Testing for Interchangeability of Imaging Tests

Nancy A. Obuchowski, PhD, Naveen Subhas, MD, Paul Schoenhagen, MD

Rationale and Objectives: New tests are typically assessed by estimating their technical and diagnostic performance through comparisons with a reference standard. A valid reference standard, however, is not always available and is not required for assessing the interchangeability of a new test with an existing one.

Materials and Methods: To show interchangeability of a new test with an existing test, one compares the differences in diagnoses between the new and existing tests to differences between diagnoses made with the existing test on several occasions. We illustrate the test for interchangeability with two studies. In a transcatheter aortic valve replacement study, we test whether semiautomated analysis can be used interchangeably with manual reconstructions from three-dimensional computed tomography (CT) images. In patients with femoroacetabular impingement, we test whether magnetic resonance imaging (MRI) can replace CT to measure acetabular version.

Results: Although the semiautomated method agreed often with the manual measurement of aortic valve size (87.6%), interchanging the semiautomated method with manual measurements by an expert would lead to a 1.7%–12.2% increase in the frequency of disagreement. Interchanging MRI for CT to measure acetabular version would lead to differences in angle measurements of 2.0° to 3.1° in excess of the differences we would expect to see with CT alone.

Conclusions: Testing for agreement or correlation between a new and an existing test is not sufficient evidence of the performance of a new test. A formal evaluation of interchangeability can be conducted in the absence of a reference standard.

Key Words: Reference standard; accuracy; interchangeability; agreement; correlation.

As new medical imaging tests and procedures are developed, it is imperative that their performance be assessed before the new test can be used in clinical practice. The intended use of the new test affects how it is assessed. If the new test is to be used as an adjunct to the existing test, then its performance in combination with the existing test should be superior to the performance of the existing test alone. If the new test is to replace the existing test, then its performance as a stand alone should not be inferior to the existing test. Another role for new tests is interchangeability with an existing test. Here, we assess whether the new test can be switched with the existing test without affecting individual patients' diagnoses.

For adjunct and replacement roles, new tests are assessed by estimating their technical and diagnostic performance through comparisons with a gold or reference standard (eg, sensitivity, specificity, and receiver operating characteristic (ROC) analysis) (1–3). We compare the performance (eg area under the ROC curve) of the new and existing tests for the relevant *population of patients*. A valid reference standard, however, is not always available and is not required for assessing interchangeability. When interchanging a new test with an existing test, we must have sufficient evidence that for *individual patients*, either of the tests can be used with similar results. Replacement is a less burdensome criterion than interchangeability because the test results of individual patients do not need to agree, only the performance over the population of patients.

Consider a study of patients undergoing transcatheter aortic valve replacement (TAVR) (4). Three-dimensional computed tomography (CT) imaging is the standard approach for preprocedural selection and matching of device size to the size of the annulus. Manual reconstructions are time consuming with the potential for inter- and intra-observer variability. Semiautomated analysis is a new, faster, and potentially more reproducible approach. Investigators are interested in whether the semiautomated method can be used interchangeably with the manual measurements.

Consider a second study of patients with femoroacetabular impingement. Analysis of acetabular version is an essential part of preoperative planning in these patients, and currently CT is the best preoperative tool for its measurement. Most of these patients, however, also have an magnetic resonance imaging (MRI) examination which can be used to measure acetabular version potentially obviating the need for a second imaging study and exposure to radiation. In order for MRI to replace CT, its performance must be evaluated relative to CT, yet CT

Acad Radiol 2014; 21:1483–1489

From the Department of Quantitative Health Sciences, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195 (N.A.O.); Departments of Diagnostic Radiology (N.A.O., N.S., P.S.) and Cardiovascular Medicine (P.S.), Cleveland Clinic Foundation, Cleveland, OH. Received April 30, 2014; accepted July 7, 2014. **Address correspondence to:** N.A.O. e-mail: obuchon@ccf.org

http://dx.doi.org/10.1016/j.acra.2014.07.004

is not a true gold standard. Without a valid reference standard, investigators want to test if MRI is interchangeable with CT and thus argue that it can replace CT.

In this article, we describe a method to evaluate whether a new imaging test is interchangeable with an existing test by taking into consideration the measurement errors of the new and existing tests. We illustrate the test for interchangeability with data from the TAVR and acetabular version studies. We contrast the results with simpler, naive methods, that is, testing for correlation and measuring agreement.

METHODS

Test for Interchangeability

The idea of interchangeability of imaging tests is similar to the concept of switchability, or bioequivalence, between a test drug and a reference drug (5,6). When drugs are interchangeable, we expect them to produce the same clinical result in any given patient. To show interchangeability of a new drug with a reference drug, one compares the differences in bioavailability between the test and reference drugs to differences between two responses with the reference drug. The concept is used by the Food and Drug Administration for establishing bioequivalence (7–9).

Obuchowski (10) and Barnhart et al. (11) applied these ideas to the setting of testing the equivalence of diagnostic tests and devices. Let Y_{iTjk} denote the result or measurement with the new test (T) modality by reader j for subject i on occasion k, and let Y_{iRjk} denote the result or measurement with the existing (reference [R]) modality by reader j for subject i on occasion k. The null and alternative hypotheses for a test of interchangeability are as follows:

$$H_0:\ \gamma = E\left(Y_{iTjk} - Y_{iRjk'}\right)^2 - E\left(Y_{iRjk} - Y_{iRjk'}\right)^2 > \theta_I \text{ versus } H_1: \gamma \le \theta_I \quad (1)$$

where γ is the individual equivalence index and θ_I is the equivalence limit. θ_I would be defined during the planning phase of a study. It can be shown that the individual equivalence criterion in Equation (1) is equivalent to the squared difference in the means of the two modalities plus the difference in their variances:

$$\gamma = (\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2, \quad (2)$$

where μ_T and σ_T^2 are the mean and variance of the measurements on the new test modality, respectively, and μ_R and σ_R^2 are the mean and variance of the measurements on the reference modality, respectively. Note that if the new test modality has little bias (ie, $\mu_T \cong \mu_R$), it can have a little less precision ($\sigma_T^2 > \sigma_R^2$) and still be considered interchangeable with the reference test, or if the new test modality is more precise than the reference test ($\sigma_T^2 < \sigma_R^2$), it can have a little bias (ie, $\mu_T \neq \mu_R$) and still be considered interchangeable.

In many diagnostic device settings, there is no well-accepted definition for how close two measurements must be in order for the two measurements to be considered equivalent. For example, in the acetabular version study, it is unclear what difference in angle measurements between the CT and MR, or between two readers' measurements on CT, is acceptable. So, we would estimate γ in Equation (1) and construct a confidence interval (CI) for γ. The estimate of γ would provide an estimate of the difference in angle measurements between MR and CT that is *in excess of* the difference in angle measurements obtained with just CT.

In other diagnostic settings, there is a clear definition of clinical agreement. The test of interchangeability then becomes a comparison of the frequency with which patients receive a similar diagnosis with the two modalities, relative to the frequency with which they receive a similar diagnosis on two occasions with the reference test. For example, in the TAVR study, agreement occurs when two measurements of annular area lead to the same valve size. For these studies, we use a different form of the individual equivalence index, referred to as the probability criterion, as follows:

$$H_0:\ \gamma_{(p)} = Prob\left(Y_{iRjk} = Y_{iRjk'}\right) - Prob\left(Y_{iTjk} = Y_{iRjk}\right) > \theta_{I(p)} \text{ versus } H_1: \gamma_{(p)} \le \theta_{I(p)} \quad (3)$$

where $Prob(Y_{iRjk} = Y_{iRjk'})$ is the probability that the result (eg, valve size) determined at two occasions (eg, two readers j and j') with the reference test agree for subject i, and $Prob(Y_{iTjk} = Y_{iRjk})$ is the probability that the results of the new and reference tests agree for subject i.

Estimation of γ and $\gamma_{(p)}$ depends on the specific interchangeability question asked in the study and by the study design. We now consider each of our two clinical studies.

TAVR Study

In a study by Lou et al. (4), 110 patients with severe aortic stenosis, who were being evaluated at our center for TAVR, were identified from an imaging database as having a dedicated TAVR contrast-enhanced CT angiography protocol of the aortic root in the preprocedural evaluation. Data collection was approved by the institutional review board (IRB), with waiver of individual consent. An experienced investigator (>10 years of experience) performed both manual and semiautomated measurements in a blinded fashion each at two time points. A second less-experienced investigator (about 1 year) performed the semiautomated measurements twice and manual analysis once. The goal of the study was to determine if the new semiautomated software could be used interchangeably with the manual measurements.

Based on both the manual and semiautomated measurements, annulus area categories were classified for selection of one of three valve sizes (23, 26, and 29 valve):