensitivity of a conventional test (P_{Conv}). The statistical analysis would adopt superiority testing, typically by using the P value calculated either with McNemar test (for paired data) or Fisher exact test (for unpaired parallel data) (H_0: $P_{New} = P_{Conv}$ versus H_1: $P_{New} \neq P_{Conv}$), where $P < .05$ would reject the H_0 and demonstrate a significant difference between the two sensitivities. Although less commonly used in published articles, the same statistical testing can also be performed by using the two-sided 95% confidence interval (CI) of the difference between the two proportions ($P_{New} - P_{Conv}$). As the logical difference between superiority testing and equivalence/noninferiority testing can be more easily understood by means of the CI approach than P values, the following explanations regarding superiority testing will primarily use the CI approach. A 95% CI of the difference indicates that one can be 95% sure that the CI includes the true difference between the population proportions—that is, if the same experiment/sampling were to be performed multiple times, with a different CI calculated each time, the CIs will include the true difference 95% of the time. If the 95% CI of $P_{New} - P_{Conv}$ includes zero, it has the same meaning as $P \geq .05$ from the McNemar or Fisher exact test. Conversely, if the 95% CI does not include zero, then the P value must be less than .05.

What if the conventional test is a known standard test (typically referred to as an active control or active comparator in equivalence/noninferiority study), and one wants to determine if the sensitivity of a new test (referred to as P_T hereafter, where T stands for test method) is equivalent/noninferior to the sensitivity of the conventional test (referred to as P_{AC} hereafter, where AC stands for active control method)? Although modified special statistical tests to calculate P values for equivalence/noninferiority testing, by which not $P \geq .05$ but $P < .05$ concludes equivalence/noninferiority, exist, they are less commonly used (4,12). Statistical testing for equivalence/noninferiority is generally based on the more informative CI approach (4,9–11). Unlike the use of CI for superiority testing, equivalence/noninferiority testing requires a predefined range of outcome differences ($P_T - P_{AC}$ in the example) that will be considered equivalent/noninferior (Figs 1–3). Equivalence/noninferiority statistical testing is not to prove the exact equality of outcomes between the compared groups but to prove whether the outcomes do not differ enough to be clinically or scientifically relevant; the range of equivalence/noninferiority defines the clinically or scientifically acceptable range of differences. The boundaries of the range are referred to as equivalence margins (Fig 1) or noninferiority margins (Figs 2 and 3), depending on the nature of analysis. The halfway width of the equivalence range, that is, the distance from a $P_T - P_{AC}$ of 0 to either bound of the range, is often referred to as delta, symbolized as Δ (Figs 1–3). For equivalence testing, in which one wants to prove that P_T is equivalent to P_{AC} (H_0: $P_T - P_{AC} \geq \Delta$ or $P_T - P_{AC} \leq -\Delta$ versus H_1: $-\Delta < P_T - P_{AC} < \Delta$), a two-sided CI of $P_T - P_{AC}$ (typically a two-sided 95% CI, assuming a two-sided 5% significance level is acceptable) is used. An equivalence of P_T to P_{AC} is then inferred when the entire CI is within the equivalence range; however, if a part of or the entire CI lies outside the equivalence margins, P_T is not equivalent to P_{AC} (ie, no evidence of equivalence) (Fig 1). If the observed point estimate of outcome difference in the sample is within the equivalence range and the CI lies across the equivalence margin (as shown in B, C, and E of Fig 1), the result could be viewed somewhat inconclusive. Although the data per se do not prove equivalence, a larger sample with a narrower CI may have shown equivalence. Therefore, confirmation of whether the study was adequately powered is necessary in such a case. If the observed point estimate of outcome difference of the sample lies outside the equivalence range (as shown in F of Fig 1), a lack of equivalence is clearer. As shown in Figure 1, a lack of statistically significant difference does not assure statistical equivalence (B and C of Fig 1) and, on

Figure 1

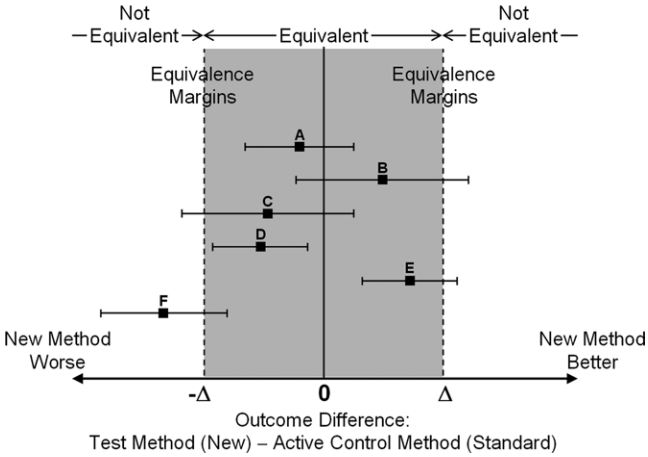

Figure 1: Interpretation of equivalence of a new method (test) to the standard method (active control). A greater outcome value indicates a better outcome. Shaded area = equivalence range ($-\Delta$ to Δ). ■ = observed point estimate of outcome difference in each sample, corresponding error bar = two-sided 95% CI (caps at each end = lower and upper bounds of CI). *A*: Test method is not significantly different from active control method because two-sided 95% CI crosses the 0 outcome difference (same as $P \geq .05$ from conventional statistical tests) and is equivalent to active control method, as entire CI is within equivalence range. *B* and *C*: Test method is not significantly different from active control method but is not equivalent to active control method because part of CI lies outside equivalence range. *D*: Test method is significantly different from active control method, as two-sided 95% CI does not cross the 0 outcome difference (same as $P < .05$ from conventional statistical tests) but is equivalent to the active control method. This puzzling case is rare, since it requires a very large sample size. It can also result from having too wide an equivalence margin. *E* and *F*: Test method is significantly different from active control method and is not equivalent to active control method. Although the data in *B*, *C*, and *E* do not prove equivalence per se, a larger sample with a narrower CI may have shown equivalence. Therefore, confirming whether the study was adequately powered is necessary in such a case. If the observed point estimate of outcome difference of the sample lies outside the equivalence range (as for *F*), a lack of equivalence is clearer.

Figure 2

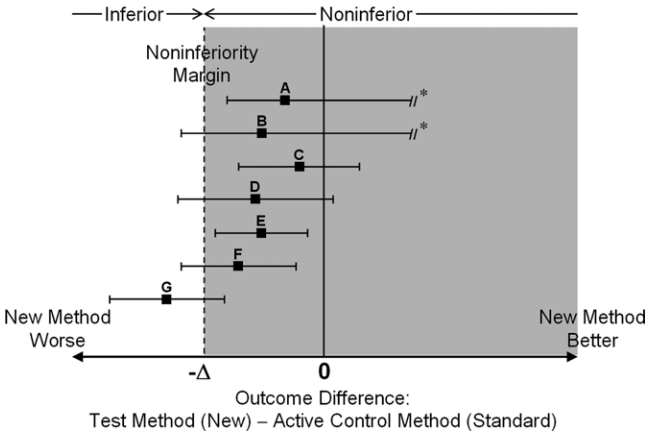

Figure 2: Interpretation of noninferiority of a new method (test) to the standard method (active control) when a greater outcome value indicates a better outcome. Shaded area = noninferiority range (above $-\Delta$). ■ = observed point estimate of outcome difference in each sample. Error bars of *A* and *B* (*) = one-sided CIs (cap at left = lower bound). All other error bars (*C*–*G*) = two-sided 95% CIs (caps at ends = lower and upper bounds of CI). *A*: Test method is noninferior to active control method because entire CI is above noninferiority margin ($-\Delta$). *B*: Test method is not noninferior to active control method because part of CI lies below noninferiority margin. *C*: Test method is not significantly different from active control method because two-sided 95% CI crosses the 0 outcome difference (same as $P \geq .05$ from conventional statistical tests) and is noninferior to active control method. *D*: Test method is not significantly different from active control method but is not noninferior to active control method. *E*: Test method is significantly different from active control method, as two-sided 95% CI does not cross the 0 outcome difference (same as $P < .05$ from conventional statistical tests) but is noninferior to active control method. This puzzling case is rare, since it requires a very large sample size. It can also result from having too generous a noninferiority margin. *F* and *G*: Test method is significantly different from active control method and is not noninferior to active control method. Although the data in *B*, *D*, and *F* do not prove noninferiority per se, a larger sample with a narrower CI may have shown noninferiority. Therefore, confirming whether the study was adequately powered is necessary in such a case. If the observed point estimate of outcome difference of the sample lies outside the noninferiority range (as for *G*), a lack of noninferiority is clearer.

the other hand, a statistically significant difference does not exclude statistical equivalence, either (D of Fig 1).

For noninferiority testing in which one is interested to know if the sensitivity (an index in which a larger value represents a better outcome) of the new test (P_T) is not worse than the sensitivity of the active control test (P_{AC}) without regard to its superiority to the active control test (H_0: $P_T - P_{AC} \leq -\Delta$ versus H_1: $P_T - P_{AC} > -\Delta$), only the relationship between the lower bound of the CI of $P_T - P_{AC}$ and the noninferiority margin (ie, the lower bound of the equivalence range in this example) matters. Therefore, both one-sided and two-sided CIs of $P_T - P_{AC}$ can be used for the analysis (Fig 2). Noninferiority of P_T to P_{AC} is inferred when the lower bound of either CI is above the noninferiority margin ($-\Delta$); however, if a part of or the entire CI lies below the noninferiority margin, P_T is not noninferior to P_{AC} (ie, no evidence of noninferiority) (Fig 2). In noninferiority testing for an outcome in which a smaller value represents a better result (Fig 3)—for example, one study (13) analyzed the noninferiority of low-dose CT to standard-dose CT in evaluating suspected appendicitis by using the negative appendectomy rate (the rate of absence of appendicitis out of all appendectomies performed for a suspicion of appendicitis) as the outcome measure, where a lower rate indicates a better diagnosis (H_0: $P_T - P_{AC} \geq \Delta$ and H_1: $P_T - P_{AC} < \Delta$)—noninferiority of P_T to P_{AC} is inferred when the upper bound of either one-sided CI

Figure 3

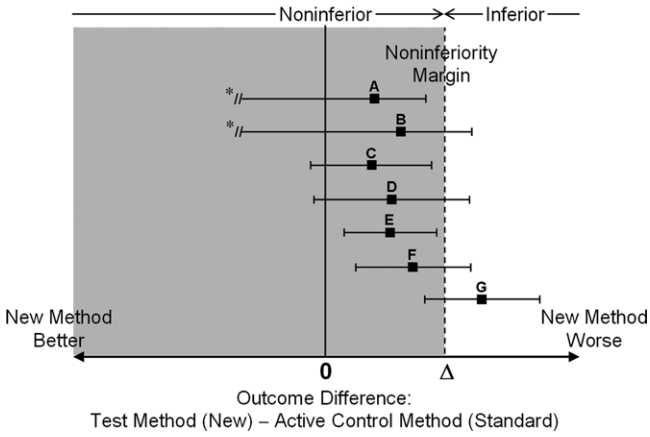

Figure 3: Interpretation of noninferiority of a new method (test) to the standard method (active control) when a smaller outcome value indicates a better outcome. Shaded area = noninferiority range (below Δ). ■ = observed point estimate of outcome difference in each sample. Error bars of *A* and *B* (*) = one-sided CIs (cap at right = upper bound). All other error bars (*C*–*G*) = two-sided CIs (caps at both ends = lower and upper bounds of CI). *A*, *C*, and *E*: Test method is noninferior to active control method because entire CI is below noninferiority margin (Δ). *B*, *D*, *F*, and *G*: Test method is not noninferior to active control method because part of CI lies above noninferiority margin. Although the data in *B*, *D*, and *F* do not prove noninferiority per se, a larger sample with a narrower CI may have shown noninferiority. Therefore, confirming whether the study was adequately powered is necessary in such a case. If the observed point estimate of outcome difference of the sample lies outside the noninferiority range (as for *G*), a lack of noninferiority is clearer.

Table 1

Type I and II Errors of Statistical Testing

Error	Description
Superiority testing	
Type I error	An error to make a false claim of significant difference when there is not a difference
Type II error	An error not to make a claim of significant difference when there is a difference
Equivalence/noninferiority testing	
Type I error	An error to make a false claim of noninferiority and equivalence when the outcomes are actually inferior and not equivalent, respectively.
Type II error	An error not to make a claim of noninferiority or equivalence when the outcomes are actually noninferior and equivalent, respectively.

or two-sided CI of $P_T - P_{AC}$ falls below the noninferiority margin (Δ). Similar to equivalence testing, when the observed point estimate of outcome difference in the sample is within the noninferiority range and the CI lies across the noninferiority margin (as shown in B, D, and F of Figs 2 and 3), confirmation of whether the study was adequately powered is necessary. If the observed outcome difference of the sample lies outside the noninferiority range, a lack of noninferiority is clearer (as shown in G of Figs 2 and 3). As shown in Figures 2 and 3, a lack of statistically significant difference (D of Figs 2 and 3) does not guarantee noninferiority and a statistically significant difference does not exclude noninferiority, either (E of Figs 2 and 3).

Although both one-sided and two-sided CIs work essentially the same for inferring noninferiority, a two-sided CI (most commonly a two-sided 95% CI of which either one side limit, depending on the nature of the outcome, is only used for noninferiority testing) may be preferred as it often provides more insight about the overall data interpretation (4). When a one-sided CI is alternatively used, a conservative one-sided 97.5% CI with one-sided 2.5% type I error (which is the same as either the lower limit or the upper limit of a two-sided 95% CI) is generally preferred to a one-sided 95% CI with one-sided 5% type I error for the analysis, although noninferiority studies are primarily interested in either one side of the equivalence region (4). If a one-sided 5% type I error is considered appropriate, a one-sided 95% CI or a two-sided 90% CI could be used for noninferiority analysis (4,14,15). The meanings of type I and II errors in equivalence/noninferiority testing compared with those in superiority testing are summarized in Table 1.

Equivalence/noninferiority analysis of quantitative or continuous variable outcomes can be likewise as explained previously except for the use of CI of the mean outcome difference instead of the difference in outcome proportions between compared groups. If the CI of the mean outcome difference lies within equivalence and noninferiority ranges, equivalence and noninferiority, respectively, of the method under test to the active control is inferred. For example, Hausleiter et al (16) conducted a noninferiority study to determine if the image quality of coronary CT angiography is maintained (ie, not worse) when 100-kVp tube voltage scan is used (test group) compared with when standard 120 kVp is used (active control group). Image quality was determined on a 4-point grading system (scores 1 to 4). The

mean image quality scores ± standard deviation (SD) was 3.30 ± 0.67 and 3.28 ± 0.68 for 100-kVp and 120-kVp groups, respectively. Therefore, the mean image quality score difference between the two groups (100 kVp − 120 kVp) was 0.02; and the two-sided 95% CI and one-sided 97.5% CI of the mean difference was −0.11 to 0.15 and greater than or equal to −0.11, respectively. As the lower limit of either CI (−0.11) was above the noninferiority margin of −0.2, noninferiority of the quality of 100-kVp images to that of 120-kVp images was concluded.

Equivalence/noninferiority testing of quantitative or continuous variable outcomes is appropriate when the average outcome of a group is the main concern because it makes an inference about the mean outcome difference between compared methods for a group of subjects. However, equivalence/noninferiority testing is inappropriate for analyzing the agreement between a pair of continuous outcome measurements such as an analysis of intermethod or interobserver agreements (17,18). Complete inclusion of the CI of the mean measurement difference within a predefined equivalence range will exclude any substantial systematic over- or undermeasurement by one method/observer compared with the other but does not necessarily prove good agreement (18). The measurement agreement can be more appropriately analyzed by using the Bland-Altman method (19).

Noninferiority Margin: General Principles

As true equivalence clinical studies are rare, the following explanations will focus on noninferiority margin unless specified otherwise. As mentioned previously, noninferiority testing is not to prove the equality of the outcome between compared groups but to prove whether the outcome of the test group is not worse enough compared with the active control group to be clinically or scientifically relevant. The noninferiority margin, thus, defines the clinically or scientifically acceptable range of differences and is a crucial component of noninferiority testing. The margin should be justified on both clinical and statistical grounds (8–11). It is important to avoid a margin that is too generous because such a margin would make a false claim of noninferiority for an inferior outcome, which will adversely affect patients. No single answer exists regarding how to define the margin because a clinically or scientifically acceptable difference would vary according to the particular clinical or scientific issue.

Some guidance regarding determining the noninferiority margin has been proposed for therapeutic drug trials (8–11). To some extent, the logical principles used could be utilized in noninferiority radiology research studies, although their direct application to radiology research studies may be difficult in view of the inherent differences in therapy and diagnosis (this issue will be further discussed in the next section). First, Δ must be smaller than the expected difference in the outcome between the active control and placebo state to ensure that the method being tested has a clinically relevant superiority over placebo. In other words, in terms of therapeutic drugs, a test drug that is noninferior to the standard active control drug has to be at least therapeutically more efficacious than no treatment (ie, natural improvement) as, otherwise, the test drug would be useless or even harmful. The noninferiority margin must be chosen to assure this. As a diagnostic example, suppose a noninferiority comparison between low-dose CT (test method) and standard-dose CT (active control method) for diagnosing acute appendicitis. For low-dose CT to be noninferior in terms of diagnostic accuracy to standard-dose CT, low-dose CT must have at least higher accuracy than does diagnosing appendicitis without CT (diagnostic placebo state). Therefore, Δ needs to be smaller than the difference in diagnostic accuracy between standard-dose CT and diagnosis without CT. When multiple previous studies comparing the active control and placebo are available, a statistical lower bound of the pooled estimate of the outcome differences between the active control and placebo in the previous studies could be used as a conservative measure of the efficacy of the active control method over placebo. Second, the methodologic similarity between the current study and historical data, if referenced in determining the noninferiority margin, such as similarity in patient population characteristics, techniques used, and reference standards needs to be confirmed, which is referred to as the constancy assumption. If any concerns exist regarding the constancy, the estimated outcome difference between the active control and placebo should be adjusted appropriately or, if impractical, the effect of the lack of constancy on the study results should be clearly explained. For example, suppose again that one is designing a noninferiority study to compare low-dose CT (test method) and standard-dose CT (active control method) for diagnosing acute appendicitis. Assume that past data comparing standard-dose CT (active control) and diagnosis without CT (placebo) for diagnosing appendicitis are available for reference in determining the noninferiority margin. However, what if the historical data were obtained in the era of nonspiral CT, whereas the study to be conducted will use 64-dector (or higher) CT scanners? Considering the remarkable advances in CT technology over time but presumably unchanged clinical diagnosis, the accuracy difference between standard-dose CT and no CT would now be greater than it was in the historical data. This lack of constancy should be considered in determining the noninferiority margin. Next, a judgment is made concerning how much of the active control–to-placebo outcome difference should be preserved—that is, Δ—which is dependent on the trade-off between the consequences of the loss in the main outcome of interest (in the previous example, increased unnecessary surgery due to false-positive diagnosis or complications due to delayed diagnosis) and the compensatory gain such as safety, less invasiveness, better tolerability, greater availability, cost or time saving, convenience, ease of performance, or better patient compliance (in the previous example, benefits from less radiation exposure to the population).

When the primary outcome does not involve a serious irreversible result such as death and the compensatory gain is substantial, Δ could be more flexible. Δ is sometimes chosen as a fraction of the estimated outcome difference between the active control and placebo (7), for example, less than 50% of the difference; however, this approach may not always be appropriate (8–10).

Despite the theoretical principles, ambiguity still exists regarding how specifically the principles can be applied to actual research studies, and it is sometimes impractical to strictly follow the guidance. In fact, most equivalence/noninferiority therapeutic studies published thus far did not exactly conform to the principles (7,20,21). According to several systematic reviews (7,21), only 42.7% and 45.7%, respectively, of the published equivalence/noninferiority therapeutic studies provided a justification for their equivalence/noninferiority margin, of which only 16%–19% calculated the margins on the basis of prior active control–to-placebo comparison studies and approximately one-half merely stated that the margin was based on clinical judgment without further details. Considering the extreme diversity of clinical and scientific issues in research studies, some degree of divergence from the principles may be natural. Even if so, research studies should try to provide clear explanations of the rationale for Δ.

Noninferiority Margin: Radiologic Perspective

Some complexities and impracticalities exist in applying the aforementioned principles to radiology research studies, particularly to diagnostic performance studies. First, placebo-controlled studies (ie, diagnostic performance or benefit with versus without a radiologic examination) have been rarely performed because radiologic examinations have customarily been assumed to be diagnostically beneficial. Second, it is not always straightforward to define the placebo state of a diagnosis. Assuming a diagnosis by pure chance (50%-50% chance, like coin flipping) to be the placebo state is likely inappropriate in many cases since medical diagnosis is not solely dependent on radiologic diagnosis and clinical diagnostic capability without radiologic diagnosis is likely greater than diagnosis by pure chance. Third, even if any historical active control–to-placebo comparison exists, a constancy assumption may be difficult to make on some occasions owing to rapid development of imaging technologies. The active control of a current study may include different (generally more advanced) techniques compared with those of historical studies—for example, multidetector CT in the current study versus single detector CT in historical data. Fourth, unlike therapeutic trials that directly investigate ultimate patient outcomes, because diagnostic studies generally evaluate intermediate outcomes such as diagnostic performance or technical quality of a diagnostic image, it is more difficult to estimate how much compromise in these intermediate metrics will or will not cause a clinically relevant negative impact on the ultimate patient outcomes. Maybe, partly related to the fourth factor, many noninferiority diagnostic radiology studies published thus far have used ultimate patient outcome associated with the diagnostic procedures rather than diagnostic accuracy as the primary study endpoint (Table 2). Table 2 summarizes select published noninferiority radiology research studies and shows how Δ was defined. Δ was chosen in various ways in published radiologic noninferiority studies, and similar to the trend in therapeutic studies (7,21), many studies simply stated that Δ was chosen according to "clinical judgment" (15,16,22) or did not clearly explain the rationale for Δ (14,23,24). Additionally, even if a reference to historical data were made in determining Δ, studies did not exactly follow the formal guidance (13,25,26).

Calculation of Sample Size and Confidence Interval

Sample size calculation for equivalence/noninferiority studies depends on general factors similar to the sample size estimation for superiority studies (27), such as (estimated) measurement variability, type I error (significance criterion), statistical power (equivalent to 1 − type II error), and paired (eg, the same group of subjects examined with both test A and test B) versus parallel (eg, two separate groups of subjects, examined with test A or test B) design, and additionally Δ. The meanings of type I and II errors in equivalence/noninferiority testing compared with those in superiority testing are summarized in Table 1. A generous noninferiority margin will require a smaller sample size, whereas a conservative noninferiority margin will require a larger sample. The following will explain how to calculate the sample size and CI for equivalence/noninferiority studies involving several types of data that are frequently observed in radiology research studies. For simplicity of the explanations, a study design that intends to allocate equal numbers of subjects to two compared groups was assumed. Those readers who are interested in other study designs or analysis of other data formats should refer to other articles or books more dedicated to the topic (28).

Comparison of Binary Outcome between Two Groups in the Parallel Design

The total sample size *(N)*—including both the test and active control groups each with an equal number of subjects—for an equivalence/noninferiority study to assess a binary outcome can be estimated as follows:

$$N = 4\frac{(Z_{crit} + Z_{pwr})^2 P(1-P)}{\Delta^2},$$

where Z_{crit} denotes the absolute value of Z-score for the significance criterion, Z_{pwr} is the absolute value of Z-score for the statistical power, and P is the prestudy estimate of the outcome proportion assuming that the outcome proportions of the two groups are the same. Two-tailed Z_{crit} and Z_{pwr} values are used for calculating the sample size for an equivalence study, whereas one-tailed Z_{crit} and Z_{pwr} values are used in the sample size estimation for a noninferiority

Table 2

Examples of Study Design and Δ in Published Noninferiority Radiology Research Studies

Author, Year, and Citation	Study Design	Test vs Active Control	Primary Outcome	Data Type of Primary Outcome	Δ and Criteria to Determine Noninferiority of Test to Active Control*
af Geijerstam et al, 2006 (14)	Parallel	Rapid patient triage using immediate posttrauma head CT vs observation after hospitalization for mild head injury patients	Full vs incomplete recovery 3 months after head injury	Binary	One-sided 95% CI of the difference in incomplete recovery rate < 5% (Δ)
Schaefer et al, 2006 (23)	Parallel	MR angiography with gadodiamide vs gadopentetate for evaluating arterial stenosis	Diagnostic accuracy	Binary	Two-sided 95% CI of the accuracy difference greater than −15% (−Δ)
Anderson et al, 2007 (25)	Parallel	CT pulmonary angiography versus ventilation-perfusion scanning to exclude pulmonary embolism	Development of symptomatic pulmonary embolism or deep vein thrombosis during 3-month follow-up after initial negative test	Binary	Two-sided 95% CI of the difference in the event rate < 2.5% (Δ)
Vincken et al, 2007 (24)	Parallel	Conservative vs arthroscopic treatment in patients with nonacute knee symptoms without MR imaging abnormalities	Effective treatment at 6-month follow-up	Binary	Two-sided 95% CI of the difference in the effective treatment rate greater than −15% (−Δ)
Righini et al, 2008 (26)	Parallel	D-dimer test and CT without vs with lower extremity US to exclude pulmonary embolism	Development of venous thromboembolic events during 3-month follow-up after initial negative test	Binary	Two-sided 95% CI of the difference in the event rate < 1.5% (Δ)
Kim et al, 2012 (13)	Parallel	Low-dose CT vs standard-dose CT for evaluating suspected appendicitis	Rate of uninflamed appendix out of all intended appendectomies (negative appendectomy rate)	Binary	Two-sided 95% CI of the difference in the negative appendectomy rate < 5.5% (Δ)
Hausleiter et al, 2010 (16)	Parallel	Low-dose coronary CT angiography vs standard-dose coronary CT angiography	Semiquantitative score of CT angiography image quality	Continuous	One-sided 97.5% CI of the mean image quality score difference greater than −0.2 score points (−Δ)

* Difference represents outcome of the test method minus outcome of the active control method.

study (Tables 3 and 4), which also applies likewise to other types of data to be discussed later. Therefore, if a noninferiority study aims to achieve a 2.5% one-sided type I error and 90% statistical power, for which Z_{crit} and Z_{pwr} are 1.960 and 1.282, respectively, the equation is reduced to:

$$N \approx \frac{42P(1-P)}{\Delta^2}.$$

After completing the data collection, the two-sided 95% CI for the outcome difference between the two groups to be used for statistical testing is calculated from the study data as follows:

$$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

where p_1 and p_2 represent observed proportions in the test and active control groups, respectively, and n_1 and n_2 are the actual sample sizes of the respective groups.

Table 3

Z_{crit} Values for Different Significance Criteria (ie, Type I Error Levels)

Significance Criteria	Z_{crit}
Noninferiority study	
One-sided 2.5%	1.960
One-sided 5%	1.645
Equivalence study	
Two-sided 5%	1.960
Two-sided 10%	1.645

Note.—Z_{crit} = the absolute value of Z-score for the significance criterion.

Table 4

Z_{pwr} Values for Different Levels of Statistical Power

	Z_{pwr}	
Statistical Power	Noninferiority Study	Equivalence Study
80%	0.842	1.282
90%	1.282	1.645
95%	1.645	1.960

Note.—Z_{pwr} = the absolute value of Z-score for the statistical power.

Suppose, as an example, a hypothetical randomized controlled trial exists to determine noninferiority of the sensitivity of low-dose CT (test) to the sensitivity of standard-dose CT (active control) for diagnosing hepatic tumors

where enrolled patients are to be randomly allocated to either group at a 1:1 ratio. If the expected sensitivity of standard-dose CT for hepatic tumors is 90% and Δ is chosen as 10% (ie, noninferiority of low-dose CT will be inferred if the sensitivity of low-dose CT is not lower than that of standard-dose CT by greater than 10%), a total sample size of 378 patients (ie, 189 patients for each group) is required to achieve 90% power and a 2.5% one-sided type I error: $378 = 42 \times 0.9 \times (1 - 0.9) \div 0.1^2$. As the sample size is inversely proportional to the square of Δ, increase in Δ would rapidly decrease the required sample size (eg, a total sample size of 96 patients for Δ of 20%); however, a Δ that is too large would not be clinically acceptable. If the observed sensitivities after conducting the randomized trial were 85% (161 of 189) for the low-dose CT and 91% (173 of 190) for the standard-dose CT, the two-sided 95% CI for the difference between the two sensitivities is calculated as −12.5% to 0.5% ($0.85 - 0.91 \pm 1.96 \times [0.85 \times 0.15 \div 189 + 0.91 \times 0.09 \div 190]^{1/2}$). Because the lower limit of the 95% CI is below the noninferiority margin (−10%), the study failed to prove the noninferiority of low-dose CT to standard-dose CT regarding the sensitivity.

Comparison of Binary Outcome between Two Groups in the Paired Design

A paired design to assess a binary outcome is the most commonly used study design for diagnostic accuracy studies—that is, comparison of sensitivity, specificity, or accuracy of different diagnostic procedures after having all study subjects undergo all the compared diagnostic tests. Unfortunately, however, statistical procedures for analyzing equivalence/noninferiority between paired binary outcomes have yet to be popularized. Several statistical methods (29–32) have been developed based on the McNemar test and were adopted in clinical studies (22,33,34). Of those, the method by Liu et al (32) is well accepted for estimating the sample size and for statistical testing. As the mathematical details are beyond the scope of this review, we avoid further mathematical discussion in this review and recommend referring to the original articles (29–32) for the details of sample size estimation and statistical testing.

Comparison of Continuous Outcome between Two Groups in Parallel and Paired Designs

The total sample size *(N)*, including both the test group and separate active control group each with an equal number of subjects, for a study to assess a continuous outcome can be estimated as follows:

$$N = 4\frac{(Z_{crit} + Z_{pwr})^2\sigma^2}{\Delta^2},$$

where σ is the prestudy estimate of the SD of outcome values in the population (the larger SD can be used as a conservative approach when the SDs of the two groups are expected to be different). Two population means are assumed to be the same for the sample size calculation. For a noninferiority study with a 2.5% one-sided type I error and 90% statistical power, the equation is reduced to:

$$N \approx \frac{42\sigma^2}{\Delta^2}.$$

After conducting the study, the two-sided 95% CI of the mean outcome difference between the two parallel groups, which will be used for statistical testing, is calculated from the study data as follows:

$$m_1 - m_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where m_1 and m_2 represent observed means in the test and active control groups, respectively, s_1 and s_2 are observed SDs in the respective groups, and n_1 and n_2 are the actual sample sizes of the respective groups.

As an example, Hausleiter et al (16) conducted a randomized study of a parallel design to determine the noninferiority of the image quality of coronary CT angiography performed with a low radiation dose (100 kVp) to the image quality of the standard-dose (120 kVp) examination. The image quality was assessed by using a semiquantitative four-point visual grading system, and the expected SD of the image quality scores was 0.65 in both groups. Δ was set at 0.2 (ie, noninferiority of low-dose CT angiography will be inferred if its mean image quality score is not lower by greater than 0.2 compared with the mean image quality score of standard-dose CT angiography). For a one-sided type I error of 2.5% and 80% statistical power, for which Z_{crit} and Z_{pwr} are 1.960 and 0.842, respectively (Tables 3 and 4), 332 patients were needed for the study: $332 = 4 \times (1.960 + 0.842)^2 \times 0.65^2 \div 0.2^2$. The study showed that the mean ± SD of the image quality scores was 3.30 ± 0.67 for 202 patients examined with low-dose CT angiography and 3.28 ± 0.68 for 198 patients examined with standard-dose CT angiography. The two-sided 95% CI was calculated to be −0.11 to 0.15 ($3.30 - 3.28 \pm 1.96 \times [0.67^2 \div 202 + 0.68^2 \div 198]^{1/2}$). As the lower limit of the two-sided 95% CI (or equivalently, the lower limit of the one-sided 97.5% CI) was greater than the predefined noninferiority margin of −0.2, noninferiority of low-dose CT angiography to standard-dose CT angiography was concluded.

For a paired-design study in which a pair of continuous outcome data are obtained from the same population, the estimated sample size for the study *(N)* can be calculated as follows:

$$N = 2\frac{(Z_{crit} + Z_{pwr})^2\sigma_d^2}{\Delta^2},$$

where σ_d is the prestudy estimate of the SD of the outcome differences between the two comparing methods in individual subjects of the population. For a noninferiority study with a 2.5% one-sided type I error and 90% statistical power, the equation becomes:

$$N \approx \frac{21\sigma_d^2}{\Delta^2}.$$

The two-sided 95% CI for the mean difference between the two paired continuous outcomes to be used for statistical testing is calculated from the study data as follows:

$$m_1 - m_2 \pm 1.96\sqrt{\frac{s^2}{n}},$$

where m_1 and m_2 represent the observed mean outcome in the test and active control groups, respectively, s is the observed SD of the outcome differences between the two comparing methods in the study subjects, and n is the sample size.

Comparison of Receiving Operator Characteristic Curves

Receiver operating characteristic (ROC) analysis is widely used in diagnostic radiology research studies in addition to the analysis of sensitivity, specificity, and/or accuracy since radiologic interpretation is oftentimes not truly binary (ie, presence versus absence of a disease) but is better described by (semi-)quantitative continuum of probabilities of the presence of a disease state (35,36). Statistical methods for equivalence/noninferiority analysis of the area under the ROC curve as the primary endpoint are available (37–39). However, these methods are not widely used yet, similar to the rather arcane statistical methods for equivalence/noninferiority analysis of paired binary data. The mathematical details will not be discussed in the review, for which we recommend referring to the original literature (37–39). As equivalence/noninferiority studies are expected to become more prevalent in radiology research in the future, the relevant statistical methods for analyzing ROC data as well as paired binary data would be more recognized in the near future.

Conclusion

The need for research studies to demonstrate the similarity of a radiologic test or treatment, which has ancillary advantages, to traditional standard methods regarding their main effect is expected to increase. Such studies should be appropriately designed and analyzed using equivalence/noninferiority study methodology as explained in this review. A good understanding of the fundamental principles of equivalence/noninferiority statistical testing will be helpful for conducting as well as for interpreting such research studies.

Disclosures of Conflicts of Interest: S.A. Author stated no relevant conflicts of interest to disclose. **S.H.P.** Author stated no relevant conflicts to disclose. **K.H.L.** Financial activities related to the present article: grant to institution from Ministry of Health and Welfare, Republic of Korea. Financial activities not related to the present article: grant to institution form GE Healthcare Medical Diagnostics, National Research Foundation of Korea, and Ministry of Health and Welfare. Other relationships: none to disclose.

References

1. Costa LJ, Xavier AC, del Giglio A. Negative results in cancer clinical trials: equivalence or poor accrual? Control Clin Trials 2004;25(5):525–533.
2. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? Arch Surg 2001;136(7):796–800.
3. Krysan DJ, Kemper AR. Claims of equivalence in randomized controlled trials of the treatment of bacterial meningitis in children. Pediatr Infect Dis J 2002;21(8):753–758.
4. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295(10):1152–1160.
5. Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? Ann Intern Med 2000;132(9):715–722.
6. Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. ACP J Club 2004;140(2):A11.
7. Lange S, Freitag G. Choice of delta: requirements and reality—results of a systematic review. Biom J 2005;47(1):12–27; discussion 99–107.
8. ICH Steering Committee. ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials E10. ICH Web site. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf. Accessed March 14, 2012.
9. ICH Steering Committee. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9. ICH Web site. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf. Accessed March 14, 2012.
10. EMEA. CHMP Guideline on the choice of the non-inferiority margin. European Medicines Agency Web site. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf. Accessed March 14, 2012.
11. FDA. Guidance for industry: non-inferiority clinical trials. FDA Web site. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Accessed March 14, 2012.
12. Welleck S. Testing statistical hypotheses of equivalence and noninferiority. 2nd ed. Boca Raton, Fla: Chapman & Hall/CRC, 2010; 1–368.
13. Kim K, Kim YH, Kim SY, et al. Low-dose abdominal CT for evaluating suspected appendicitis. N Engl J Med 2012;366(17):1596–1605.
14. af Geijerstam JL, Oredsson S, Britton M; OCTOPUS Study Investigators. Medical outcome after immediate computed tomography or admission for observation in patients with mild head injury: randomised controlled trial. BMJ 2006;333(7566):465.
15. Lee SJ, Park SH, Kim AY, et al. A prospective comparison of standard-dose CT enterography and 50% reduced-dose CT enterography with and without noise reduction for evaluating Crohn disease. AJR Am J Roentgenol 2011;197(1):50–57.
16. Hausleiter J, Martinoff S, Hadamitzky M, et al. Image quality and radiation exposure with a low tube voltage protocol for coronary CT angiography results of the PROTECTION II Trial. JACC Cardiovasc Imaging 2010;3(11):1113–1123.
17. MERCURY Study Group. Extramural depth of tumor invasion at thin-section MR in patients with rectal cancer: results of the MERCURY study. Radiology 2007;243(1):132–139.
18. Park SH. Degree of error of thin-section MR in measuring extramural depth of tumor invasion in patients with rectal cancer. Radiology 2008;246(2):647; author reply 647–648.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307–310.
20. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA 2006;295(10):1147–1151.
21. Wangge G, Klungel OH, Roes KC, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting noninferiority randomized controlled trials on drugs: a systematic review. PLoS ONE 2010;5(10):e13550.
22. Kerrou K, Pitre S, Coutant C, et al. The usefulness of a preoperative compact im-

ager, a hand-held gamma-camera for breast cancer sentinel node biopsy: final results of a prospective double-blind, clinical study. J Nucl Med 2011;52(9):1346–1353.

23. Schaefer PJ, Boudghene FP, Brambs HJ, et al. Abdominal and iliac arterial stenoses: comparative double-blinded randomized study of diagnostic accuracy of 3D MR angiography with gadodiamide or gadopentetate dimeglumine. Radiology 2006;238(3):827–840.

24. Vincken PWJ, ter Braak APM, van Erkel AR, et al. MR imaging: effectiveness and costs at triage of patients with nonacute knee symptoms. Radiology 2007;242(1):85–93.

25. Anderson DR, Kahn SR, Rodger MA, et al. Computed tomographic pulmonary angiography vs ventilation-perfusion lung scanning in patients with suspected pulmonary embolism: a randomized controlled trial. JAMA 2007;298(23):2743–2753.

26. Righini M, Le Gal G, Aujesky D, et al. Diagnosis of pulmonary embolism by multidetector CT alone or combined with venous ultrasonography of the leg: a randomised non-inferiority trial. Lancet 2008;371(9621):1343–1352.

27. Eng J. Sample size estimation: how many individuals should be studied? Radiology 2003;227(2):309–313.

28. Julious SA. Sample sizes for clinical trials. Boca Raton, Fla: Chapman & Hall/CRC, 2009; 236–242.

29. Lu Y, Bean JA. On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. Stat Med 1995;14(16):1831–1839.

30. Nam JM. Establishing equivalence of two treatments and sample size requirements in matched-pairs design. Biometrics 1997;53(4):1422–1430.

31. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat Med 1998;17(8):891–908.

32. Liu JP, Hsueh HM, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired binary data. Stat Med 2002;21(2):231–245.

33. Durkalski VL, Palesch YY, Pineau BC, Vining DJ, Cotton PB. The virtual colonoscopy study: a large multicenter clinical trial designed to compare two diagnostic screening procedures. Control Clin Trials 2002;23(5):570–583.

34. Itoh A, Ueno E, Tohno E, et al. Breast disease: clinical application of US elastography for diagnosis. Radiology 2006; 239(2):341–350.

35. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229(1):3–8.

36. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol 2004;5(1):11–18.

37. Jin H, Lu Y. A non-inferiority test of areas under two parametric ROC curves. Contemp Clin Trials 2009;30(4):375–379.

38. Liu JP, Ma MC, Wu CY, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. Stat Med 2006;25(7):1219–1238.

39. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley, 2002; 165–221.