The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions

Larry G Kessler, Huiman X Barnhart, Andrew J Buckler, Kingshuk Roy Choudhury, Marina V Kondratovich, Alicia Toledano, Alexander R Guimaraes, Ross Filice, Zheng Zhang, Daniel C Sullivan and QIBA Terminology Working Group

Stat Methods Med Res published online 11 June 2014
DOI: 10.1177/0962280214537333

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–18

DOI: 10.1177/0962280214537333
smm.sagepub.com
SAGE

The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions

Larry G Kessler,[1] Huiman X Barnhart,[2] Andrew J Buckler,[3] Kingshuk Roy Choudhury,[2] Marina V Kondratovich,[4] Alicia Toledano,[5] Alexander R Guimaraes,[6] Ross Filice,[4] Zheng Zhang,[7] Daniel C Sullivan[2] and QIBA Terminology Working Group

Abstract

The development and implementation of quantitative imaging biomarkers has been hampered by the inconsistent and often incorrect use of terminology related to these markers. Sponsored by the Radiological Society of North America, an interdisciplinary group of radiologists, statisticians, physicists, and other researchers worked to develop a comprehensive terminology to serve as a foundation for quantitative imaging biomarker claims. Where possible, this working group adapted existing definitions derived from national or international standards bodies rather than invent new definitions for these terms. This terminology also serves as a foundation for the design of studies that evaluate the technical performance of quantitative imaging biomarkers and for studies of algorithms that generate the quantitative imaging biomarkers from clinical scans. This paper provides examples of research studies and quantitative imaging biomarker claims that use terminology consistent with these definitions as well as examples of the rampant confusion in this emerging field. We provide recommendations for appropriate use of quantitative imaging biomarker terminological concepts. It is hoped that this document will assist researchers and regulatory reviewers who examine quantitative imaging biomarkers and will also inform regulatory guidance. More consistent and correct use of terminology could advance regulatory science, improve clinical research, and provide better care for patients who undergo imaging studies.

[1]University of Washington, Seattle, WA, USA
[2]Duke University, Durham, NC, USA
[3]Elucid Bioimaging Inc, Wenham, MA, USA
[4]Food and Drug Administration, Silver Spring, MD, USA
[5]Biostatistics Consulting, LLC, Kensington, MD, USA
[6]Harvard-Massachusetts General Hospital, Boston, MA, USA
[7]Brown University, Providence, RI, USA

Corresponding author:
Daniel C Sullivan, Duke University, Durham, NC, USA.
Email: daniel.sullivan@duke.edu

Keywords
Quantitative, terminology, definitions, bias, precision

1 Introduction of Quantitative Imaging Biomarkers Alliance (QIBA) and the Terminology Working Group

In response to the need for reliable and reproducible quantification of biomedical imaging data, the Radiologic Society of North America (RSNA) in 2007 organized the QIBA to unite researchers, healthcare professionals, industry stakeholders, and regulatory scientists and reviewers in the advancement of quantitative imaging and the use of quantitative imaging biomarkers (QIBs) in clinical trials and practice.[1] *QIBA's mission is to improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, patients and time*.

QIBA's emphasis is on building "measuring devices" rather than "imaging devices." That is, QIBA focuses on the signals that are recorded by an imaging device, processed into digital information, and then by means of mathematical algorithms turned into a number that can be interpreted by clinicians. During the first years of QIBA committee activity, it became apparent that there were inconsistencies and ambiguities across the committees in how they expressed quantitative measurements, concepts, or designed experiments to obtain needed data. Although there was recognition that it would make sense to adapt the long-accepted methodology for laboratory assay technical performance evaluation and validation, it was not always clear what the most appropriate adaptation would be for certain imaging methods or contexts of use. Therefore, QIBA and National Institute of Biomedical Imaging and Bioengineering decided to co-sponsor a workshop on metrology to obtain expert advice on terminology and methodology relevant to the above concerns. Three working groups were formed to address three main but interdependent issues: terminology, technical performance, and methods for algorithm comparison.

This paper is the result of the Terminology Working Group and is organized as follows. Section 2 presents the process of this working group; Section 3 discusses the concepts of biomarkers and quantitative imaging which lead to QIBs and terms related to their use; Section 4 lists recommended definitions of terms in alphabetical order; Sections 5 through 7 provide the working group's understanding and rationale behind these definitions, their proper use, as well as examples; and we conclude with a summary in Section 8.

2 Process of the Terminology Working Group

Between April 2012 and November2013, the Terminology Working Group held two face-to-face meetings and numerous conference calls. The terminology defined in this paper emerged from several sources. The terms come from the existing literature or authors (see the author list) who have direct experience in the development, analysis, and regulatory review of quantitative biomarkers, and from the other working groups on technical performance and algorithm comparison who suggested terms that they use for their work. Finally, we consulted various international standards bodies that have developed similar terms.

One of our guiding principles was to use internationally standardized and accepted terminology that is applicable in the context of QIBs. In this paper, we draw definitions from several sources, including the International Vocabulary of Metrology (VIM[2]), International Organization for Standardization (ISO[3]), Clinical and Laboratory Standards Institute,[4] and National Institute of Standards and Technology (NIST[5]), and relate them to QIBs. When there are discrepancies between sources, we describe them and make a recommendation or modification. When providing examples, we use the International System of Units (SI units) and their abbreviations.

In all other cases, we developed definitions, circulated them repeatedly throughout the working group, and then used conference calls to ensure we achieved consensus among the members of the working group. The definitions and explanations herein have been agreed to by all members of the working group.

3 QIB and its use

A widely accepted definition of a biomarker, used by Food and Drug Administration (FDA), is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or biological responses to a therapeutic intervention."[6] NIH uses a very similar definition for a biomarker.[7] The term "quantitative imaging" has recently been formally defined as

> the extraction of quantifiable features from medical images for the assessment of normal or the severity, degree of change, or status of a disease, injury, or chronic condition relative to normal. Quantitative imaging includes the development, standardization, and optimization of anatomical, functional, and molecular imaging acquisition protocols, data analyses, display methods, and reporting structures. These features permit the validation of accurately and precisely obtained image-derived metrics with anatomically and physiologically relevant parameters, including treatment response and outcome, and the use of such metrics in research and patient care. (https://www.rsna.org/QIBA.aspx)[8]

QIB is a term we apply to a numerical characteristic extracted from quantitative imaging and this numerical characteristic has properties of quantitative measurements.

3.1 What is a QIB?

Consider tumor volume (the measurand), which may be considered as a biomarker to describe a disease process, or, for example, a patient's response to therapy. Is tumor volume a QIB? Before we arrive at a formal definition of QIB, let's consider the following VIM definitions (all VIM citations found at Joint Committee for Guides in Metrology[2]):

> *Quantity [VIM, 1.1]: a property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference.*
> *For example of tumor volume, this reference can be a measurement unit, e.g. a cubic centimeter (cm^3) or a Becquerel (Bq).*
>
> *Quantity value [VIM, 1.19]: a number and reference together that express the magnitude of a quantity.*
> *For example: the volume of a given tumor, 2.0 cm^3, is a quantity value.*
>
> *Measurement [VIM, 2.1]: the process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.*
> *Measurand [VIM, 2.3]: the quantity intended to be measured.*

We define a QIB as "A quantitative imaging biomarker (QIB) is an objective characteristic derived from an in vivo image MEASURED on a ratio or interval scale as indicators of normal biological processes, pathogenic processes or a response to a therapeutic intervention."

QIBs will *consist either only of a measurand (variable of interest) or a measurand and other factors that may be held constant, and the difference between two values of the measurand is meaningful.*

Note that in some cases, *a clear definition of zero such that the ratio of two values of the measurand is meaningful* is required. In this context, "a clear definition of zero" means that zero indicates no signal is present. For example, in computed tomography (CT) a value of zero Hounsfield Units (HU) is not a meaningful zero because in the HU scale zero is arbitrarily defined as the CT attenuation of water. In other words, there is a CT signal present from the water, but it is arbitrarily assigned the value of zero.

Some imaging biomarkers are composed solely of a measurand, e.g. tumor volume. Other imaging biomarkers are functions of a measurand and other factors. An example of this is the standardized uptake value (SUV) obtained from positron emission tomography (PET) scans. Here, the measurand is tissue radioactivity concentration at some time after injection, and 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

$$SUV = \frac{tissue\ radioactivity\ concentration\ at\ time\ t}{injected\ dose\ at\ time\ zero/body\ weight}$$

3.2 Where are QIBs of use?

QIBs can be useful in both regulatory and clinical settings. In order to determine their applicability and validity, it is crucial that the framework in which they are acquired is well described including context of use, acquisition parameters, measurement methodology, and quantification of variability and error. Knowledge of these factors enables clinicians to reliably compare measurements over time and across imaging platforms. For example, comparing two or more SUV measurements over time could be clinically useful in cancer patients if the SUV measurements were reliable. Reliable QIBs can also help advance medical product development in the regulatory setting, for example if an imaging biomarker were qualified by the FDA for drug development, described by the FDA in their Biomarker Qualification Program (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284076.htm[6]). FDA-qualified imaging biomarkers could help move novel, safe, and effective medical therapies to the public through either traditional or accelerated approval pathways with appropriate postapproval follow-up.[9] Qualification refers to the rigorous evaluation of biomarkers for use in medical product development within the regulatory process. An important component of developing and validating a QIB is 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 following section gives the definition of QIB and terms related to its use in alphabetical order. These terms also serve as input to the groups reporting on consensus regarding performance measures and algorithmic comparisons for QIBs.

4 List of recommended definitions of terms in alphabetical order

Text box for all important definitions: Alphabetical

Agreement: ***Agreement is the degree of closeness between measurements made on the same experimental unit.***

Bias: ***Bias is an estimate of systematic measurement error; it is the difference between the average (expected value) of measurements made on the same object and its true value.*** **Percent bias is bias divided by the true value in percent.**

Biomarker: *A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or a response to a therapeutic intervention.*

Interval variable: *Measures for which the difference between two values is meaningful, but the ratio of two values is not, are called interval variables.*

Limit of blank: *A threshold above which measurements from the quantity with true state of measurand* $= 0$ *are obtained with probability* α *(probability of falsely claiming that the true state of measurand*>0*).*

Limit of detection: *Limit of detection (LoD) for a QIB is the measured quantity value, obtained by a given measurement procedure, for which the probability of falsely claiming that the true state of measurand* $= 0$ *is* β*, given a probability* α *of falsely claiming that the true state of measurand* >0*.*

Limit of quantitation: *The limit of quantitation (LoQ) is defined as the lowest value of measurand that can be reliably detected and quantitatively determined with {stated} acceptable precision and {stated, acceptable} bias, under specified experimental conditions.*

Linearity: *Linearity [ISO 18113[3]]: The ability to provide measured quantity values that are directly proportional to the value of the measurand in the experimental unit.*

Measurand: *Measurand [VIM, 2.3]: The quantity intended to be measured.*

Measurement: *Measurement [VIM, 2.1]: The process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.*

Measuring interval: The set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions.

Monotonicity: *The property of a variable such that it has relation of the form* $Y = f(X)$*, where* f *is a strictly increasing or decreasing function.*

Precision: *Precision is the closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar experimental units under specified conditions [VIM, 2.15].*

Profile claim: *A profile claim tells a user of the product or a biomarker what quantitative results can be achieved by the use of that product/biomarker in a clinical context.*

Quantitative Imaging Biomarker: An objective characteristic derived from an in vivo image MEASURED on a ratio or interval scale as indicators of normal biological processes, pathogenic processes, or a response to a therapeutic intervention.

Quantity: Quantity [VIM, 1.1]: A property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference.

Quantity value: *Quantity value [VIM, 1.19]: A number and reference together that express the magnitude of a quantity.*

Ratio variable: *A variable such that the difference between any two measures is meaningful and any two values have a meaningful ratio, making the operations of multiplication and division meaningful. A ratio variable possesses a meaningful (unique and nonarbitrary) zero value.*

Reliability: *Reliability is defined as the ratio of variance based on between-subject measurement to total variance based on the observed measurement.*

Repeatability: *Repeatability represents the measurement precision under a set of repeatability conditions of measurement.*

Repeatability condition of measurement: *The repeatability condition of measurement is derived out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same physical location, and replicate measurements on the same or similar experimental units over a short period of time [VIM, 2.20].*

Reproducibility: ***Reproducibility is measurement precision under reproducibility conditions of measurement [VIM, 2.25].***

Reproducibility condition of measurement: ***The reproducibility condition of measurement is derived from a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.***

Trueness: ***Trueness is the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value [VIM, 2.14].***

Uncertainty: ***Uncertainty is a nonnegative parameter characterizing the dispersion of the quantity values being attributed to a measurand.***

5 Characteristics of QIBs

5.1 QIBs are ratio or interval variables

The description earlier requires that to be considered a QIB, the corresponding measurands must be ratio or interval variables as defined by Stevens.[10] For example, tumor volume is a QIB 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 have real meanings: (1) the larger tumor is 1.0 cm^3 bigger than the smaller tumor; and (2) the larger tumor is three times the size of the smaller tumor. Tumor volume is thus a ratio variable as defined by Stevens.

For another example, PET SUV is a QIB because all factors for obtaining its value (i.e. injected dose, body weight, and time of measurement (t)) other than the measurand (i.e. concentration of radioactivity at time t) can be held constant, and the measurand is a ratio variable as defined by Stevens.[10] Consider two tumors receiving the same ratio of injected dose to body weight. If we look at SUV at time t and the tissue radioactivity concentration at that time (the measurand) is 5 MBq/kg for one tumor and 10 MBq/kg for the other tumor, then the following statements have real meanings: (1) the second tumor has 5 MBq/kg more radioactivity than the first tumor; and (2) the second tumor has two times as much radioactivity per unit mass as the first tumor.

Measures for which the difference between two values is meaningful, but the ratio of two values is not, are called interval variables.[10] We will refer to these types of variables as interval variables throughout this document. Examples of interval variables are described by Coxson[11] and Dirksen[12] where imaging biomarkers based on computed tomography (CT) are used to estimate the severity of emphysema, percent emphysema, and percentile density.

This paper does not address non-QIBs, i.e. measures for which values are assigned a magnitude, but neither the difference between two values nor the ratio of two values is meaningful. The scale of such values is sometimes called ordinal, because the ordering of values does have meaning (see VIM, 1.26; and Stevens[10]). Examples include, in mammography, the Breast Imaging Reporting and Data System (BI-RADS) Assessment Categories 1 (Negative) through 5 (Highly suggestive of malignancy) and Breast Composition Categories 1 (Almost entirely fat) through 4 (Extremely dense),[13] and in CT Colonography and Data Reporting System (C-RADS) finding categories C1 (Normal Colon or Benign Lesion; Continue Routine Screening) through C4 (Colonic Mass, Likely Malignant; Surgical Consultation Recommended).[14]

5.2 Properties of QIBs

5.2.1 Differences between variable types

The following table demonstrates some key differences between these types of variables and the statistical parameters that can be computed from each of these variable types. That certain variable types allow quantitative comparisons and calculations as shown below illustrates why this working group considers only interval and ratio variables as legitimate for QIBs.

Meaningful computation	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Addition or subtraction	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratio or coefficient of variation	No	No	No	Yes

5.2.2 Linearity

In order for the measurements to represent physical reality, they must bear a well-defined relationship to the underlying measurand. For that reason, it is desirable that the QIB has the property of linearity:

> Linearity [ISO 18113[3]]: the ability to provide measured quantity values that are directly proportional to the value of the measurand in the experimental unit.

In image analysis, linearity is intrinsically driven by many factors associated with both the formation as well as quantitation of the image from the device. For example, the scanner's spatial fidelity has a profound impact on accuracy of measurements as the size of the biological phenomenon to be quantified approaches the resolution of the scanner. However, due 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 that the measurements are in a linear relationship with measurands.

We consider linearity in the general sense that the measured quantity values (Y) bear a linear relationship of the form $Y = a + bX$, where X is the reference (or true) value of the measurand. A definition of linearity in the [ISO 18113] is proportionality, when the intercept (or offset) term $a = 0$. An example is the measured change in HU in iodinated contrast CT for measuring extravascular leakage.[15] A relationship of $Y = bX$ corresponds to QIB as a ratio variable and relationship of $Y = a + bX$ $(a \neq 0)$ corresponds to QIB as an interval variable.

While linearity is a sufficient condition for the measurements to bear a well-defined relationship to the measurand, more generally a strictly monotonic relation of the form $Y = f(X)$, where f is a strictly increasing or decreasing function, is necessary and sufficient to associate a unique measured value Y with every distinct value X of the measurand. This holds true when the form of f is known, because it is possible to transform the measurement Y as $Z = f^{-1}(Y)$, to produce a proportional relationship between Z and X and therefore the QIB has the property of linearity. An example of a multivariate monotonic relationship is the measurement of apparent diffusion coefficient (ADC) from diffusion weighted Magnetic Resonance Imaging (MRI) studies. ADC has been shown to depend on the concentration of fluid (as a sixth degree polynomial) and phantom temperature (as a quadratic) in phantom studies.[16]

5.3 QIBs and risk scores

The type of marker with numerical values which is a combination of several features or measurements (such a marker is sometimes called a "score") with a primary purpose of measuring current or future risk but does not represent any underlying biological quantity is not considered a QIB.

5.4 Need for QIB-related terminology

Early QIBA claims were inconsistent and had deficiencies, and that was the motivation for this Metrology Working Group. More recently developed QIBA Profile Claims, those that the QIBA work groups have used to describe the use of QIBs, illustrate the value of a more precise terminology with regards to QIB.

QIB claims refer to representations about the biomarker that provide a description of use and its benefits, including clinical value, safety, and effectiveness. Many publications illustrate problems both in the use of terminology and statistical measurements for QIBs. For example, authors frequently use Pearson Correlation Coefficient (r) as a measure of ‘‘accuracy’’ where ‘‘accuracy’’ is not clearly defined. The Pearson Correlation Coefficient only assesses the linear association between the two measurements and it does not assess the closeness or agreement between the measurements. In addition, the correlation coefficient depends on the range of the true quantity values: if the range is wide, the correlation will be greater than if the range is narrow. Studies where investigators compare two QIBs over a broad range of values may show a high correlation but the two QIBs can be in poor agreement.

QIBA Profiles state a QIBA Claim and also the specifications needed to meet that Claim. In the initial versions of QIBA Profile documents the structure of QIBA Claims varied widely across the QIBA Profiles and many of the Claims were incomplete. Therefore, the QIBA Metrology Working Group was asked to develop a template to structure QIBA Claims. It was decided that QIBA Claims should state the measurand, the clinical context, and the bias and precision that could be obtained by following the specifications in the QIBA Profile. Other metrics listed in the terminology earlier (such as linearity, limits of detection, etc.) could also be listed in a QIBA Claim, but for pragmatic reasons the QIBA Metrology Working Group recommended that bias and precision be listed at a minimum.

6 Assessment of QIBs

6.1 Measuring uncertainty of a QIB

To fully characterize a QIB, we believe it is necessary to present an indication of its uncertainty. Uncertainty is a nonnegative parameter characterizing the dispersion of the quantity values being attributed to a measurand. Uncertainty combines many components. Some components of uncertainty arise from systematic effects, e.g. bias (see definition below if true value is known). Other components of uncertainty arise from random factors, e.g. a lack of precision from the imaging device that makes the measurement, or biological variability of the measurand over time.

For these reasons, the QIBA Terminology Working Group recommends the use of clear definitions of terms when assessing the measures of dispersion made on the same experimental unit. To measure uncertainty, two typical approaches are used: disaggregated and aggregated approaches. A disaggregated approach uses more than one parameter to summarize different components of uncertainty, e.g. bias and precision while an aggregated approach uses one parameter, e.g. mean squared error (MSE), limits of agreement, intra-class correlation coefficient (ICC), etc. to summarize the uncertainty. Significant sources of uncertainty should be identified and the parameter measuring any of these sources should be stated explicitly. It is not sufficient to say, e.g. ‘‘Uncertainty is 10%.’’ To adequately characterize a QIB, it is necessary to say, e.g. ‘‘In this group of patients, the average coefficient of variation in PET SUV was 10%." The terminology associated with the two different approaches is presented in the following sections.

6.2 Measuring uncertainty via disaggregated approach

6.2.1 Bias

One characteristic of the technical performance of a device may be bias, which is an estimate of a systematic measurement error [VIM, 2.18].

Bias describes the difference between the average (expected value) of measurements made on the same object and its true value. In particular, for a laboratory measurement, bias is the difference (generally unknown) between a laboratory's average value (over time and presumably over a hypothetically infinite number of measurements) for an experimental unit and the average that would be achieved by the reference laboratory if it undertook the same measurements on the same experimental unit (http://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm[17]). Percent bias is bias divided by true value presented in percent (%Bias = Bias/(True Value)). If the true value is unknown, then the bias cannot be evaluated. In this situation, only precision can be evaluated.

6.2.2 Precision

Precision deals with variability. Variability is the tendency of the measurement process to produce different measurements on the same experimental unit, where conditions of measurement are either stable or vary over time, temperature, operators, etc.

Variability occurs even when conditions of measurement are not changed in any apparent manner and it is compounded when those conditions differ. Variability in measurements is often related to the performance characteristics of the imaging device when the same experimental unit is measured under stable test conditions. Variability in measurements is related to factors beyond just the performance characteristics of the imaging device if the same test item is measured under different test conditions. For example, if a tumor has a certain volume at diagnosis and has a smaller volume after some course of treatment, we expect the measured tumor volume to decrease. If a tumor has a certain volume and it is measured independently by two different radiologists, and one of those radiologists tends to include more voxels at the margins, we expect the two measurements of tumor volume to be different. In a clinical setting, all of these sources of variability are included, and it is usually impossible to separate them.

The QIBA Terminology Working Group therefore recommends evaluating separately the components contributing to variability, explicitly identifying each source and stating the parameter being used to describe it (e.g. standard deviation, coefficient of variation, ICC, etc.).

Precision is the closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar experimental units under specified conditions [VIM, 2.16]. As noted in the VIM, precision is usually expressed numerically by measures of imprecision under the specified conditions of measurement, for example, within-subject standard deviation (wSD), within-subject coefficient of variation (wCV), or 95% precision limit. Note that within-subject standard deviation for *imaging precision studies* means the standard deviation from measurements from the same (or similar) subject under specified conditions; the term "within-subject SD" in another context may be used to describe biological variation of subject true measurand values. This variation may occur in particular when measurements are taken in very similar experimental conditions but taken over a period of time and this variation might then reflect not measurement error or lack of precision, but rather reflect biological change when homeostasis had been assumed.

We recommend when reporting significant change for a QIB profile, only precision, not biological variation, be included within an individual.

Sometimes the precision is expressed by a scaled agreement index such as ICC.[18] The common term for 95% precision limit is repeatability coefficient (RC) or reproducibility coefficient (RDC) (see below for difference between repeatability and reproducibility).[18] In practice, sometimes examination of the difference between two device results is required and for this, a critical difference (or another term for this is "significant difference") should be considered. The value of this critical difference depends on the probability level with which it is associated (usually, this level is 95%) and on the shape of the underlying distribution. For example, if measurements of the same subject follow a normal distribution and the probability level is specified as 95%, this critical difference is 2.77σ where σ is the wSD. If σ corresponds to the wSD under repeatability conditions, the value 2.77σ is sometimes called the RC and it is called a RDC if σ corresponds to the wSD under reproducibility conditions. Precision may be different for different experimental units with different mean values (or true values if known and available) and thus the precision values over a range of mean values (or true values) provide the precision profile. The precision profile is a function between QIB mean values (or other factors) and precision parameters such as wSD (or wCV). If the precision is assumed to be the same across a range of mean values (true values), the pooled precision may be reported for this range of mean values (true values) with this assumption stated.

The number of significant digits in the measurement obtained should reflect the precision. The "specified conditions" can be, for example, repeatability conditions of measurement or reproducibility conditions of measurement where their differences are clarified below.

> Repeatability represents the measurement precision under a set of repeatability conditions of measurement.
>
> The repeatability condition of measurement is derived out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same physical location, and replicate measurements on the same or similar experimental units over a short period of time [VIM, 2.21].

An important and related but different concept is reproducibility.

> Reproducibility is measurement precision under reproducibility conditions of measurement [VIM, 2.25].
>
> The reproducibility condition of measurement is derived from a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.

Compared with repeatability, reproducibility still requires the same measurement procedure, the same operating conditions, and a short period of time between measurements. It is only location, operator, and/or measuring system that may differ.

Precision studies can make their measurements on a single phantom, a single lesion/subject, or a group of similar subjects (e.g. healthy individuals). Precision studies performed under repeatability conditions are sometimes called test–retest studies.

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, within-site precision can be used for a set of conditions that includes different operators (technologists, radiologists), measuring systems, and replicate measurements on the same or similar objects within a single location (site). In that case, between-operator differences and between-instrument differences will contribute to parameters measuring precision (e.g. standard deviation, coefficient of variation). Other parameters that might vary in a within-site precision study could include date, time of day for scan, and/or different scanner acquisition settings. Examples of variation in parameters for image analysis include scanner hardware changes, scanner software changes, scan protocol errors, patient motion, patient hydration state, and other sources of variability between patients.

The QIBA Terminology Working Group recommends explicitly describing the conditions of measurement; identifying each source of uncertainty, variability, or imprecision; reporting the range of measurand values for which precision was evaluated; and stating the parameter being used to describe it. We do not recommend making vague statements such as: "The uncertainty is 10%," "The variability is ±5," or "The precision is $\pm 0.1\,\mathrm{cm}^3$." We do recommend making statements such as: "The coefficient of variation for these tumor volume measurements was 10% for tumors of 2 cm in diameter," "The standard deviation of these tumor volume measurements was $0.1\,\mathrm{cm}^3$ for tumors of $2\,\mathrm{cm}^3$ in diameter" "The standard deviation of these PET SUV measurements was 2.5," or "The 2.5th and 97.5th percentiles of PET SUV values are 4.2 and 14.2, respectively."

6.3 Measuring uncertainty via an aggregated approach

6.3.1 Accuracy and trueness

The degree of uncertainty can be measured using bias and precision separately. Due to the inherent trade-off between bias and precision, it is often of interest to evaluate the aggregated impact of bias and precision on the overall closeness of measurements with truth. Here, we introduce related terms.

Two terms related to precision and bias that should be used with care if they must be used at all are accuracy and trueness. Accuracy is a commonly used term. It has been variously used to describe a range of characteristics including how a measured value relates to a known physical reference or the ability of a diagnostic test or other measure to identify or characterize a complex disease process.

Although the conventional uses of the term "accuracy" are most often related to bias, precision and trueness (defined below), the term accuracy should not be used as a synonym for any of those. If the term accuracy is used, it should be defined in such a way that anyone reading the report can reproduce the calculation.

> *Trueness is the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value [VIM, 2.15].*

As noted in the VIM, trueness is not a quantity and thus cannot be expressed numerically. Trueness is inversely related to systematic measurement error, but is not related to random measurement error. If the reference quantity value is the truth, then trueness is described by bias (or by percent bias). Although related to accuracy, trueness should not be used as a synonym for accuracy.

6.3.2 Agreement and reliability

When the true value may or may not be known, there is a broader term called agreement that has much utility. In certain cases, the narrower term reliability is often used in practice when the true value is not known. We recommend the following definitions of agreement and reliability:

> *Agreement is the degree of closeness between measurements made on the same experimental unit.*

If the true value is available, the agreement between the measured value and the true value provides the overall aggregated impact of bias and precision. If the true value is not available, only precision components can be assessed. Different precision parameters defined earlier, such as wSD, wCV, RC, RDC, ICC, are examples of indices for assessing agreement between repeated measurements on the same experimental unit under the same or different conditions. More generally, agreement is a broad term that encompasses an aggregated assessment of bias and precision.

It should be noted that agreement is an abstract concept. In absolute/literal terms, measurements agree only if they are identical and disagree if they are not identical. Agreement is not used literally as a binary concept here as yes/no (i.e. perfect agreement or not), rather it is used to describe the degree of closeness between measurements.

Many different agreement indices exist in the literature.[18] Each index has a specific meaning for its interpretation because it depends on the specific conditions under consideration. Some of the indices have direct interpretation in terms of the measurement value, e.g. wSD, limits of agreement. Other indices may be dimensionless and do not have direct interpretation in terms of the measurement value, e.g. ICC. However, all indices theoretically have a value that corresponds to perfect agreement, e.g. 0 for limits of agreement and 1 for ICC. The concept of agreement may be used in settings both with and without a known true value, although it has traditionally been used for the case without a known true value. As is the case with many imaging and biomarker situations, the true value is a difficult value to measure and to have available for study.[19] Agreement is a broader concept than reliability in the sense that reliability is a subset of indices for assessing agreement.

> *Reliability is defined as the ratio of variance based on between-subject measurement to total variance based on the observed measurement.*

In nonmathematical terms, reliability is used to describe the overall relative consistency of a measure to between-subject variability. A measure is said to have a high reliability if it produces similar results relative to between-subject variability. Reliability is usually the focus of the performance of the QIBs in settings without a known true value and is often defined with additional assumptions. Different assumptions lead to different versions of the ICC for assessing reliability. Reliability is an agreement measure because a value of 1 corresponds to perfect agreement and values less than 1 correspond to degrees of agreement where smaller values indicate less agreement. Reliability is a dimensionless index that does not have direct interpretation in terms of measurement value. Therefore, it is difficult to understand and judge how large a value of reliability constitutes adequate reliability. Although Landis and Koch[20] provided adjectives to describe ranges of reliability values as (0, 0.20) – slight, (0.21, 0.40)—fair, (0.41, 0.60)—moderate, (0.61, 0.80)—substantial, (0.80, 1.0)—almost perfect, these ranges are subjective. Contrary to intuition, sometimes very different values of reliability may result from data with the same difference between measurements for different populations.[21] Reliability is a scaled index that is relative to the between-subject variability, and it recently has been recognized that its magnitude may be better interpreted as the ability of a measurement to differentiate between experimental units (e.g. subjects or objects).[22]

6.3.3 The use of the MSE concept

MSE or mean standard deviation[18] is another agreement index which can be easily understood statistically in terms of the trade-off of bias and precision. In statistical terminology, a QIB $\hat{\theta}$ is an estimate of truth, an underlying physiological quantity or parameter θ. The agreement between a biomarker $\hat{\theta}$ and truth can be quantified as the MSE: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$. For imaging biomarkers, this average is typically taken over the distribution of image noise, which is assumed to have some probability distribution (across space and/or across replicate images). The amount of this error can also depend on a number of other factors: (a) the nature of the biomarker itself; (b) the imaging parameters set on the device; (c) the measurand related to the underlying physiological parameter, which can vary across subjects, time, spatial region, etc.[23] The bias-variance decomposition, $MSE(\hat{\theta}) = Var(\hat{\theta}) + B^2(\hat{\theta})$, enables the decomposition of MSE into two components: systematic error or bias, $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ and variance, $Var(\hat{\theta})$, where the variance component may or may not be the precision parameter. Specifically, the variance is wSD^2, a precision parameter, only if the MSE is averaged over the distribution across replicated images. The variance would contain both variations due to replication as well as due to subjects if the MSE is averaged over the distribution of subjects. Imaging biomarkers $\hat{\theta}$ typically involve a tuning parameter which can be adjusted to affect the value of MSE. An example of such a tuning parameter is the size of region of interest used to define the SUV peak measurement in PET imaging.[24] Adjusting the tuning parameter usually has reciprocal effects on the bias and variance components of MSE: lowering variance implies increasing bias and vice versa. However, the changes in bias and variance are not necessarily equal in magnitude. Hence it may be possible to choose an optimal value of λ which minimizes MSE by trading off bias against variance.[25]

By definition MSE can only be evaluated in settings where the truth θ is known. However, only the bias term involves truth. This implies that different types of studies can be used to evaluate the two components of MSE: (a) variance, which does not require knowledge of θ, can be evaluated from studies where no independent assessment of θ is available (precision studies); (b) bias does require knowledge of θ. Reliable independent assessment of θ can be difficult or impossible to obtain clinically for some biomarkers. For this reason, bias may need to be evaluated in somewhat artificial studies. Two strategies can be used. One strategy uses computer simulation: this involves the construction of a mathematical model of the underlying physiological process (sometimes called a digital phantom) as well as the imaging device. Data are simulated from this combined mathematical model and the biomarker is computed using an objective procedure which does not involve knowledge of the underlying physiological parameter θ. Bias and variance can be evaluated by appropriately averaging over repeated simulations.[26] The applicability of results from computer simulation is limited by the extent to which the physiological/anatomic and imaging models are realistic. The second strategy employs physical phantoms: a physical phantom is an artificial construct which is intended to mimic the behavior of the human/animal body or part thereof. The advantage here is that it is possible to invasively measure values of the physiological parameter of interest (θ). The physical phantom is placed in the actual imaging (scanner) setup and imaged similar to a clinical study. Like computer simulation, it is typically possible to image the phantom repeatedly and/or under a variety of conditions, hence it is possible to evaluate both bias and variance from these repeat studies.[27] The applicability of results from physical phantom studies is limited by the extent to which the phantom model realistically represents human physiology and/or anatomy. The bias-variance decomposition enables the independent evaluation of the two components of MSE at different levels of realism, e.g. variance in a clinical study, bias in a physical phantom. However, care must be exercised to ensure that similar values of the tuning parameter are used in all assessments and caution is needed in interpreting the MSE with estimated values of bias and variance from two different studies under different conditions.

The MSE provides an intuitive assessment that combines both components of bias and precision. Its magnitude is often difficult to interpret for judging adequate agreement. For example, it is not easy to choose a cutoff for MSE such that we know the measurements are close to the truth for majority of the measurements. There exist other agreement indices[18] whose magnitudes are easier to interpret, e.g. limits of agreement, coverage probability (CP) and total deviation index (TDI). For example, if we require 95% of the measurements to be close to the true value by some distance, the limits of agreement provide the interval of such distances (centered around bias) between measurement and truth for 95% of measurements. Similarly, TDI provides the same type of distances but centered around zero. The CP is the percent of measurements whose values are not farther than a prespecified distance from truth.

6.3.4 Traceability and commutability

Two desirable properties of QIBs are *traceability [VIM, 2.41]* and *commutability [VIM, 5.15]*. Traceability occurs, for example, when we calibrate the machine using a reference standard which is recognized by international organizations (e.g. NIST). This reference can be a physical test object (phantom) or some digitally generated (synthetic) data. Consider, for example, a phantom containing a lesion with volume 5.0 cm^3 with some low level of uncertainty. Traceability implies that if it were possible to take an infinite number of identically repeated measurements of that lesion by different quantitative imaging systems measuring that lesion, for every system with traceability, the average of system measurements would match the reference value. Commutability occurs when the reference material reflects the "routine samples" (e.g. physical lesions in patients being imaged). It is what allows us to say that if a machine can measure tumor volume on a phantom with some precision and lack of bias, measurements of tumor volumes made on that same machine using a corresponding measurement procedure have the same precision and lack of bias for physical lesions in patients.

6.3.5 Limit of blank, LoD, and LoQ

Other characteristics of a QIB include the limit of blank (LoB), LoD, and LoQ [Linnet et al.,[28] CLSI EP17-A2[29] and VIM 3.18]. LoQ related to the measuring interval is an important characteristic for QIBs. *The LoQ is defined as the lowest value of a measurand that can be reliably detected and quantitatively determined with acceptable precision and bias, under specified experimental conditions.*

To prespecify an acceptable difference between any two measurements on the same subject is similar to prespecifying a clinically significant difference when designing a randomized clinical trial. The choice of acceptable difference must be based on clinical knowledge; it is not an intrinsic part of the QIB or its measurement. First of all, we want to choose an acceptable difference that is small enough so that we can be sufficiently confident that any change that is larger than this difference is not due to measurement error. On the other hand, we do not want to set an acceptable difference that is too small to be humanly achievable. If some magnitude of difference in the measurement is strongly related to clinical outcomes, this would be a good choice for an acceptable difference. Typically the acceptable difference should be smaller than a clinically significant difference because there is no point in detecting any difference that is due to measurement error.

Because precision and bias can depend not only on the value of the measurand but also on other factors, LoQ can be different for different combinations of these factors. *LoD for a QIB is the measured quantity value, obtained by a given measurement procedure, for which the probability of falsely claiming that the true state of measurand = 0 is β, given a probability α of falsely claiming that the true state of measurand >0.* The LoD of a measurement procedure is the lowest amount which can be detected (above LoB) (minimal detectable amount) but not necessarily quantitated as an exact value. The definitions of LoQ and LoD imply that LoQ $\geq$LoD. An example of stating LoQ in imaging (from the QIBA CT Volumetry Profile for Solid Tumors) is as follows. The LoQ is 10 mm, with lesions where:

(1) The tumor possesses sufficient conspicuity to allow boundaries to be adequately demarcated from surrounding tissue.
(2) The tumor morphology is not unduly complex.

(3) The tumor composition is sufficiently homogeneous, or various tissue types within a mass can be segmented from each other.
(4) These tumors can be measured with %CV of precision of less than 20%, as an example of acceptance criterion. [N.B. the actual difference or the SD could be used instead. They have the advantage that they are not scaled measurements.]

Measuring interval is a set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions.

Sometimes other terms such as reportable range or working interval are used, but we do not recommend usage of these terms. The lower limit of measuring interval is the LoQ; therefore, LoQ is called sometimes LLoQ (lower limit of quantitation). The upper limit of the measuring interval is called ULoQ (upper limit of quantitation).

Information about measuring interval for QIBs should be provided along with information concerning deviations from linearity, acceptable precision, and acceptable bias (if applicable).

7 Measuring change in a QIB

Clinical decision making, one of the most important reasons for developing, defining, and using QIBs, often depends on information about a change over time. For example, we know that parenchymal changes in a lung nodule may be a marker of serious illness, such as lung cancer.

Making decisions about true states of change in a measurand within a subject is the most common situation we face that requires the ability to measure change. To properly understand the nature of change, one must provide four kinds of information: definition of change, number of directions, precision profile, and error rates. For simplicity, consider a subject for whom a measurand was measured at two times: the measurements are X1 at time T1 and X2 at time T2. Further, let both X1 and X2 be above the LoQ, such that numerical values are reported.

Change can be defined as either absolute or relative. Absolute change is the difference X2 − X1. Relative change can be defined relative to baseline, average, or nadir. Change relative to baseline is (X2 − X1)/X1 if X1 is the baseline measurement, or as (X2 − X1)/X0 if the baseline measurement is X0 obtained at time T0 prior to T1. Change relative to average is (X2 − X1)/XA, where the average measurement XA may include only X1 and X2, XA = (X1 + X2)/2; or may also include measurements in larger time windows. Change relative to nadir is (X2 − X1)/XN, where XN is the lowest measurement in a defined time window. Each definition has uses in different clinical settings. It is therefore important to be explicit in defining how change is calculated whenever one is measuring change.

The second aspect of measuring change is directionality, i.e. whether one is interested in change in a single direction or in change in either direction within a subject. Measuring change in one direction involves determining whether true change in measurand = 0 versus true change in measurand <0 (measuring a decrease), or determining whether true change in measurand = 0 versus true change in measurand >0 (measuring an increase). Measuring change in two directions involves determining whether true change in measurand = 0 versus true change in measurand ≠ 0.

In order to know whether a change has occurred, we need to know about the distribution of change values. This relates to the precision profile of a measurand. Precision should include components of variability which 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 T1 and T2 by different operators, the precision profile that includes a between-operator component of variance should be considered. Finally, two types of error could be considered: type I error and type II error. If we are interested in change in one direction, significant change is the value of observed change C such that if true change = 0, observed change is further from 0 than C with a low probability. That probability is the type I error rate; type I error is deciding that a change occurred when in fact true change = 0. The type I error rate for determining a value of significant change is usually set at 5%, i.e. $\alpha = 0.05$. The distribution of change values determines the relationship between change values and the probability of their occurrence. This distribution should be described when reporting a value of significant change—for example, stating that one assumed a normal distribution with some mean and standard deviation. If true change = 0 and we are measuring change in two directions, we will report the thresholds $[C_{\text{decrease}}, C_{\text{increase}}]$ such that the probability that observed change is outside this interval is equal to the desired type I error rate.

The concept of "minimal detectable change" relates to type II error, which happens if we decide that no change exists when in fact a true change in measurand of some amount, C_{true}, has occurred. The amount of true change determines the center of the distribution of change values and sometimes also plays a role in determining the spread of that distribution, such that different values of minimal detectable change must be obtained for each different value of true change. Given a distribution of change values, the amount of observed change in excess of which we will decide that change has occurred 95% of the time is referred to as the 95% minimal detectable change. This is associated with a type II error rate of 100% − 95% = 5%, i.e. $\beta = 0.05$.

An example of measuring change comes from the QIB of PET FDG SUVmax. ^{18}F-FDG (fludeoxyglucose) is a glucose analogue, often abbreviated as FDG or ^{18}F-FDG. The rationale for its use in oncology is based on the typically increased rate of glycolysis in tumors compared to normal tissue. FDG-PET scans are sensitive and specific for detection of most malignant tumors.[30] FDG-PET scans reflect glucose metabolic activity of cancers and this metabolic activity can be measured with a high degree of reproducibility over time. Longitudinal changes in tumor ^{18}F-FDG accumulation during therapy often can predict clinical outcomes earlier than changes in standard anatomic measurements.[31] Based on the QIBA Profile for FDG-PET (http://rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA_FDG-PET_Profile_v105_Publicly_Reviewed_Version_FINAL_11Dec2013.pdf),[32] tumor glycolytic activity as reflected by the maximum standardized uptake value (SUVmax) should be measurable from FDG-PET/CT with a within-subject coefficient of variation of 15% and mean value of M. A within-subject coefficient of variation of 15% implies a 95% RC of±41.6%*M. Because M is unknown one should consider a relative change as $2*(X2 - X1)/(X2 + X1)$ (M is estimated by the average of X2 and X1) or by a relative change $(X2 - X1)/X1$ (M is estimated by X1). Significant change for the $2*(X2 - X1)/(X2 + X1)$ is −42.5% for significant decrease and +42.5% for significant increase. This implies that any observed relative change $|2*(X2 - X1)/(X2 + X1)|$ above 43% may be interpreted as the true change.

8 Summary

The development and implementation of QIBs has been hampered by the inconsistent and often incorrect use of terminology related to these markers. With recent initiatives by the RSNA and by the FDA, (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentTools QualificationProgram/ucm284076.htm),[6] interest in biomarkers in general and QIBs in particular is growing.

There are rather daunting scientific challenges for QIBs in the development, validation, and measurement of their clinical utility. For example, QIBs may have the potential to serve as surrogate endpoints in drug (or device, or biologic, or other clinical) trials. However, the bar that has been set from a statistical perspective is quite high[33] and many potential QIBs may not yet be at the stage of development to meet that bar.

To address the inconsistent use of metrology definitions, where possible we have drawn our definitions from existing national or international standards rather than invent new definitions for these terms. We provide recommendations for appropriate use of QIB terminological concepts. We hope that this document will assist researchers and regulatory reviewers who examine QIBs and inform regulatory guidance. More consistent and correct use of terminology could advance regulatory science, improve clinical research, and provide better care for patients who undergo imaging studies.

Acknowledgement

The authors acknowledge and appreciate the Radiological Society of North America and NIH/NIBIB contract # HHSN268201000050C for supporting two workshops and numerous conference calls for the authors' Working Group. We would also like to thank Michael Boss of NIST and Jingjing Ye and Shashi Amur of the FDA for comments on earlier versions of the paper.

References

1. Buckler AJ, Bresolin L, Dunnick NR, et al. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* 2011; **258**: 906–914.
2. 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 27 November 2011).
3. International Organization for Standardization, http://www.iso.org/iso/home.html (accessed 1 August 2012).
4. Clinical and Laboratory Standards Institute (CLSI), http://www.clsi.org/ (accessed 1 August 2012)
5. *Guidelines for evaluating and expressing the uncertainty of NIST measurement results*, http://physics.nist.gov/Pubs/guidelines/appd.1.html (1993, accessed 27 November 2011).
6. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284395.htm.
7. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001; **69**: 89–95.
8. https://www.rsna.org/QIBA.aspx.
9. http://www.fda.gov/forconsumers/byaudience/forpatientadvocates/speedingaccesstoimportantnewtherapies/ucm128291.htm#accelerated.
10. Stevens SS. On the theory of scales of measurement. *Science* 1946; **103**: 677–680.
11. Coxson HO. Quantitative chest tomography in COPD research: Chairman's summary. *Proc Am Thorac Soc* 2008; **5**: 874–877.
12. Dirksen A. Monitoring the progress of emphysema by repeat computed tomography scans with focus on noise reduction. *Proc Am Thorac Soc* 2008; **5**: 925–928.
13. *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*, A.C.o.R. (ACR), Editor. Reston, VA: American College of Radiology, 2003.
14. Zalis ME, Barish MA, Choi JR, et al. CT colonography reporting and data system: A consensus proposal. *Radiology* 2005; **236**: 3–9.
15. O'Connor JP, Jackson A, Asselin MC, et al. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *Lancet Oncol* 2008; **9**: 766–776.
16. Matsuya R, Kuroda M, Matsumoto Y, et al. A new phantom using polyethylene glycol as an apparent diffusion coefficient standard for MR imaging. *Int J Oncol* 2009; **35**: 893–900.
17. http://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm.
18. Barnhart HX, Haber MJ and Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat* 2007; **17**: 529–569.
19. Sullivan DC and Gatsonis C. Response to treatment series: part 1 and introduction, measuring tumor response—challenges in the era of molecular medicine. *Am J Roentgenol* 2011; **197**: 15–17.
20. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
21. Nevill AM and Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med* 1997; **31**: 314–318.
22. Kottner J, Gajewski BJ and Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *Int J Nurs Stud* 2011; **48**: 659–660.
23. Malyarenko D, Galban CJ, Londy FJ, et al. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *J Magn Reson Imaging* 2012; **37**: 1238–1246.
24. Vanderhoek M, Perlman SB and Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response. *J Nucl Med* 2012; **53**: 4–11.
25. Ramani S, Liu Z, Rosen J, et al. Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods. *IEEE Trans Image Process* 2012; **21**: 3659–3672.

26. Doot RK, Muzi M, Peterson LM, et al. Kinetic analysis of 18F-fluoride PET images of breast cancer bone metastases. *J Nucl Med* 2010; **51**: 521–527.
27. Korukonda S and Doyley MM. Visualizing the radial and circumferential strain distribution within vessel phantoms using synthetic-aperture ultrasound elastography. *IEEE Trans Ultrasonics Ferroelectrics Frequency Control* 2012; **59**: 1639–1653.
28. Linnet K and Kondratovich M. A partly nonparametric approach for determination of the limit of detection. *Clin Chem* 2004; **50**: 732–740.
29. CLSI EP17-A2 (2012). *Evaluation of detection capability for clinical laboratory measurement procedures*; Approved Guideline – 2nd ed.
30. Fletcher JW, Djulbegovic B, Soares HP, et al. Recommendations on the use of 18F-FDG PET in oncology. *J Nucl Med* 2008; **49**: 480–508.
31. Weber WA. Assessing tumor response to therapy. *J Nucl Med* 2009; **50**: 1S–10S.
32. http://rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA_FDG-PET_Profile_v105_Publicly_Reviewed_Version_FINAL_11Dec2013.pdf.
33. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989; **8**: 431–440.