Radiology

Metrology Standards for Quantitative Imaging Biomarkers[1]

Daniel C. Sullivan, MD
Nancy A. Obuchowski, PhD
Larry G. Kessler, ScD
David L. Raunig, PhD
Constantine Gatsonis, PhD
Erich P. Huang, PhD
Marina Kondratovich, PhD
Lisa M. McShane, PhD
Anthony P. Reeves, PhD
Daniel P. Barboriak, MD
Alexander R. Guimaraes, MD, PhD
Richard L. Wahl, MD
For the RSNA-QIBA Metrology Working Group[2]

Although investigators in the imaging community have been active in developing and evaluating quantitative imaging biomarkers (QIBs), the development and implementation of QIBs have been hampered by the inconsistent or incorrect use of terminology or methods for technical performance and statistical concepts. Technical performance is an assessment of how a test performs in reference objects or subjects under controlled conditions. In this article, some of the relevant statistical concepts are reviewed, methods that can be used for evaluating and comparing QIBs are described, and some of the technical performance issues related to imaging biomarkers are discussed. More consistent and correct use of terminology and study design principles will improve clinical research, advance regulatory science, and foster better care for patients who undergo imaging studies.

[1] From the Department of Radiology, Duke University Medical Center, Box 2715, Durham, NC 27710 (D.C.S., D.P.B.); Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, Ohio (N.A.O.); Department of Public Health, University of Washington, Seattle, Wash (L.G.K.); Department of Informatics, ICON Medical, Washington, Pa (D.L.R.); Center for Statistical Sciences, Brown University, Providence, RI (C.G.); National Cancer Institute, Bethesda, Md (E.P.H., L.M.M.); Center for Devices and Radiological Health, U.S. Food and Drug Administration, White Oak, Md (M.K.); Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY (A.P.R.); Department of Radiology, Oregon Health & Science University, Portland, Ore (A.R.G.); and Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (R.L.W.).
Received October 1, 2014; revision requested November 10; revision received June 21, 2015; accepted July 11; final version accepted July 11. **Address correspondence to** D.C.S. (e-mail: daniel.sullivan@duke.edu).

[2] The members of the RSNA-QIBA Metrology Working Group are listed in the Acknowledgments.

In the past 2 decades, there has been a dramatic increase in our knowledge about the molecular basis of disease. The transformative effect of this ever-increasing genotypic and phenotypic molecular information on the management of disease is referred to as *precision medicine*. Precision medicine has recently been identified as a national priority *(http://www.nih.gov/precision-medicine/)*. Tailoring and monitoring therapy to an individual's molecular signature of disease necessitates measuring signals from a variety of anatomic, physiological, and biochemical characteristics of the body. These characteristics are called *biomarkers*. A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or a response to a therapeutic intervention" (1).

Concomitantly, during the past 2 decades, remarkable advances in medical imaging technology have made it possible to obtain many anatomic, functional, metabolic, and physiological measurements from clinical images, all of which reflect in some way the molecular substrate of the healthy or diseased tissue, organ, or person undergoing imaging. With appropriate calibration, most of these imaging technologies can provide quantitative information about some properties of the material from which the imaging signal has emanated. Thus, such imaging methods also constitute biomarker measurement processes and are conceptually similar to laboratory or physiological assays.

Advance in Knowledge

- The development and implementation of quantitative imaging biomarkers (QIBs) have been hampered by the inconsistent or incorrect use of terminology for technical performance and statistical concepts; the Quantitative Imaging Biomarkers Alliance Metrology Working Group developed recommendations on terminology and methods for assessing the technical performance of a QIB.

The term *quantitative imaging* has been defined as "the extraction of quantifiable features from medical images for the assessment of normal [findings] or the severity, degree of change, or status of a disease, injury, or chronic condition relative to normal [findings]" (2). Therefore, by combining the two concepts of biomarkers and quantitative imaging, a quantitative imaging biomarker (QIB) can be defined as an objectively measured characteristic derived from an in vivo image as an indicator of normal biological processes, pathogenic processes, or response to a therapeutic intervention.

Investigators in the imaging community have been active in developing and evaluating imaging biomarkers. A search for the keywords "quantitative AND imaging AND biomarkers" in PubMed produces more than 43 000 results. Most imaging journals include several articles related to QIBs every month. For example, a search conducted by using the same keywords in the journal *Radiology* yields more than 200 primary articles on QIBs since 2010. However, few of these imaging biomarkers have been rigorously evaluated, and even fewer are used routinely in clinical trials or clinical care. This problem is not unique to imaging biomarkers. While appraising the field of biomarkers in general, Poste wrote that "thousands of papers have been written, but too few clinically useful biomarkers have been produced" (3). He argued that the research community must "adopt common standards and a cross-disciplinary, systems-based approach to biomarker discovery and validation."

For imaging biomarkers to play an important role in the future evolution of precision medicine, both technical performance and clinical performance need to be evaluated rigorously. Technical performance is an assessment of how a test performs in reference objects or subjects under controlled conditions. Once technical performance is established for a given biomarker, considerable additional research needs to be performed to determine clinical validation (how it performs in a human population—for instance, clinical sensitivity and specificity) and clinical usefulness (benefit to a subject in terms of accepted clinical or regulatory outcomes). The relationship between an imaging biomarker and a clinical outcome determines which biomarkers are deemed clinically relevant, but we do not address clinical validation issues in this article. Also, when quantitative biomarkers are used in clinical practice, cutoff points or thresholds are generally established for clinical decision making, but in this article we do not address the methods for making those determinations.

To determine the technical validity and clinical usefulness of QIB measurements, it is crucial that the framework in which they are acquired is described rigorously, including context of use,

Implication for Patient Care

- More consistent and correct use of terminology and study design principles will improve clinical research, advance regulatory science, and foster better personalized care for patients who undergo imaging studies.

Published online before print
10.1148/radiol.2015142202 **Content code:** BQ

Radiology 2015; 277:813–825

Abbreviations:
QIB = quantitative imaging biomarker
QIBA = Quantitative Imaging Biomarkers Alliance
SUV = standardized uptake value

Author contributions:
Guarantors of integrity of entire study, D.C.S., D.L.R.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, D.C.S., N.A.O., D.L.R., C.G., E.P.H., L.M.M., A.P.R., D.P.B.; experimental studies, M.K.; statistical analysis, N.A.O., C.G., E.P.H., M.K., L.M.M., A.R.G.; and manuscript editing, D.C.S., N.A.O., L.G.K., D.L.R., C.G., E.P.H., M.K., L.M.M., A.P.R., D.P.B., A.R.G.

Funding:
This research was supported by the National Institutes of Health (grant no. HHSN268201000050C).

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

acquisition parameters, and measurement methods. After that framework is described, then the variability and error according to those settings can be quantified. Knowledge of these factors will enable clinicians to reliably compare measurements over time and across imaging platforms.

The concepts and methods for evaluating and comparing the technical performance characteristics of imaging biomarkers are not well understood by many in the imaging community. Standard terminology and methods have become established in medicine to describe, evaluate, and validate laboratory assays. The same concepts and approaches could and should be applied to imaging assays, and this has begun to occur in an organized way in imaging in the past few years. In this article, we review some of the important statistical concepts relevant to technical performance, describe methods that can be used for evaluating and comparing QIBs, and discuss some of the technical performance issues related to imaging biomarkers. A glossary of recommended definitions for key terms is provided in Table 1. Additional terms, definitions, and discussion on these concepts can be found in the work of Kessler et al (1).

Table 1

Recommended Terminology for Describing the Technical Performance of QIBs

Term	Definition
QIB	A characteristic derived from one or more in vivo images and objectively measured according to a ratio or interval scale as an indicator of normal biological processes, pathogenic processes, or response to a therapeutic intervention.
Measurand	The quantity intended to be measured (VIM clause 2.3).
Bias	An estimate of a systematic measurement error (VIM clause 2.18).
Linearity	The ability to provide measured quantity values that are directly proportional to the value of the measurand in the experimental unit (ISO standard 18113).
Precision	The closeness of agreement between measured quantity values obtained by means of replicate measurements of the same or similar experimental units with specified conditions (VIM clause 2.15). Repeatability and reproducibility are types of precision.
Reference value	A value, generally accepted as having a suitably small measurement uncertainty, to be used as a basis for comparison with values of quantities of the same kind (eg, the mean of a large number of replicate measurements) by using a reference method (VIM clause 5.18).
Repeatability	The measurement precision with conditions that remain unchanged between replicate measurements (repeatability conditions) (VIM clause 2.20).
Repeatability conditions	The set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions, same physical location, and replicate measurements of the same or similar experimental units over a short period of time.
Reproducibility	The measurement precision with conditions that vary between replicate measurements (reproducibility conditions) (VIM clause 2.25).
Reproducibility conditions	The set of conditions that includes *(a)* different locations, operators, and measuring systems and *(b)* replicate measurements of the same or similar objects.
Truth or true value	In metrology, truth is the real or actual value of a quantity associated with some object. Because each physical measurement has some uncertainty in terms of whether it agrees with the real quantity value, the true value cannot be known with certainty.

Note.—ISO = International Organization for Standardization, VIM = International Vocabulary of Metrology.

Quantitative Imaging Biomarkers Alliance

In response to the need for reliable and reproducible quantification of biomedical imaging data, in 2007 the Radiological Society of North America organized the Quantitative Imaging Biomarkers Alliance (QIBA), with the mission of improving the value and practicality of QIBs by reducing variability across devices, patients, and time (2). QIBA participants span a wide range of expertise, including clinical practice, clinical research, physics, statistics, engineering, marketing, regulatory practices, pharmaceuticals, and computer science. With QIBA, a systematic, consensus-driven approach is used to produce a QIBA profile that includes one or more claims and specifications for the image acquisition and processing necessary to achieve that claim. QIBA profiles are based on published data whenever such data are available and on expert consensus opinion for specifications when no data exist.

During the first few years of QIBA committee activity, it became apparent that participants were expressing quantitative concepts inconsistently and/or ambiguously. Therefore, QIBA convened experts to advise on terminology and methods relevant to the concerns listed earlier. The information and recommendations presented in this article are derived from the deliberations and publications of the QIBA Metrology Working Group (1,4–7).

QIB Characteristics

The QIBA Metrology Working Group recommends that to be considered a QIB, the measurand (an underlying quantity of interest) must be a ratio or interval variable, as defined by Stevens (8). A ratio variable is one for which there is a clear definition of zero and for which the ratio of two values can be meaningfully interpreted. In this context, "a clear definition of zero" means that zero indicates that no signal intensity is present from the feature being

measured. For example, tumor volume as measured with computed tomographic (CT) volumetry is a ratio variable because if one tumor has a volume of 0.5 cm^3 and another tumor has a volume of 1.5 cm^3, the following statements based on arithmetic operations have real meaning: *(a)* The larger tumor is 1.0 cm^3 bigger than the smaller tumor, and *(b)* the larger tumor is three times the size of the smaller tumor. Furthermore, a tumor volume of zero means there is no tumor mass.

Measures for which the difference between two values is meaningful but the ratio of two values is not and for which the scales do not have a "meaningful zero" are called *interval variables* (8). Examples of scales in everyday use that do not have "meaningful zeros" are the centigrade and Fahrenheit temperature scales, where a temperature measurement of zero does not mean that the entity has no heat energy. The Kelvin temperature scale, on the other hand, does have a meaningful zero, because absolute zero in that scale means there is no thermal motion present. In the imaging context, the CT Hounsfield scale is one that does not have a meaningful zero. By definition, zero in the Hounsfield scale is the density of water, so substances that measure zero in Hounsfield units do have some density. Therefore, imaging biomarkers such as lung densitometry based on Hounsfield unit measurements are examples of interval variables. Examples of interval variable imaging biomarkers based on CT are those used to estimate the severity of emphysema, such as percentage emphysema index and percentile density (9,10).

Stevens describes two other types of scales: ordered and nominal. Ordered scales are those for which values are assigned a magnitude and for which the ordering of values does have meaning, but neither the difference between two values nor the ratio of two values is meaningful. These are not QIBs. Examples in imaging include the Breast Imaging Reporting and Data System, or BI-RADS, assessment categories 1 (negative) through 5 (highly suggestive of malignancy) and breast composition categories 1 (almost entirely fat) through 4 (extremely dense) (11). At first glance, these might seem to be examples of quantitative biomarkers, but they are not. The order of the category numbers follows the increasing clinical significance, in the sense that BI-RADS category 2 is worse than category 1, category 3 is worse than category 2, and so on, but category 4 is not twice as bad as category 2, for example (ie, arithmetic operations in these category numbers have no real meaning).

With a nominal scale, numbers are arbitrarily assigned to categories, and neither the ordering nor the arithmetic operations on the numbers have real meaning. For example, spiculated nodules might be called category 1, part-solid nodules category 2, nodules with indistinct margins category 3, and so on. The numbers are assigned arbitrarily for convenience, and the order does not necessarily convey clinical or other significance. Although categorizations such as these have clinical usefulness in terms of consistency of communication and the like, they are not QIBs as defined here.

QIBs consist of only a measurement of a measurand or a measurement obtained while other specified or relevant factors are held constant. An example of the former is the volume of a tumor obtained from a CT image. An example of the latter concept is the standardized uptake value (SUV) obtained from positron emission tomography (PET) images. Here, the measurand is tissue radioactivity concentration at some time after injection. The SUV is calculated as the ratio of the value of the measurand to the injected dose at the time of injection, divided by body weight (injected dose and weight being the relevant factors that are held constant).

Truth and Reference Values

In metrology (ie, measurement science), measurements need to be compared in some way to "truth." However, the concept of "truth" in metrology is more philosophical than mathematically specific. Truth can be thought of as the set of "true values" of whatever is being measured. Because all measurements have some inherent error (uncertainty), no single measurement can exactly equal the "true value." In publications in the biomedical literature, authors sometimes use the terms "ground truth" or "gold standard." Neither of these has a standardized or unambiguous definition and should therefore not be used. These terms are sometimes used to imply that truth is known with certainty (in the case of "ground truth") or that a measurement has no error (in the case of "gold standard"). Neither the absence of uncertainty nor the absence of error in quantitative values is accepted in metrology.

The prescribed method for obtaining the set of true values is called a *reference method*, "a methodology that has exact and clear descriptions of the necessary conditions and procedures that provide sufficiently accurate and precise laboratory data for it to be used to assess the validity of other laboratory methods" (International Organization for Standardization standard 20776–2) (12). A reference method may also involve the use of one or more reference materials—that is, materials with sufficiently homogeneous and stable properties for use in assignment of a value to another material. The term *reference standard* can have one of two meanings, depending on the context. In the International Vocabulary of Metrology, it is defined as a physical object, synonymous with a *reference material* (13). However, many standards organizations use the phrase to mean the description of procedures for performing the reference method and using the reference material. The reference method also has random measurement errors, but the mean of replicates (ideally a large number of replicates) is generally accepted as the *reference value*.

True values, or even reference values, are difficult to obtain in humans and are therefore usually not available in the clinical setting. True values may require confirmation of disease measurements, and such measurements are often not reliable or possible even at surgery or autopsy. Obtaining reference values can also be challenging owing to

the constraints or expense of imaging subjects multiple times. Pragmatic alternatives include the use of digital or physical phantoms.

Consider these two examples. In a study by Reeves et al (14), the performance of imaging procedures to measure the size and change in size of pulmonary nodules on CT images was assessed. In one part of the study, measurements were made of a synthetic phantom nodule. The volume of spherical synthetic nodules can be mathematically determined on the basis of the radius (true value), while the volume of irregularly shaped nodules can be determined as the mean of multiple measurements of the amount of displaced water (reference value). In the same study, the authors assessed measured change in nodule size for patients who underwent imaging two times within a short time interval. The authors assumed that the true value of change was zero, since the time interval between measurements was too small for a biological change to occur in tumor volume. In a second study (15), an evaluation of an automated system to measure aortic annular area from three-dimensional CT images involved the use of an expert radiologist's manual measurements from the CT images as the reference standard. It was recognized that the expert radiologist's measurements of the annular area contained measurement error, which was quantified in the study by assessing the radiologist's intrareader variability. The magnitude of the intrareader variability was taken into account in the assessment of the automated system.

The set of measurements must be determined by using a method with a suitably small measurement uncertainty. In other words, the reference values must be highly concordant with the true values. Otherwise, comparisons of the measurement values with the reference values will produce significantly different results from comparisons with the true values. Simulation studies from Obuchowski et al (6) indicate that the intraclass correlation coefficient that reflects the concordance between the reference values and the true values needs to be at least 0.99 for the reference values to be an appropriate substitute for the true values in this setting.

Accuracy is a commonly used term that does not have a single unambiguous definition. In some publications, it is used synonymously with *bias* (16), whereas it is often used in a way that combines both bias and precision. However, there is no single agreed-upon way to combine bias and precision, which is why there are multiple aggregate measures of technical performance. Thus if the term *accuracy* is used to describe a quality of the measurement presented, it should also be accompanied by a more detailed description of the components of uncertainty that include both the central tendency (ie, bias) and data dispersion (ie, precision) it encompasses. This is in contrast to the field of diagnostic imaging, where investigators rely on trained readers to interpret an image; in that field, the term *accuracy* is synonymous with the area under the receiver operating characteristic curve.

Investigators in QIB research studies report a wide array of statistical and descriptive metrics to demonstrate the reliability of QIB for clinical use. However, these diverse metrics are sometimes difficult to interpret, may be inappropriate for the recommended use, occasionally lead to misinterpretation of accepted statistical definitions, and may cause exaggerated conclusions of quantitative reliability to be drawn. Studies that are poorly designed and do not adhere to established experimental design principles can and do lead to confounding of results and yield incomplete or incorrect conclusions. The following paragraphs provide recommendations about the evaluation and validation of QIBs that are in accordance with standard statistical principles and are intended to provide a framework for standardized evaluation of QIB performance.

Bias

Bias describes the difference between the mean (expected value) of measurements determined from the same object and its true value. Percentage bias is bias divided by the true value, presented as a percentage (percentage bias = [bias/true value] · 100).

Inherent in the definition of *bias* is knowledge of the true value of the measurand. In situations where the true value is unknown, which is often the case for in vivo imaging, estimation of bias is not possible (1). As noted earlier, pragmatic alternatives include the use of an agreed-upon reference value or evaluation of bias by using digital or physical phantoms. When using a reference value, it should be understood that only bias relative to the reference value is being assessed; similarly, linearity (see the following) relative to a reference value is assessed. Limitations of digital or physical phantoms are that they do not possess all of the characteristics of a human target and might not adequately account for measurement distortion induced by physical properties of living tissues. For example, even a complex magnetic resonance (MR) imaging phantom could not completely represent the magnetic properties of a human; therefore, bias and linearity from measurements acquired in a phantom may not represent the true quantity. However, phantom measurements provide a practical estimate of bias, and it is a reasonable assumption that poor phantom results with respect to bias would most likely be reflected as poor results in patients. Conversely, good results with a phantom or reference standard do not guarantee that one will get good results with respect to the (biologically) true value. When there is concern that phantom measurements may lead to false conclusions, then consideration should be given to using a reference method. It may also be helpful to test for sensitivity of the measurement to changes in the human acquisition environment.

When possible, when conducting assessments of bias, at least two and preferably three or more experimental replicates should be performed for each of several settings of "truth" (measurand levels) as measured according to a reference method. Various numbers of measurand levels have been proposed,

but sources for laboratory assays recommend at least five to seven levels (17). In phantom studies, these should be appropriately spaced over the measuring interval to adequately characterize the bias.

A similar effect can be achieved in patient studies by selecting patients to represent the whole spectrum of clinical characteristics (such as extent of disease or age) that are related to the QIB. When there is reason to believe that bias would change more rapidly in certain regions than in others, a higher concentration of measurand values in those regions of expected rapid change is necessary to complete the performance assessment. Commonly used metrics to characterize bias and related properties are summarized in Table 2.

Table 2

Metrics of Concordance with Truth

Metric	Comment
Estimate of bias	The difference between the mean of QIB measurements and the reference value (true value)
Limits of agreement	An interval that is expected to contain 95% of future differences between the QIB measurement and the measurand, centered at the mean difference (18)
Mean squared deviation	The expectation of the squared difference between a QIB measurement and the measurand (5)
Coverage probability	The proportion of cases for which the QIB measurement falls within a prespecified allowable difference between the QIB measurement and measurand (19)
95% Total deviation index	The threshold below which the difference between the QIB measurement and measurand is expected to fall with 95% probability

Linearity

Linearity is the ability to deliver measured values that are directly proportional to the true value of the measurand in the experimental unit (1). In image analysis, linearity is intrinsically driven by many factors associated with the formation of the image, as well as the quantification of it, with the device. For example, the scanner's spatial fidelity has a profound effect on measurements as the size of the biological phenomenon to be quantified approaches the resolution of the scanner. However, owing to the complexity of imaging systems technology, there are many other measurement characteristics that may affect the ability to transform computed results in such a way that the measurements are linearly related to measurands. The relationships between measured and true values may be proportional, nonproportional, or nonconstant (nonlinear). Examples of these possible relationships are illustrated in Figure 1.

Figure 1

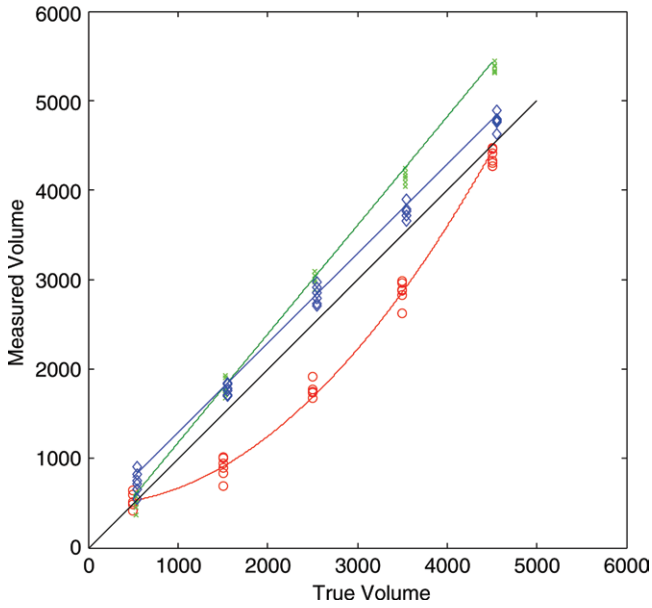

Figure 1: Plot of measured phantom volumes versus true volumes. The identity line (black line) with slope equal to 1 and intercept at zero would represent a biomarker with no bias. The blue line, where the slope is equal to 1 but the intercept is not zero, represents a biomarker with fixed or constant bias (ie, not proportional to the true values). The green line, where the slope is constant but not equal to 1, represents a biomarker with proportional bias. The red line, which is curvilinear, represents a biomarker with nonconstant (nonlinear) bias. See reference 4 for more detail.

The term *measuring interval* is defined as the range of the measurand in which bias, linearity, and precision are within acceptable bounds (1). In other words, the measuring interval is the range in which a change in the value of the measurand should result in a reliable change in the QIB measurement. Acceptable measurement interval limits will be determined by the clinical context of use to ensure that the QIB measurements can reliably indicate clinically important true changes in the QIB. Examples of determination of varying degrees of measuring intervals are found in the work of Echeverría et al in quantifying liver iron content with MR imaging (20). Methods differ in the ability to reliably estimate liver iron content over the entire range of values expected in patients with hemochromatosis. For example, the methods proposed by Gandon and others demonstrate linearity in patients

with normal to high liver iron content but are clearly nonlinear much beyond 200 μmol per gram of dry weight (20). Therefore, changes in liver iron content will not be reliable when including patients with extremely high iron content, and a reasonable measuring interval would be limited at the high end to 200 μmol per gram of dry weight.

Assessment of linearity should follow an approach similar to that for assessment of bias. The conditions under which the bias and reliability study is conducted should be representative of the conditions under which the QIB would be used clinically to provide confidence that performance estimated in the study will be representative of that experienced in clinical practice. The examination should include a preliminary assessment of whether the measured values are linearly related to the measurand on the basis of a simple scatterplot. Additional factors to consider include the range of true measurand values that will be covered, the sample size (number of distinct measurand values and number of replicate measurements obtained at each measurand value), where those QIB measurements will be obtained along the range of true measurand values, and under what conditions the measurements will be obtained. Measurements should adequately cover the measuring interval of the imaging measurement system. When comparing the QIB to phantom or truth data, nine to 11 levels of the true value are consistent with recommendations that have been proposed for linearity assessment for laboratory assays (21), but the optimal number will ultimately depend on the specific characteristics of the QIB. The measurand levels should typically be roughly equally spaced over the measuring interval, but it may be necessary to concentrate additional levels in areas that may violate the linear assumptions, such as the upper and lower ends of the measuring interval. Ideally, three replicates should be run at each level of the measurand, but the actual number of replicates needed depends on the repeatability and acceptable deviation from linearity.

Precision

Precision refers to variability of the measurement process regarding the expected value for different measurements on the same experimental unit, where conditions of measurement are either stable or vary over time according to temperature, operators, and so on (National Institute of Standards and Technology guideline 2.1.1.4) (22). Measurement variability is present even when conditions of measurement are not changed in any apparent manner and may be confounded with other factors when those conditions change. Variability in repeated measurements is dependent on the technical performance of the imaging device when the same experimental unit is measured according to stable test conditions. However, it is often difficult to maintain identical conditions, and measurement variability may also include contributions from other factors beyond variability inherent to the performance characteristics of the imaging device. For example, a repeated measurement of a tumor may occur within an interval of time that allows the tumor to grow and scanner conditions to vary, if even slightly, thereby contributing additional variability to the precision assessment. In a clinical setting, all of these sources of variability are included, and it is usually impossible to separate them.

While it is necessary to demonstrate that a QIB can be used to repeat a measurement reliably, it is also important to demonstrate that a QIB can be used with a more general set of conditions. *Repeatability* is a measurement of precision that occurs with identical or near-identical conditions. *Reproducibility*, in contrast, is a measurement of precision when location, operator, measuring system, or other factors differ. In reproducibility studies, the objective is to measure the effects of different conditions on the performance of the QIB with the goal of demonstrating equivalent performance in less restrictive study conditions. An example would be to demonstrate that scanners made by two different manufacturers result in equivalent tumor measurements and that both scanner types can be used in a single study.

In precision studies, measurements can be derived in a single phantom, a single lesion or subject, or a group of similar subjects (eg, healthy individuals). Precision studies performed with repeatability conditions are sometimes called *test-retest studies*. Truth or reference values are not necessary for repeatability measurements, since they are acquired with an assumption of no change in the measured object. Commonly used statistical metrics for repeatability are summarized in Table 3.

Repeatability studies are typically conducted with all measurements performed at a single clinical site with a specific imaging device. The measurements may be conducted in phantoms, animals, or human subjects. Phantom scans can be repeated several times in sequence, either immediately following one another or separated by a defined time interval. However, phantoms do not represent the complexity of human targets; thus, precision is often overestimated. The ability to perform repeated patient scans is limited because of safety concerns related to radiation exposure for CT or PET procedures, use of contrast media or tracers, or the need to account for kinetic behavior of those agents (eg, a washout period is needed before it is possible to rescan by using contrast media or a tracer). Furthermore, additional patient consent is required for scans that are performed for research purposes and not for clinical care. Human repeatability tests are often limited to two scans performed as test-retest, sometimes called "coffee break" experiments, with only a short break (eg, on the order of minutes or hours) between scans. See Table 4 for a comparison of phantom versus human test-retest study characteristics.

The attributes of the imaging system will define the repeatability conditions and the minimum time interval between repeat imaging sessions. Shorter time intervals minimize the variability introduced by ancillary factors. However, factors such as scanning period, radiation dose, contrast material washout, radioactive half-life, and subject

Table 3

Repeatability Metrics

Metric	Comment
Within-subject standard deviation	Standard deviation of replicate measures for a subject.
Limits of agreement, repeatability coefficient	Originally proposed by Altman and Bland (18), limits of agreement represent the interval that is expected to contain 95% of future differences between replicate measures. According to the assumption that test-retest measures are independent and identically distributed, the repeatability coefficient is equal to the half-width of the limits of agreement interval.
Intraclass correlation coefficient	The ratio of the variance of between-subject measurements to the total measurement variance (variance of between-subject measurements plus variance of within-subject measurements) in the sampled population. Comparison of intraclass correlation coefficients estimated from groups of subjects sampled from different populations can be misleading because intraclass correlation coefficients are scaled relative to the subjects in the study sample; thus, comparisons based on different populations can be invalid (5, 6).
Within-subject coefficient of variance	The within-subject coefficient of variance is the standard deviation of the replicate measures (within-subject standard deviation) divided by the mean.

Table 4

Inherent Characteristics of Phantom versus in Vivo Studies

Characteristic	Phantom or Digital Reference Image Studies	In Vivo Studies
Subject description	Physical or simulated models of the target of interest	Living human subjects
Realism	Approximation of real targets	Targets in living subjects
True value of the measurand known	Yes	No
Assessment of bias and linearity	Bias and linearity are assessable	Can only provide estimates of bias in zero-change scenarios
Level of evidence that the study provides	Serves as minimum performance requirement for bias and precision	Provides a more definitive assessment of precision
Expense	Low	High

fatigue place restrictions on the conditions of repeat scans. Other factors that are specific to each modality may have to be considered. For example, many imaging modalities have spatially varying measurement quality characteristics. CT scanners generally have better spatial resolution near the isocenter of the image than at the more peripheral regions of the image. If the subject is repositioned within the scanner, variability due to even slight differences in subject positioning will also be recorded. When determining the bias and precision of CT image measurements, it is important to indicate the region for which that characterization is valid. Options include either taking measurement samples over the whole field of view and characterizing the region with the poorest characteristics or specifying explicitly the subregion of the image for which the characterization is valid. For PET scanners, precision will depend on the relative count rate, so measurements with different amounts of radioactivity may need to be performed. Sequential repeat scanning within the same imaging session records effects due to scanner adjustments and image noise that define a base level of noise only above which a change in the QIB can be reliably measured and detected. A smaller change will be reliably detectable with optimal conditions in which there is no repositioning involved. With more variable conditions, minimum detectable change will likely be larger than that with the optimal conditions.

Although repeatability and reproducibility are presented here as distinct concepts, in practice they are very often inseparable, and together, they comprise the total variability of the QIB. Repeatability studies can be embedded within a larger study design that also includes multiple scanner or site reproducibility testing. Reproducibility studies are designed with the goal of evaluating different factors that may affect the QIB measurement. These factors may include clinical sites, scanner models or manufacturers, operators, technologists, radiologists, standard procedures, or any variable within a clinical trial that may affect the study results. Designs for reproducibility studies fall into two main categories:

Repeated Measurement Design

In each study subject or experimental unit, QIB measurements are acquired with multiple sets of reproducibility conditions. Replicate measurements may also be acquired within each reproducibility condition for individual subjects to estimate both reproducibility and repeatability.

Cohort Measurement Design

QIB measurements for different reproducibility factors are acquired in different subjects. This is especially useful when evaluating reproducibility between clinical sites because it is rarely feasible to transfer subjects from their local site to acquire images of the same subject at two or more clinical sites.

When reporting the results of a precision study, a description of the conditions of measurement should be provided. This is especially true if repeatability does not strictly apply. For example, the term *within-site precision*

can be used in reference to a set of conditions that includes different operators (technologists and/or radiologists), measuring systems, and replicate measurements of the same or similar objects within a single location (site). In that case, between-operator differences and between-instrument differences are expected to contribute to measures of imprecision (eg, standard deviation, coefficient of variation). Other factors that might vary in a within-site precision study could include date, time of day, and/or different scanner acquisition settings. Examples of image analysis factors that may vary include scanner hardware changes, scanner software changes, scanning protocol errors, patient motion, patient hydration state, and other patient characteristics that may affect performance of the imaging system.

Commonly used statistical metrics for reproducibility are summarized in Table 5. The estimated values for these metrics serve as indicators of whether studies conducted with varied reproducibility conditions can produce reliable results and can be confidently compared with other studies conducted with different conditions. If there is explicit interest in comparing reproducibility between two different imaging methods, formal statistical hypothesis tests may be performed to establish whether the methods have equivalent reproducibility or whether one is noninferior, inferior, or superior (5).

Table 5

Reproducibility Metrics

Metric	Comment
Reproducibility coefficient; limits of agreement	The interval that is expected to contain 95% of future differences between measurements obtained with different measuring conditions
Concordance correlation coefficient	A modified measure of correlation used to measure agreement between two continuous variables by penalizing the standard correlation coefficient for differences in the means (23)

Figure 2

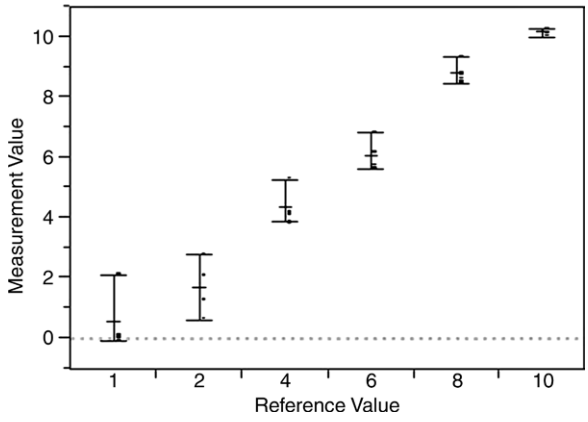

Figure 2: Precision profile box plot. At reference values (ie, the true value of the measurand) from 1 to 10, the mean and quartiles of the measured values are plotted. The heights of the box plots illustrate how the variance of the measured values decreases as the reference value increases.

Measuring Biological Change

One of the reasons we focus so much on precision is that a common clinical role for imaging biomarkers is to use QIB measurements to estimate biological change within a subject over time. To determine whether a measured change is likely a biological change, we must know with confidence how much of the measured change is due to technical variability. Change can be defined as either the simple difference between two QIB values or a difference relative to a reference QIB. Relative change can be defined with respect to baseline, average, nadir, peak, or reference value. Each definition has uses in different settings. It is therefore important to be explicit in defining how change is calculated (1,4).

To reliably assess whether a biological change has actually occurred, we need to understand the measurement precision profile. Precision is often not constant over the range of values of the measurand that are of clinical interest. This can be captured in a precision profile, which conveys how precision varies over the range of interest. An example is shown in Figure 2. Precision assessments should include components of variability that are relevant to the calculation of change in the measurand in the real-life setting or clinical study in which change is being measured. For example, if the same subject will be measured at times t_1 and t_2 by different operators, then both within- and between-operator components of variance will have an effect on what measured amount of change can be reliably interpreted as a true biological change. The span of time over which the measurements are obtained should also be considered because variability between measurements may increase with length of time between measurements.

Study Design Considerations

Figure 3 summarizes how study design recommendations differ, depending on whether one set of investigators is interested in assessing the performance of one imaging procedure to measure a QIB, comparing several procedures, or synthesizing performance across

Figure 3

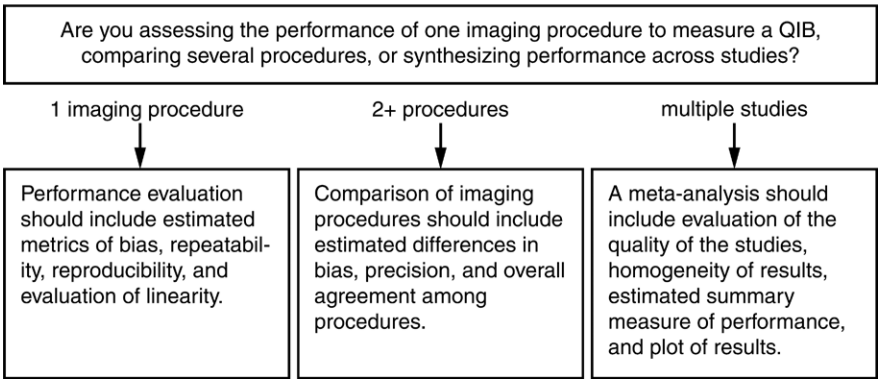

Figure 3: Flow diagram of the three main types of QIB performance evaluation studies. Estimates should also be accompanied by measures of uncertainty, such as 95% confidence intervals.

Table 6

Steps in Designing a QIB Technical Performance Study

Step	Example
Step 1: Define the QIB and its relationship to the measurand	Fluorodeoxyglucose uptake in gastric lesions will be measured as the SUV adjusted for lean body mass, or SUV_{lbm}, and is a measure of the integrated metabolic rate within a specified region of interest
Step 2: Define the study claim or question to be addressed in the analysis	Fluorodeoxyglucose uptake in gastric lesions as measured with SUV_{lbm} will increase proportionally with concentration of fluorodeoxyglucose, and the repeatability coefficient of changes in SUV will be no more than 20%
Step 3: Define the experimental unit	Lesions (gastric cancers)
Step 4: Define the technical performance metrics to be estimated	Linearity and precision (repeatability)
Step 5: Specify the elements of the statistical design and data requirements	Sample size, number of reviewers, technologists, radiologists, patient population, choice of reference measurement method (for bias and linearity studies), range of measurand values, choice of metrics for measuring performance, reproducibility conditions to be measured, repeatability time intervals, washout periods, and other conditions that will have implications in the eventual employment of the QIB in a controlled study; number of experimental subjects, data range, number of repeat measurements for repeatability and the number of conditions with which reproducibility will be assessed
Step 6: Statistical analysis	Choice of random or fixed effects to represent various factors that affect variability; specification of stratification and/or blocking factors; superiority, noninferiority or equivalence alternative hypotheses; estimation of performance metrics and measurement of uncertainty

studies. Table 6 lists the steps we recommend when designing studies to evaluate the technical performance of a single QIB. Not all steps will be applicable for every study but are presented here as useful guidelines for most QIB performance assessments. More detail is provided by Raunig et al (4).

Investigators often want to compare the technical performance of two or more competing imaging procedures to assess the typical performance of the procedures, to identify the best procedure, to test the noninferiority of a procedure relative to a standard procedure, or to identify procedures that provide similar measurements (6). Table 7 summarizes some common research questions asked in QIB procedure comparison studies and possible study designs used with each. When designing comparison studies, it is imperative that the performance metrics to be used for the primary analysis and the statistical hypothesis to be tested are specified in the planning phase of the study. It should be noted that procedures often have different strengths, for example, with one procedure being less biased but more precise than another. It is both possible and likely that one can reach different conclusions about the relative performance of procedures by using different types of metrics. For example, the study results may depend on whether "disaggregate" (bias and precision considered separately) or "aggregate" (bias and precision considered together in one summary measure, such as coverage probability) metrics of performance are chosen. Thus, it is important to specify a priori which metric will be used for the primary analysis in a study. Statistical methods for estimating the performance metrics and constructing the confidence intervals and for estimating differences in performance between procedures and corresponding confidence intervals should also be specified in the design phase.

Table 7 also provides references to statistical methods for comparing performance of QIB procedures. Further details and illustration of these methods can be found in references 5 and 6. Two commonly confused terms in the statistical analysis of comparison studies are *agreement* and *correlation*. Correlation is a weak criterion for assessing technical performance. The measurements from two procedures might be correlated with each other but may never agree with each other or agree with the true value of the measurements. Agreement of the findings of a new procedure with those of an existing one is

Table 7

Common Research Questions, Study Designs, and Statistical Methods for Comparing the Technical Performance of QIB Procedures

Research Question	Common Study Design	Common Statistical Methods
Which values from procedure(s) that provide measurements for individual patients or the population are closest to the true mean?	Phantom studies and test-retest studies to measure and compare bias	Repeated measures analysis or nonparametric tests (24) when the true value is known; error-in-variable models (25) or regression without truth (26,27) when the true value is unknown
Which procedure provides the most precise measurements with the same testing conditions?	Clinical studies that include replicate measurements for each procedure for each subject	Levene test (28) and mixed-effects models for comparing procedures
Which procedure(s) provides the most precise measurements with different testing conditions?	Clinical studies with at least one measurement for each procedure with each testing condition for each subject	Variance component analysis
Which procedure(s) is interchangeable with a commonly used reference procedure?	Clinical studies that involve replicate measurements with the reference standard for each subject and at least one measurement with each procedure for each subject	Estimation of the individual equivalence ratio and its confidence interval; bootstrap methods to construct the confidence interval for the individual equivalence ratio (29)
Which procedure findings agree with each other?	Clinical studies with at least one measurement with each procedure for each subject	Estimation of limits of agreement (30) or coverage probability (19) and their confidence intervals; bootstrap methods to compare procedures

stronger evidence, but when the findings from two procedures disagree, there is no way to determine which procedure is correct. Studies that include a comparison of the measurements from the procedures to the true value (eg, comparison of biases) provide the best assessment of the relative performance of procedures.

Statistical models and methods can sometimes be misleading when the bias and/or precision of algorithms vary in a systematic way over the range of measurements (5,18). For example, QIB procedures in which pulmonary nodule volume is measured often perform best for medium-sized lesions and may be biased and imprecise for small and large nodules (6). In these cases, the bias and precision profiles may need to be evaluated in subpopulations, such as those with different ranges of the measurements where performance is more homogeneous. Sometimes data transformations, such as the log transformation, can be used to remove dependence of bias and/or precision on the true value of the measurand (18).

Combining Results from Multiple Technical Performance Studies

Individual studies about the technical performance of an imaging procedure are often small, frequently containing as few as 10–20 patients (31–33) and typically focusing on a very specific subset of the population. Improved inferences on relevant technical performance metrics, as well as a more comprehensive understanding of technical performance across a variety of imaging technical configurations and clinical settings, can be achieved by combining results from multiple studies in a meta-analysis. Huang et al (7) provide a detailed description of the steps for conducting a meta-analysis of the technical performance of an imaging procedure.

Conclusion

QIBs offer tremendous promise for improving disease diagnosis, staging, and treatment and for expediting the regulatory approval of new therapies. However, while much progress has been made in the development of QIBs, few have been rigorously evaluated in terms of technical and clinical performance. The development and successful implementation of QIBs have been hampered by the inconsistent and often incorrect use of terminology and methods related to evaluation of the technical performance of these markers. We believe that implementation of the terminology and study design recommendations presented here for evaluation of technical performance will improve clinical research, advance regulatory science, and, in turn, lead to better care for patients who undergo imaging studies.

We have focused on metrics that summarize the bias and precision of QIBs and the agreement of QIBs with current measurements or clinical tests, because comparison of alternative or competing QIBs is often the question of interest. We have distinguished between statistical methods on the basis of knowing the true value and methods where the reference standard (ie, reference method) provides values measured with error. More work is needed to develop methods to predict QIB performance in actual target clinical populations from indirect measures of QIB performance. Shared data sets, improved realism of phantoms, and hybrid approaches (for example, simulations of realistic pathologic findings on normal images) are all areas worthy of further investment.

The emphasis of this article has been on unidimensional quantitative imaging measures and the associated algorithms. The future of QIBs will continue to move toward higher dimensional quantities, including architectures of tissue morphology, microstructure, or architecture manifested

in image textures, vector or tensor measures of fluid flows, localization and distance measures, perfusion, diffusion, and so on. Development of such QIBs is an active area of investigation and will require extensions of the methods considered in this article.

Many groups and organizations around the world are now devoting more attention to the evaluation and validation of QIBs. For example, the Foundation for the National Institutes of Health supports the Biomarkers Consortium, which funds several imaging biomarker projects, as do many of the professional imaging organizations in the United States and Canada. In Europe, the European Society of Radiology, the European Association of Nuclear Medicine, the European Organization for Research and Treatment of Cancer, and the Innovative Medicines Initiative all have active programs related to imaging biomarkers (34,35). With rigorous attention devoted to accurately and consistently describing exactly what physical phenomenon (and its relation to disease progress or outcome) is being measured, under what circumstances, and with what error, the clinical potential of QIBs can be achieved.

Acknowledgments: The authors acknowledge and appreciate the Radiological Society of North America and the National Institutes of Health (NIH) and National Institute of Biomedical Imaging and Bioengineering contract no. HHSN268201000050C for supporting two workshops and numerous conference calls for the authors' working groups.

The members of the RSNA-QIBA Metrology Working Group are Tatiyana V. Apanasovich, PhD, The George Washington University; Daniel P. Barboriak, MD, Duke University; Huiman X. Barnhart, PhD, Duke University; Andrew J. Buckler, MS, Elucid Bioimaging; Paul L. Carson, PhD, University of Michigan; Patricia E. Cole, PhD, MD, Takeda Pharmaceuticals; Ross Filice, MD, MedStar Georgetown University Hospital; Brian Garra, MD, Washington DC VA Medical Center and U.S. Food and Drug Administration (FDA); Constantine Gatsonis, PhD, Brown University; Maryellen L. Giger, PhD, University of Chicago; Robert J. Gillies, PhD, Moffitt Cancer Center; Dmitry B. Goldgof, PhD, University of South Florida; Mithat Gönen, PhD, Memorial Sloan-Kettering Cancer Center; Alexander Guimaraes, MD, PhD, Oregon Health Sciences University; Erich Huang, PhD, National Cancer Institute and NIH; Edward F. Jackson, PhD, University of Wisconsin–Madison; Jayashree Kalpathy-Cramer, PhD, Harvard–Massachusetts General Hospital; Larry G. Kessler, ScD, University of Washington; Hyun J. (Grace) Kim, PhD, University of California, Los Angeles; Paul E. Kinahan, PhD, University of Washington; Marina V. Kondratovich, PhD, FDA; Brenda F. Kurland, PhD, University of Pittsburgh; Lisa M. McShane, PhD, NIH and National Cancer Institute; Kyle J. Myers, PhD, FDA and Center for Devices and Radiological Health (CDRH); Nancy A. Obuchowski, PhD, Cleveland Clinic Foundation; Kevin O'Donnell, MASc, Toshiba Medical Research Institute–USA; Gene Pennello, PhD, FDA and CDRH; Nicholas Petrick, PhD, FDA, CDRH, and Office of Science and Engineering Laboratories; David L. Raunig, PhD, ICON Medical Imaging; Anthony P. Reeves, PhD, Cornell University; Kingshuk Roy Choudhury, PhD, Duke University; Lawrence H. Schwartz, MD, Columbia University; Adam J. Schwarz, PhD, Eli Lilly; Daniel C. Sullivan, MD, Duke University; Alicia Toledano, ScD, Biostatistics Consulting; James T. Voyvodic, PhD, Duke University; Richard L. Wahl, MD, Johns Hopkins University; Xiao-Feng Wang, PhD, Cleveland Clinic Foundation; Jingjing Ye, PhD, FDA and CDRH; Gudrun Zahlmann, PhD, F. Hoffmann–La Roche; and Zheng Zhang, PhD, Brown University.

Disclosures of Conflicts of Interest: D.C.S. disclosed no relevant relationships. **N.A.O.** disclosed no relevant relationships. **L.G.K.** disclosed no relevant relationships. **D.L.R.** disclosed no relevant relationships. **C.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author received payment from Wilex, Endocyte, Genentech, Philips Healthcare, and EBG Advisors for consulting; author received payment from Medical Imaging & Technology Alliance for lectures; author is a member of the medical advisory board for Wilex. Other relationships: disclosed no relevant relationships. **E.P.H.** disclosed no relevant relationships. **M.K.** disclosed no relevant relationships. **L.M.M.** disclosed no relevant relationships. **A.P.R.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author receives payment as the president and sole owner of D4Vision; author received grants from the Flight Attendant Medical Research Institute; author received payment for a patent; author received royalties from the Center for Technology Licensing at Cornell University; author has stock in Visiongate. Other relationships: disclosed no relevant relationships. **D.P.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is a member of the GE Medical Systems neuro-MRI advisory committee and receives nonfinancial support. Other relationships: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: author served on the Siemens speaker's bureau and served as an expert witness for the U.S. Department of Justice. **R.L.W.** disclosed no relevant relationships.

References

1. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res 2015;24(1):9–26.
2. Quantitative Imaging Biomarkers Alliance. http://rsna.org/QIBA.aspx. Accessed July 27, 2015.
3. Poste G. Bring on the biomarkers. Nature 2011;469(7329):156–157.
4. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res 2015;24(1):27–67.
5. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat Methods Med Res 2015;24(1):68–106.
6. Obuchowski NA, Barnhart HX, Buckler AJ, et al. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. Stat Methods Med Res 2015;24(1):107–140.
7. Huang EP, Wang XF, Choudhury KR, et al. Meta-analysis of the technical performance of an imaging procedure: guidelines and statistical methodology. Stat Methods Med Res 2015;24(1):141–174.
8. Stevens SS. On the theory of scales of measurement. Science 1946;103(2684):677–680.
9. Coxson HO. Quantitative chest tomography in COPD research: chairman's summary. Proc Am Thorac Soc 2008;5(9):874–877.
10. Dirksen A. Monitoring the progress of emphysema by repeat computed tomography scans with focus on noise reduction. Proc Am Thorac Soc 2008;5(9):925–928.
11. American College of Radiology. Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, Va: American College of Radiology, 2003.
12. International Organization for Standardization. http://www.iso.org/iso/home.html. Accessed August 30, 2014.
13. Joint Committee for Guides in Metrology. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms. http://www.nist.gov/pml/div688/grp40/upload/International-Vocabulary-of-Metrology.pdf. Accessed August 30, 2014.

14. Reeves AP, Jirapatnakul AC, Biancardi AM, et al. The VOLCANO'09 challenge: preliminary results. In: Brown M, de Bruijne M, van Ginneken B, et al, eds. The Second International Workshop on Pulmonary Image Analysis: London, UK, September 20, 2009. Scotts Valley, Calif: CreateSpace, 2009; 353–364.

15. Lou J, Obuchowski NA, Krishnaswamy A, et al. Manual, semiautomated, and fully automated measurement of the aortic annulus for planning of transcatheter aortic valve replacement (TAVR/TAVI): analysis of interchangeability. J Cardiovasc Comput Tomogr 2015;9(1):42–49.

16. Frey EC, Humm JL, Ljungberg M. Accuracy and precision of radioactivity quantification in nuclear medicine images. Semin Nucl Med 2012;42(3):208–218.

17. Clinical and Laboratory Standards Institute (CLSI). http://www.clsi.org/. Accessed August 30, 2014.

18. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2):135–160.

19. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. J Am Stat Assoc 2002; 97(457):257–270.

20. Alústiza Echeverría JM, Castiella A, Emparanza JI. Quantification of iron concentration in the liver by MRI. Insights Imaging 2012; 3(2):173–180.

21. Clinical and Laboratory Standards Institute/ National Committee for Clinical Laboratory Standards. Evaluation of the Linearity of Quantitative Measurement Procedures: a Statistical Approach; Approved Guideline. CLSI/NCCLS document. Wayne, Pa: National Committee for Clinical Laboratory Standards, 2010.

22. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. http://physics.nist.gov/Pubs/guidelines/appd.1.html. Published 1993. Accessed August 30, 2014.

23. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255–268.

24. Brunner E, Domhof S, Langer F. Test statistics. In: Nonparametric analysis of longitudinal data in factorial experiments. New York, NY: Wiley, 2001.

25. Dunn G, Roberts C. Modelling method comparison data. Stat Methods Med Res 1999;8(2):161–179.

26. Kupinski MA, Hoppin JW, Clarkson E, Barrett HH, Kastis GA. Estimation in medical imaging without a gold standard. Acad Radiol 2002;9(3):290–297.

27. Hoppin JW, Kupinski MA, Kastis GA, Clarkson E, Barrett HH. Objective comparison of quantitative imaging modalities without the use of a gold standard. IEEE Trans Med Imaging 2002;21(5):441–449.

28. Levene H. Robust tests for equality of variances. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB, eds. Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford, Calif: Stanford University Press, 1960; 278–292.

29. Barnhart HX, Kosinski AS, Haber MJ. Assessing individual agreement. J Biopharm Stat 2007;17(4):697–719.

30. Eden J, Levit L, Berg A, Morton S, eds. Finding what works in health care: standards for systematic reviews. Committee on Standards for Systematic Reviews of Comparative Effectiveness Research. Institute of Medicine. Washington, DC: National Academies Press, 2011.

31. Weber WA, Ziegler SI, Thödtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. J Nucl Med 1999; 40(11):1771–1777.

32. Hoekstra CJ, Hoekstra OS, Stroobants SG, et al. Methods to monitor response to chemotherapy in non-small cell lung cancer with 18F-FDG PET. J Nucl Med 2002;43(10): 1304–1309.

33. Minn H, Zasadny KR, Quint LE, Wahl RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. Radiology 1995;196(1):167–173.

34. European Society of Radiology (ESR). ESR statement on the stepwise development of imaging biomarkers. Insights Imaging 2013; 4(2):147–152.

35. Waterton JC, Pylkkanen L. Qualification of imaging biomarkers for oncology drug development. Eur J Cancer 2012;48(4):409–415.