Statistical Methods in Medical Research
http://smm.sagepub.com/

Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment

David L Raunig, Lisa M McShane, Gene Pennello, Constantine Gatsonis, Paul L Carson, James T Voyvodic, Richard L Wahl, Brenda F Kurland, Adam J Schwarz, Mithat Gönen, Gudrun Zahlmann, Marina Kondratovich, Kevin O'Donnell, Nicholas Petrick, Patricia E Cole, Brian Garra, Daniel C Sullivan and QIBA Technical Performance Working Group

Stat Methods Med Res published online 11 June 2014
DOI: 10.1177/0962280214537344

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2014/05/30/0962280214537344

Published by:
SAGE
http://www.sagepublications.com

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–41

DOI: 10.1177/0962280214537344
smm.sagepub.com
SAGE

Quantitative imaging biomarkers: A review of statistical methods for technical performance assessment

David L Raunig,[1] Lisa M McShane,[2] Gene Pennello,[3] Constantine Gatsonis,[4] Paul L Carson,[5] James T Voyvodic,[6] Richard L Wahl,[7] Brenda F Kurland,[8] Adam J Schwarz,[9] Mithat Gönen,[10] Gudrun Zahlmann,[11] Marina Kondratovich,[3] Kevin O'Donnell,[12] Nicholas Petrick,[3] Patricia E Cole,[13] Brian Garra,[3] Daniel C Sullivan[14] and QIBA Technical Performance Working Group

Abstract

Technological developments and greater rigor in the quantitative measurement of biological features in medical images have given rise to an increased interest in using quantitative imaging biomarkers to measure changes in these features. Critical to the performance of a quantitative imaging biomarker in preclinical or clinical settings are three primary metrology areas of interest: measurement linearity and bias, repeatability, and the ability to consistently reproduce equivalent results when conditions change, as would be expected in any clinical trial. Unfortunately, performance studies to date differ greatly in designs, analysis method, and metrics used to assess a quantitative imaging biomarker for clinical use. It is therefore difficult or not possible to integrate results from different studies or to use reported results to design studies. The Radiological Society of North America and the Quantitative Imaging Biomarker Alliance with technical, radiological, and statistical experts developed a set of technical performance

[1]ICON Medical Imaging, Warrington, USA
[2]National Cancer Institute, Bethesda, USA
[3]Food and Drug Administration/CDRH, Silver Spring, USA
[4]Brown University, Providence, USA
[5]University of Michigan Health System, Ann Arbor, USA
[6]Duke University BIAC, Durham, USA
[7]Johns Hopkins Medical Institute, Baltimore, MD, USA
[8]University of Pittsburgh, Pittsburg, USA
[9]Eli Lilly and Co, Indianapolis, USA
[10]Memorial Sloan Kettering Cancer Center, New York, USA
[11]Hoffman-La Roche Ltd., Basel, CH
[12]Toshiba Medical Research Institute, Vernon Hills, USA
[13]Takaeda, Deerfield, USA
[14]Duke University School of Medicine, Durham, USA

Corresponding author:
David L Raunig, ICON Medical Imaging, 2800 Kelly Rd., Warrington, PA 18976, USA.
Email: David.Raunig@iconplc.com

analysis methods, metrics, and study designs that provide terminology, metrics, and methods consistent with widely accepted metrological standards. This document provides a consistent framework for the conduct and evaluation of quantitative imaging biomarker performance studies so that results from multiple studies can be compared, contrasted, or combined.

Keywords
quantitative imaging, imaging biomarkers, reliability, linearity, bias, precision, repeatability, reproducibility, agreement

1 Background

Most medical imaging procedures were originally developed and are used to detect and diagnose disease with only limited attempts to quantify what was seen. Technical improvements in instrumentation and software provide opportunities to quantify disease features including measurement of changes in physical or functional response.[1] Use of biomarkers is being pursued vigorously to better evaluate preclinical in vivo features[2,3] and to personalize clinical medicine[4] by quantitatively measuring morphological changes and defining function down to the cellular level. Development of quantitative imaging biomarkers (QIBs) includes an assessment of the performance of the QIB under study conditions. Each QIB is the end result of a defined image acquisition process of a quantifiable image from an evaluable subject, a computer processing algorithm to reconstruct that image, an automated or manual identification of the relevant regions, and usually an algorithm used to measure and report the QIB (Figure 1). QIBs may be categorized into five general types: structural, morphological, textural, functional, or physical property QIBs. The methods of measurement may be as simple as electronic or physical calipers (e.g. length) or may be a complex derived measurement of a functional parameter that describes the dynamic relationship of the image measurement to an external factor such as time or stimulus (e.g. apparent diffusion coefficient). Inherent to each QIB are factors that affect the measurement and consequently reduce

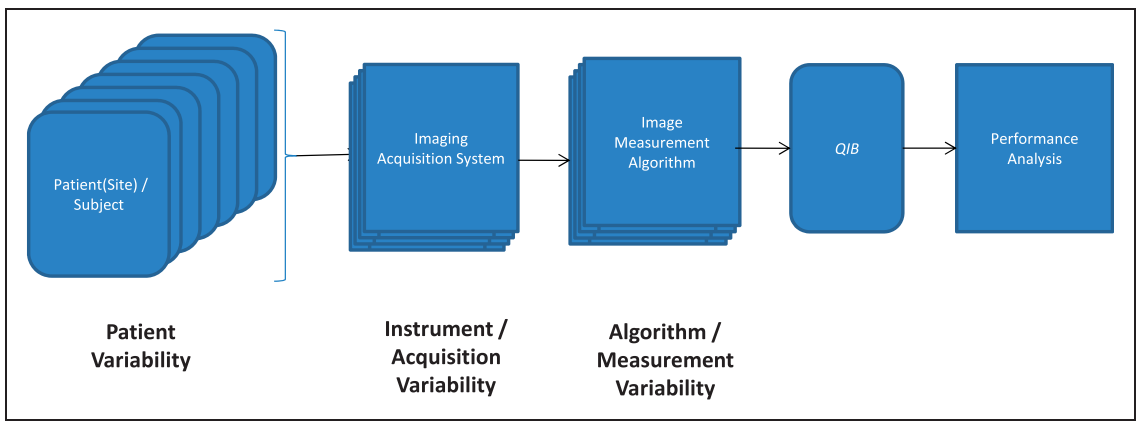

Figure 1. The quantitative image biomarker analytical process.

its reliability. Each QIB is typically evaluated for reliability under study conditions by measuring both the identifiable components of the random variability and any bias relative to a reference value.

Assessment of variability strives to evaluate each of the predicted major sources of variability against some criteria that depend on the application. However, performance evaluations are too often different from study to study resulting in confusing terminology and ad hoc performance metrics that sometimes are actually unique to the study. Consequently, multiple studies of the same or competing QIBs often cannot be compared against each other. In some cases, reported QIB evaluations involve novel reliability metrics constructed solely for use with that study. Inconsistent use of statistical metrics, use of different terms for the same metric, or use of the same term for different metrics can be confusing and even contradictory and may be inappropriate for adequate representation of the properties of the QIB.

Here we discuss how to subject QIBs to the same rigor as applied in the assessment of other quantitative clinical measurements to ensure the reliability of the image measurements to objectively evaluate ''normal processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.''[5,6]

Two summaries of QIB measurement variability are provided by the *repeatability* and the *reproducibility* of the QIB. Repeatability refers to variability of the QIB when repeated measurements are acquired on the same experimental unit under identical or nearly identical conditions. Thus, the concept of repeatability attempts to capture the ''pure'' measurement error of the QIB and can be assessed only approximately in practice. For example, repeatability can be assessed as the variance of the QIB measurement when the marker is obtained from repeated imaging of a nonbiological reference object (or phantom) under identical experimental conditions. However when the QIB is obtained by repeatedly imaging subjects and is based on actual biological features, the observed variance of the measurement also includes subject-related variability due to a variety of behavioral, physiological, psychological, and other factors that may have varied across individual measurements, even if the actual imaging acquisition process remained unchanged.

Reproducibility refers to variability in the QIB measurements associated with using the imaging instrument in real-world clinical settings which are subject to a variety of external factors that cannot all be tightly controlled. For QIBs to be robust enough to enable reliable medical decision-making or to use in controlled clinical trials, the QIB must be able to be reproducibly measured over a range of conditions. Reproducibility conditions may be identified and separated from measurement error. Some examples of different QIB reproducibility conditions are different scanner types, different patient types (e.g. ethnicities), different clinical sites, different image reviewers, and other conditions that may be specific to a study or to analysis software.

In addition to repeatability and reproducibility variability, the QIB must also be assessed for its ability to provide measurements that reliably represent the actual value of the targeted quantity. A difference between the expected value of the QIB and a known or accepted reference value (or measurand[7]) is the bias of the QIB and may be constant or nonconstant, varying by conditions or true measurand values. The reference values used to measure this bias can include knowledge of true measurements that may be available from physical or digital phantoms, for example, or from a ''gold standard'' that is very likely a QIB that has been widely used and accepted for that particular use. For the QIB to be used as a predictor of true biological feature change or difference, the QIB must predictably reflect the true and biologically relevant feature measurement (size, function, etc.) and any QIB bias should be quantified over the entire measurable range of values.

QIB technical performance for assessment of reliability is thus described by both the ability to represent the true or accepted reference measurement without bias and to do so with minimum variability.

This paper will present the issues and methodologies associated with assessment of the technical performance of a QIB. In Section 2 the motivation and objectives of technical performance assessment are discussed. Section 3 provides a brief summary of the metrology terminology used in this paper. Section 4 outlines the basic steps in designing a study to assess technical performance. Section 5 covers assessment of linearity and bias; Section 6 discusses repeatability and Section 7 discusses reproducibility. Section 8 summarizes the technical performance issues and Section 9 discusses future directions for developing QIB technical performance metrology concepts. References are found in Section. Section 10 is the appendix and contains data examples for repeatability and reproducibility analyses.

2 Motivation and objectives

This paper and the companion metrology papers[7–10] are motivated by a need to standardize and optimize metrological terminology and performance evaluations of QIBs and is distinguished from an evaluation of diagnostic, prognostic, or predictive performance for clinical utility. This effort is sponsored by the Radiological Society of North America QIBA activities to develop QIBs for use in assessing the extent and change of disease over time. QIBA is an initiative to advance quantitative imaging and the use of imaging biomarkers in preclinical studies, in clinical trials, and in clinical practice by engaging researchers, healthcare professionals, and industry with the specific mission to ''improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, patients and time.''[11]

The goal of this paper is to provide a framework for researchers developing QIBs (for example, in the context of a QIBA Profile Claim) and assessing their technical performance. To date, evaluation of technical performance methodologies for QIBs has been at the discretion of individual investigators. This is in contrast to development of laboratory measurements from blood or tissue, which are guided by established standards organizations such as the Clinical Laboratories Standards Institute (CLSI). To provide this framework in a consistent, rigorous, and broadly acceptable form, the QIBA Technical Performance Working Group has worked within metrology standards set by several international organizations including the Bureau of International Weights and Measures (BIPM) and the National Institutes of Standards and Technology. Metrology is defined by the BIPM as ''the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology.''[12] This paper addresses the technical performance of a QIB once a validated algorithm derives the measurement. Though technical performance is tied to clinical utility, the performance of the QIB to predict clinical outcome is not addressed here.

The QIBA Technology Performance Working Group was established to arrive at a reasonable consensus among clinical, technology, and statistical imaging experts to establish the metrology standards necessary to evaluate a QIB for use in a patient population. The objectives of the group's efforts are to achieve the following:

- Define technical performance metrics needed to measure and report technical performance of a QIB;
- Define the methodologies to arrive at those metrics; and

- Describe the study designs and considerations necessary to arrive at a meaningful and interpretable assessment of technical performance of a QIB.

These objectives directly relate to the development of a QIBA Claim by establishing the conditions under which the QIB can be reliably used. QIBA Profiles are syntheses of prior research and original research that propose standards for the use of specific QIBs in in vivo studies including animal studies, clinical trials, and medical decision making and are meant to specifically document the performance of a QIB and the conditions for which that performance applies.[13] Implementation follows directly from the design of the algorithm and technical performance assessments. Therefore, since study design is the underlying foundation for any technical performance study, every statistical concept or methodology discussed here will be directly tied to considerations surrounding study design.

Specific QIBA Profiles and finalization progress including public comment status may be found at the QIBA Wiki website.[14]

3 QIB technical performance terminology

General terminology for metrology concepts involved in the assessment of technical performance is found in the companion paper on QIB terminology.[7] A distinction is made here between quantitatively meaningful and clinically meaningful when defining the technical performance of the QIB. Quantitative technical performance drives the QIBA Profile that defines the claim whereas assessments of the clinical meaningfulness of a QIB must be presented to qualify the QIB for a claimed clinical use.

This paper will use the following terminology for the different values used in a technology performance assessment for experimental unit $i = 1 \ldots, n$, and k repeated measurements:

- QIB Measurement: $Y_i = \{y_{i1}, y_{i2}, \ldots y_{ik}\}$;
- Measurand (True Reference Value): $X_i = \{x_{i1}, x_{i2}, \ldots x_{ik}\}$;
- Imperfect Reference Value: $Z_i = \{z_{i1}, z_{i2}, \ldots z_{ik}\}$.

The definition of a QIB performance metric follows the basic premise that for a given set of measurements, the difference

$$\Delta yi(1,2) = yi1 - yi2 \tag{1}$$

or ratio

$$y_{i1}/y_{i2} = y_{j1}/y_{j2} \tag{2}$$

between two measurements has the same biological meaning for all i and j. This is distinguished from the difference between two ordinal scale scores that assign increasing integers to increasing levels of severity (e.g. 0 = none, 1 = slight, 2 = mild, 3 = moderate, 4 = severe). The relationships shown in equations (1) and (2) are true for ratio variables, which may be simply defined as when a value of zero (0) represents the absolute absence of the feature being measured. Ratio variables are further described in detail in Kessler.[7]

The three metrology areas that most directly address the question of technical performance can be found in detail in International Organization for Standardization (ISO) guidelines[12] as well as QIB specific guidelines described in Kessler.[7] We offer the following as practical clarification.

Bias/Agreement/Linearity: The ability of the QIB to unambiguously reflect a change in the disease state as represented by an adequate change in the QIB. In many QIB cases, the truth may not be known or even technically feasible to measure and the imaging metric may not have a linear response against a reference measurement.
Repeatability: The ability of the QIB to repeatedly measure the same feature under identical or near-identical conditions. These studies are often referred to as test–retest, scan–rescan, or ''coffee-break'' experiments.
Reproducibility: The reliability of the QIB measuring system in different conditions that might be expected in a preclinical study or clinical trial or in clinical practice (e.g. multiple sites, etc.). The technical assessment of algorithm reproducibility performance is specifically dealt with in Obuchowski.[8]
Measurand/Reference: A measurand is the true value of the quantity intended to be measured; a reference is the true or accepted value of the measurand; a theoretical or established value based on scientific principles; an assigned value based on experimental work of some national or international organization; or a consensus value based on collaborative experimental work under the auspices of a scientific or engineering group (ISO 5725-1). To avoid confusion, the term ''measurand'' will be used when referring to a prescribed true value and the term ''reference'' as the actual value used as a comparator to the QIB measurement.
Repeated measures/Replicates: A measure is repeated if it is independently acquired on the same experimental unit and, if no change in the measurement is expected; the repeated measurements most often represent the total measurement error. Replicate measurements are obtained from different experimental units for the same or equivalent measurand and represent the between-subject variability.

4 Technical performance assessment design

The following steps are recommended when designing a performance analysis of a QIB. These steps reflect fundamental statistical experimental design principles and permit appropriate collection of the data necessary to address the different technical performance objectives defined in the preceding sections. Not all steps may be applicable in every technical performance study, but they are presented here as a general guideline for most QIB reliability assessments.

Step 1: Define the QIB and its relationship to the measurand

Define the measurement to be acquired, the quantity to be measured (measurand), and the expected relationship between the QIB and the measurand. For example, ''FDG uptake in gastric lesions will be measured as the Standard Uptake Value adjusted by lean body mass (SUVlbm) and is a measure of the integrated metabolic rate within a specified region of interest (ROI).''

Step 2: Define the study claim or question to be addressed in the analysis

State the study claim either in the form of a hypothesis, general question, or statement of bounds on technical performance. Study hypotheses do not have to be in the form of a statistical hypothesis. For example, a claim may begin as a statement such as ''FDG uptake in gastric lesions as measured by SUVlbm will increase proportionally with concentration of FDG and will vary by less than 20%'' is an acceptable hypothesis. It is then translated to a more specific Profile claim that defines

conditions under which the claim is expected to be valid, including the particular patient preparation and image acquisition and analysis conditions to be investigated.

Define the statistical hypotheses, if applicable, including criteria for accepting or rejecting the hypotheses. While it is perfectly valid to formulate the problem as one of testing a statistical hypothesis, most QIB technical performance studies will focus on estimation and bounds on performance.

Define sub-strata that are identified within the claim that will be either modeled or tested in the study. For example, strata may include patient demographics, sites, or regions for testing, reader qualification, etc. depending on the applicability to the study question or limitations of the study.

Step 3: Define the experimental unit

Whether the patient/subject is the experimental unit depends on the study question. For example, tumor imaging studies may have the lesion or lesion nested within patient as the measurement unit of interest, while bone mineral density measurements of the hip have the patient as the experimental unit, though imaging just the hip or a specific region of the hip. The selection is consequential for statistical analysis and eventual inference and needs to be defined with care and synchronized to the claim and Profile specifications.

Step 4: Define the measures of variability to be estimated

The statistical measures of performance, including variability, bias, and linearity, must not only be appropriate but also useful when planning a future study that will use the QIB as a study aid or study endpoint. All QIB metrics should also include confidence bounds provided for any estimated parameters.

Step 5: Specify the elements of the statistical design

Elements of the statistical design include sample size, number of reviewers, technologists, clinicians, patient/subject population, choice of reference measurement method (for bias and linearity studies), range of measurand values, 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. Sample size should be justified by a statement on study power or precision of performance metric estimate, as applicable. When sample size is driven more by convenience (or budget) instead of statistical power or precision of estimation, this should be stated and realistic power or confidence interval (CI) width and coverage estimates should still be provided.[15] An adequate description of the study design provides the necessary context for the eventual use of the performance metrics, insuring that it will be possible to combine, compare, or contrast results from different studies.

Step 6: Determine the data requirements

Study data generally fall into two categories: prospectively and retrospectively collected data. The type and amount of data collected is often determined by the constraints of the study. While circumstances may change the final conclusions based on the data that are actually possible to collect, both retrospective and prospective QIB performance analyses should start by predetermining the desired data to set up inclusion and exclusion criteria. An example is the use

of selected archived imaging data for measurement and analysis. The number and types of acceptable images may be (and often are) different than what is available.

Each of the parameters desired as an output from the designed study must consider not only the number of experimental subjects (or experimental units) but also the data range, distribution of the data, number of repeat measurements for repeatability, and the number of conditions under which reproducibility will (or can) be assessed.

Step 7: Statistical analysis

Two common elements of a statistical plan for QIB studies are choice of random or fixed effects to represent various factors affecting variability, and specification of stratification/blocking factors, if any. Random effects models may also include the use of random slopes/random intercepts, e.g. for study subjects followed longitudinally. Repeatability studies that involve patients or subjects will very often consider the subject/patient to be a random factor. Reproducibility studies will likely include fixed effects when the possible levels of a reproducibility factor are few and not randomly chosen. Smaller reproducibility sets, such as measurement algorithms or tertiary care facilities in a defined geographic region are often not chosen at random and usually will be treated as fixed effects. A more complete discussion may be found in Kenward and Roger.[16] In general, random effects randomly choose from a large population of levels of a factor while fixed effects do not randomly select, though the distinction is not always clearly defined in practice.

Testing of a null hypothesis and estimation of specified performance metrics will be the two most common study goals. This will follow directly from whether the study question is formulated as a statistical test, as an estimation task or both. Hypothesis tests have, for various biomarkers, included superiority, noninferiority, and equivalence alternative hypotheses, and the appropriate choice is dependent on the QIB claim. A detailed examination of QIB hypothesis tests is found in the QIB algorithm comparison companion paper.[8]

One common theme throughout this discussion is the hierarchical progression of study planning for each QIB. For example, the definition of the study hypothesis directly dictates many of the downstream decisions. For this reason it is very important to consider all of the issues outlined in this section during the study design phase and to iterate through the process to ensure the claims are consistent with the study design and thus to provide the greatest chance of drawing useful conclusions from the study.

5 Bias and linearity

An ideal QIB provides an unbiased estimate of a true characteristic, or a value derived from it, over the entire range of expected values (the measuring interval) defined in the claim. For example, when measuring the volume of a solid tumor, the measured volume should, within random measurement error, represent the actual volume, or measurand, of the lesion defined, regardless of the tumor's shape, size, or composition within the context of the claim. An example of a well-defined reference is a validated phantom.

Integral to the calculation of QIB bias is defining the functional relationship of the QIB to the measurand, when available, or to an estimate of the measurand for imperfect references. Represent this relationship by

$$E(Y|X = x) = f(x)$$

If Y exhibits linearity with X, then a change in x will, on average, be reflected as a *proportional* change in Y. This is stronger than saying that a change in X will translate to a change in Y. For example, rate of change with x for a polynomial function of order greater than 1 varies depending on the true value of x; this means that the same change in x will not always correspond to the same proportional change in $E(Y|X=x)=f(x)$. Furthermore, these higher order functional relationships are not guaranteed to be monotonic.

The ideal relationship between Y and the truth, X, is linear with slope equal to 1. However, imaging biomarkers, especially in an early stage of development—and calibration—may not always yield the identity slope. Imperfect references will often not yield slopes of 1.0 although a *proportional* relationship (i.e. linearity) is likely to hold approximately within a specified range of measurand values. The proportional relationship between the measurement and the measurand defines not only the linear relationship of the QIB and truth but also provides information to the end user about the sensitivity of the QIB to measure a change in measurand value. Examples of several possible relationships between the QIB and the measurand are shown in Figure 2. For the same variance, a slope less than 1 indicates less ability to detect change and conversely, a slope greater than 1 indicates greater sensitivity to changing values. Relationships with curvature have a nonconstant sensitivity to change over the range of true values. Actual sensitivity to change also depends on variability and will be discussed in later sections.

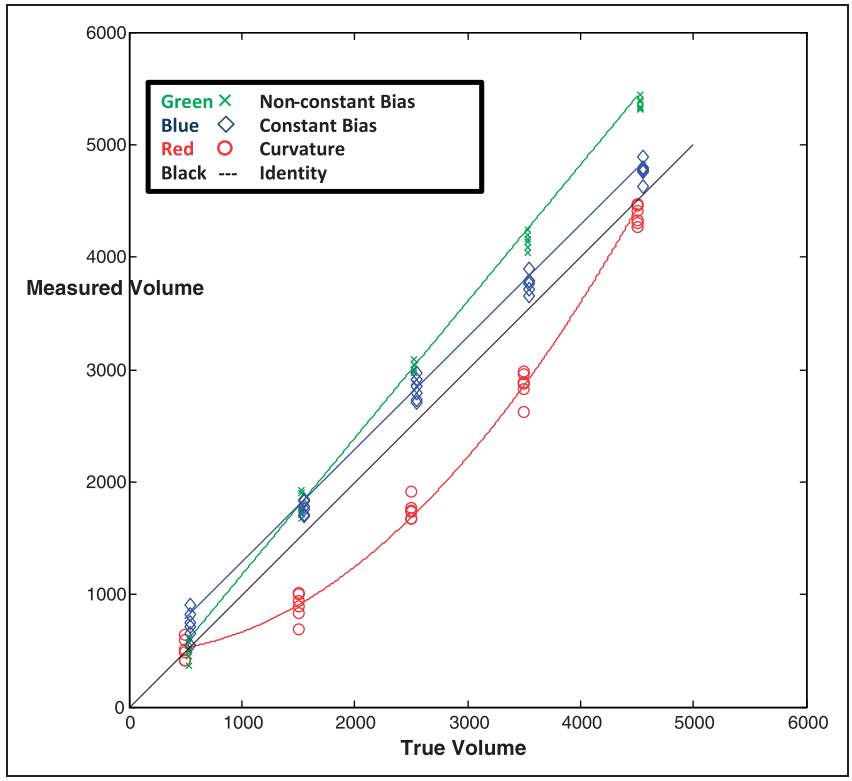

Figure 2. Plot of phantom volumes versus measured volumes with identity line (dashed black). Values were taken from QIBA 3A[17] data on phantom volumes and augmented with simulated data for illustration purposes.

Table 1. Regression results of constant bias for measurement versus phantom data.

Term	Estimate	Std Error	t Ratio	Prob > |t|
Intercept	332.087	49.228	−6.75	<.0001
Log(Volume (true))	1.034369	0.007306	141.57	<.0001

If the relationship is well approximated by a line and the estimated slope is close to 1, then the intercept of the estimated line is an estimate of the bias. The derived measurement of change, such as the measured change in tumor volume over time can then be effectively used as an estimate of the true change over time. In a clinical trial, the derived measurement can be used to compare two treatment arms for efficacy. If the slope is much different than 1 then bias is a function of the measurand and change measurements will be a function of the values themselves and the Profile claim may need to limit the use of the QIB.

The results from the regression applied to the subset of data in Figure 2 exhibiting constant bias are presented in Table 1. Determination of statistical significance may be highly influenced by the total number of samples, the range of the data, and high leverage values. Parameter estimates should also be interpreted in the context of how the QIB measurements would be used in practice. In the case presented in Table 1, the slope is significantly different than 1 ($p < 0.001$) but this minor deviation might not have important implications in practice. The technical performance assessment of the QIB measurement system should address both bias and linearity. Both performance aspects affect the application of the QIB to clinical decisions that rely on an absolute assessment of the true value of a measurand, Nonconstant bias, including nonlinearity as a special case, affects assessment of a change in the measurand. Evaluation of the impact of intercepts and slopes should not rely solely on the p-values but also on the actual estimates and their standard errors as well as the clinically relevant range of the measurand and the variability in QIB measurements.

5.1 Bias estimation

QIB bias, $\delta(x) = E(Y|x) - x$, is defined as the difference between the expected value of the measured variable Y and the true value x. If within a specified range, the QIB is can be expressed as a linear function of the true measurand plus an additive error

$$y = (y_0 + \epsilon) + x(1 + \beta) \tag{3}$$

where ε is the random measurement error with mean 0, then bias can be defined as a function of the true value x as

$$\delta(x) = y_0 + x\beta \tag{4}$$

When the truth is available through the use of digital or physical phantoms or a gold standard, the measurand is assumed to be known without error and bias is true bias. However, in practice, a true reference will rarely be known in QIB studies and a pragmatic approach using an agreed upon reference standard method must be taken knowing that only relative bias or linearity is being assessed. Therefore, it is very likely that the measurand value is determined by an

imperfect reference, z, which is measured with some error, typically known or previously estimated

$$y = (y_0 + \epsilon) + (z + v)(1 + \beta) \tag{5}$$

where v is zero mean random measurement error in z. The relative bias is defined similarly as a function of the expected value of the reference standard

$$\delta(z) = y_0 - z\beta \tag{6}$$

Reference standards such as validated phantoms may be measured with negligible error, but it may often be the case that the QIB is being considered to replace a less-reliable reference. It is important to consider the imprecision of the reference and its effect on the association of the QIB to the reference. Repeated reference measurements and error-in-variables methods are needed in these cases.

For purposes of discussion, bias is considered here to be an additive effect where ratio QIB bias may be treated as additive under a log transformation. The bias can be classified as either constant or nonconstant:

- A *constant bias* is present if the QIB measurement exhibits a fixed deviation from the measurand and is not independent of the true values. A constant bias is indicated in Figure 2 (blue diamond line) with a fitted line that is parallel to the identity line and the intercept is an estimate of the bias.
- A *nonconstant bias* is a difference from the true value that is dependent on the measurand in ways that are not always able to be determined, but would be consistently observed to have the same functional relationship to the true value in identical experiments. *Nonconstant* biases are particularly important to characterize carefully when there is interest in assessing changes in QIBs because the different biases will not "cancel out" in calculation of a change. Any curvilinear relationship as well as a linear relationship that is not parallel to the identity line will have a *nonconstant* bias.

The use of the terms systematic and nonsystematic bias is present throughout the literature and generally used to describe constant and nonconstant bias, respectively. However, the terms systematic and nonsystematic are not always used this way and so can present some confusion. Therefore, we will characterize bias as constant or nonconstant for the rest of the paper.

5.1.1 Data transformations

It may be possible to use a data transformation to stabilize nonconstant bias and reduce or eliminate the heteroscedastic nature of the QIB. Transformations such as the log-transformation may also typically dictate the form of the bias. For example, a bias that is proportional to the true value with no constant offset component becomes constant in x when a log-transformation is applied.

Very often the variance of a QIB increases with increasing measurand values and an estimate of overall variance may not be appropriate for the QIB values observed. A data transformation such as the log-transform may therefore be useful when estimating the variance as a function of the estimate of the measurand. For example, the variance of the log-transformed data may be back transformed to an estimate of the coefficient of variation where the standard deviation (SD) is directly proportional to the mean.

Appropriate transforms may be chosen from a priori knowledge of the measurand or may use power transform statistical methods, such as the Box-Cox transform to help with linearity and alleviate heteroscedasticity.[18] Some consideration should be given, however, to interpretation of the resulting measurements when transforming the QIB. Transformations may also introduce a nonconstant bias and should be approached cautiously with an assessment of the effect of the transform within the measuring interval. A common example is the custom of adding 1 to the values of data that include 0 in the measurement prior to taking the logarithm of the number. This practice has a large effect on small measurement values and should not generally be used. A practical rule for transforms is that raw measurements that have a nonzero intercept for a measurand value of 0 will not be transformable unless restricted over a limited range of measurand values.

An additional common misuse of transformations is to reduce the presence of outliers without regard to the actual underlying QIB distribution. Rank preserving data transforms are sometimes necessary in order to comply with the assumptions of normality in statistical data analyses and should not be used to hide or obscure values for convenience or to redefine a value as a nonoutlier. The treatment of outliers is addressed in more detail in Section 5.1.3.

5.1.2 Bias estimation considerations

Choosing the number, range, and distribution of measurands. The sample size, range, and distribution of the measurand will be a critical design factor and will affect the estimation of bias throughout the range of expected values. These should adequately cover the claimed *measuring interval* of the imaging measurement system. The term *measuring interval* is defined in Kessler[7] as the range of the measurand in which an incremental change in the value of the measurand should result in a change in the QIB measurement. The lower and upper endpoints of the *measuring interval* are the lower limit of quantitation and the upper limit of quantitation and can be described as the limits within which the QIB has acceptable bias and precision. Acceptable limits will be determined by the context of use. Figure 3 is a typical depiction of lower and upper limits of quantitation. Lower QIB measurement levels that approach the background noise may see an increase in variability while large QIB measurements may be physically limited and be characterized by low variability, also known as saturation bias. The measuring interval is characterized here as a region of relatively linear relationship to the measurand.

When possible, bias assessments should use at least two and preferably three or more experimental replicates carried out for each of several settings of "truth" as measured according to a standard reference. This might include phantoms with similar measurands, phantoms placed in different locations or, in the case of patient-obtained data, values that are reasonably close. Replicates are different than repeated measures which are used to define variability within a single experimental unit. Replicate values are best defined when determining the experimental unit.

Various numbers of measurand levels have been proposed, but sources for laboratory assays recommend at least five to seven levels.[19] In phantom studies these should be roughly equally spaced over the measuring interval to minimize the variance of the regressor parameter estimates and to identify departures from linearity throughout the entire measuring interval. At least three replicates, as opposed to repeated measures of the same replicate, are recommended at each phantom level to provide sufficient degrees of freedom for stable estimates of precision of the estimates.[20]

Replicates may use different phantoms if possible and may also use phantoms in different orientations, configurations, or conditions. 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

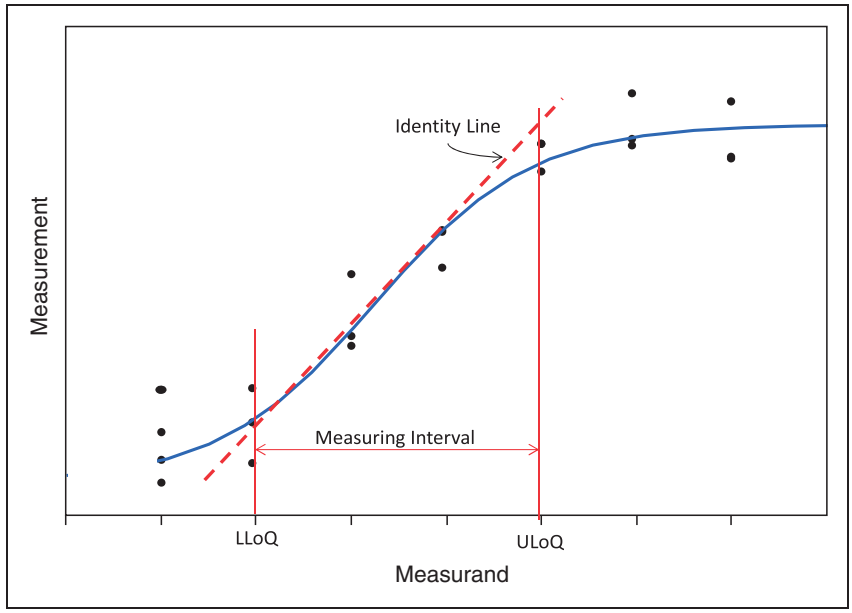

Figure 3. Measuring interval analysis of bias. Note that the values measured beyond the measuring interval are necessary to define the interval.

or age) that are related to the QIB. When there is reason to believe that bias would change more rapidly in certain regions than others a higher concentration of measurand values in those regions of expected rapid change is necessary to complete the performance assessment.

Plotting the data. A bias assessment typically begins with visual assessment of a plot of the measured values (y) against the reference values (x) (Figure 2). Individual replicate values should be plotted in addition to the means that summarize the replicates at each x value. Bias at each reference value can be assessed by comparing the mean of the replicate measurements of y to the reference value or, in the event that replicate measurements are not available, the prediction regression estimate. The bias estimates should also include confidence bounds that reflect the variability of the replicates or of the regression prediction. Their inclusion is necessary to provide the ability to assess the quality of the bias estimate and indicate a need to augment the bias data or that variability of the QIB is the priority technical performance issue to be addressed prior to an assessment of agreement to the measurand.

When bias is defined by the true values, X, then bias may also be plotted against the true values to determine the nature of the bias. An example is shown in Figure 4 where the slope is not equal to 1.0 and the bias increases with increasing true values.

5.1.3 Outlier analyses

Any assessment of technical reliability of a QIB must also include an assessment of the prevalence and impact of outliers which could preclude any further testing. Estimates of bias from either replicate means or regression prediction estimates can be greatly influenced by isolated aberrant points that may be due to chance occurrence of extreme values, procedural errors in data acquisition and recording, an unexpected mixture of populations, or an unexpected test deviation. Methods to

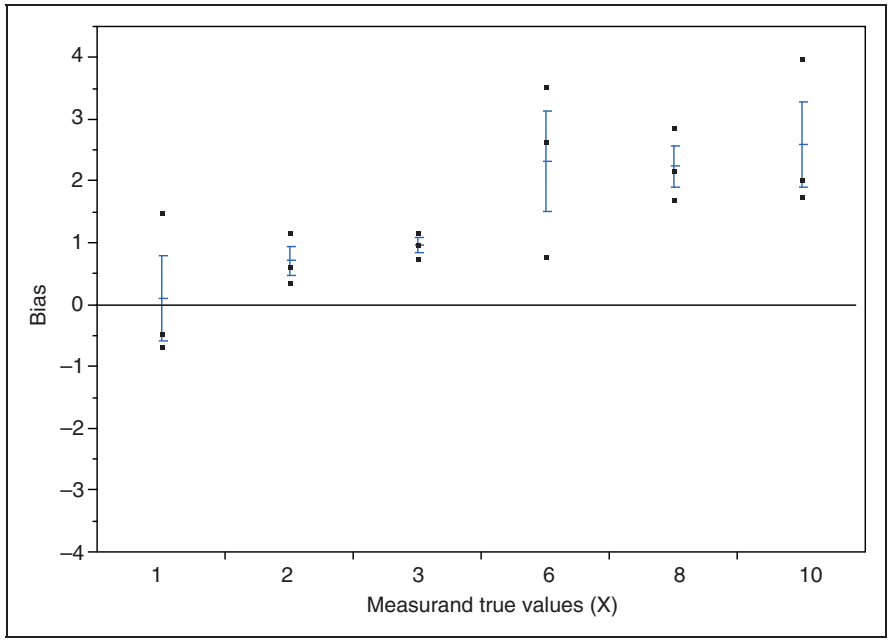

Figure 4. Bias plotted against Truth for the relationship $Y = 1.4 * X + \varepsilon$ with $\sigma^2 = 1.0$.

distinguish between chance occurrence and erroneous values include both graphical examination and statistical outlier detection tests. There is an abundance of outlier detection methods in the literature as well as substantial developments in outlier discrimination. ISO as well as the CLSI have compiled a set of commonly available plots and tests for single and multiple outliers.[21–23] While multivariate QIBs are under development by several researchers, this section describes univariate outlier detection only. However, multivariate outlier detection methods are available and may be more appropriate when employing multiple endpoints. A general review of multivariate outlier detection methods may be found in Penny and Jolliffe.[24]

A general methodology for outlier analysis and reporting is described here as a basic set of steps consistent with metrological concepts that are recommended when describing technical performance of a QIB.

(1) Screening: Visual inspection of the plotted data
  (a) Box–whisker plots with outliers plotted.
    (i) IQR = Q3–Q1, where Q3 is the third quartile, Q1 is the first quartile
    (ii) $QIB_{Outlier} < Q1 - 1.5 * IQR$ or $QIB_{Outlier} > Q3 + 1.5 * IQR$
    (iii) May be generalized for nonsymmetric distributions about the median[21]
  (b) Distribution Plots: Histogram, dot, stem-leaf, probability plots, etc.
(2) Statistical tests
  (a) Normal distribution
    (i) Generalized Extreme Studentized Deviate[25] is generalization of Grubbs test that also tests for more than one outlier that preserves the Type I error for l outliers of a predetermined maximum of m outliers where $1 \leq l \leq m$

(ii) Upper/lower fourths: A modified IQR test for normal data described in detail in ISO 16269-4.[21]

(b) Nonnormal distributions

(i) Lognormal: Logtransform data to a normal distribution

(ii) Exponential distribution: Greenwood test for existence and sequential identification of outlier observation

(c) Variance uutliers

(i) Cochran's test assesses the maximum variance against a critical value as an outlier when all SDs are estimated from identically sized cohorts found in many statistical analysis packages and easily calculated using simple spreadsheet functions as

$$C = \frac{s^2_{\max}}{\sum_{i=1}^{p} s_i^2}$$

where s_i^2 is the variance for each variance strata and $s^2_{\max}$ is the maximum variance. Critical values for C are available in many statistical texts including Snedecor and Cochran.[26]

(3) Identify procedural errors or QIB quality issues that may have led to the identified outlier. Data transcription or data translation errors that are not inherent in the acquisition of the QIB should be removed from the analysis and reported.

(4) Evaluate impact of outliers that could not be attributed to procedural or recording errors and therefore cannot be ruled out as rare but valid QIB observations.

(a) Descriptive statistics for bias and variance.

(b) Regression parameters[21,27]

(i) Difference in Fits (DFFITS)

(ii) Cook's distance

Point-wise deletion of outliers should be used with great caution when using an imperfect reference (i.e. patient data) since in most cases it will be difficult to ascertain the cause and the outlier may actually reflect the reliability of the QIB.

Bias estimation methods. If the degree of bias and variability appear approximately constant, over the measuring interval, then any bias can be estimated as the simple arithmetic average of the individual differences between measured value and reference value. If the bias is relatively constant over the measuring interval but the variability appears to depend on the true value, then a weighted mean, inversely weighted by the sample variances computed at each reference value, may be used. However, technical performance studies are very often conducted with only a few replicates at each measurand and variance estimates themselves will be unreliable. Therefore, when the variances are estimated from a small number of replicates the method of weighted means should not be used. Instead, a variance stabilizing transform is preferred. Tests for equal variances across k samples may be done by visual inspection or by formal hypothesis tests such as Bartlett[26] or Levene[28] tests, though tests of equal variance are sensitive to the number of replicates and generally are low powered and thus should be interpreted with caution.

Often a nonconstant bias is induced by incomplete calibration. Root causes may include an image processing algorithm that assumes an inappropriate form of calibration curve (e.g. linear) or failure to apply an appropriate transformation (e.g. logarithmic) to the raw signal. Nonconstant bias should first be addressed by calibration adjustments to make the bias closer to constant. Transformations to adjust bias may be difficult to interpret or not be universally applicable and

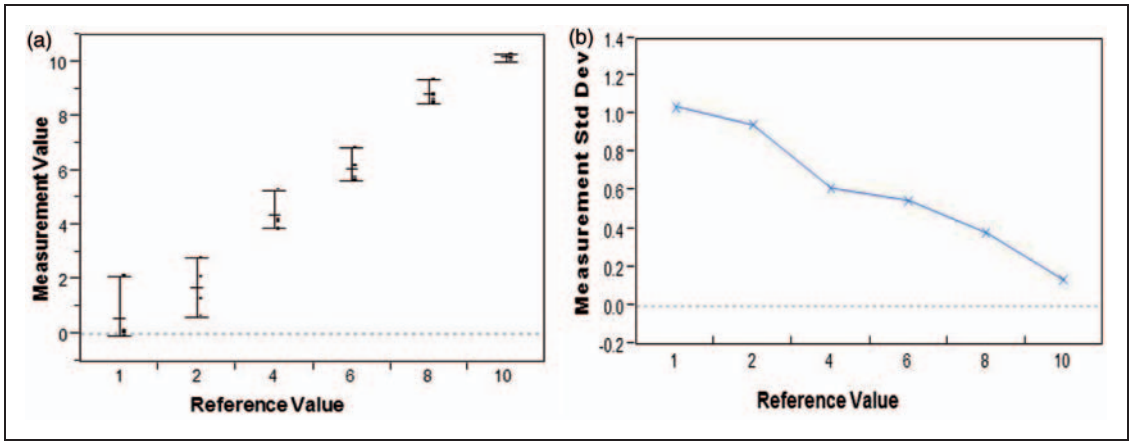

Figure 5. QIB measurement variability plots for comparison to the reference values.

so are not recommended as a first choice. If adjustments do not satisfactorily resolve the nonconstant bias, then the bias needs to be carefully characterized as a function of the true reference value (i.e. the measurand) and any other explanatory factor variables. If nonconstant variance is present in addition to nonconstant bias then sophisticated statistical methods may be required to reliably estimate the nature of the bias such as described by Carroll and Ruppert.[29]

Plots of variance with respect to the reference values (or means as appropriate) are valuable methods to assess QIB variance. Two plot options are shown in Figure 5.

Coverage Probability (CP). CP is useful in describing the impact of QIB bias and variability to an acceptable difference (d). CP(d) is the probability that the absolute difference between a measurement and the true value of the quantity is less than an acceptable difference d. Thus, CP(d) is more directed to describing the impact of the variability of the QIB about the measurand than a naïve reporting of prediction intervals and so provides the QIB user with a practical sense of the performance of the QIB with respect to the standards needed for a particular study. Technical performance of QIBs for bias should include a measure of CP(d) for a specified clinically or scientifically acceptable value of d. A detailed description of the application of CP(d) is found in Barnhart and Barboriak[30] and Barnhart et al.[15]

5.2 Linearity estimation

Bias is only independent of linearity when the measurand is zero (0). Image acquisition fundamentals and noise inherent in the image acquisition can provide a background signal that results in a QIB that is always positive and not equal to zero, or a bias at the intercept. Linear response from the intercept will result in an increasing, nonconstant bias if the slope is not exactly equal to 1.0. Nonlinear response of the QIB to the measurand will have a more complex relationship to the measurand. Therefore, the assessment of linearity is directly linked to the assessment of bias and both should always be presented when assessing either for technical performance.

5.2.1 Measuring interval and the range of linearity

Linearity may hold over the entire range of the measuring interval or only over some subinterval. The *linear range* will be defined here as the range of the measurand for which the imaging system

produces results that are within an acceptable tolerance of a linear relationship between true and measured values.

Assessment of linearity should follow an approach similar to that for assessment of bias. In addition to the steps of variability and outlier assessment described earlier for bias assessment, the examination should include an assessment of whether the measured values (y) are linearly related to the measurand (x), based initially on a simple scatterplot. The linear relationship might hold over the entire measuring interval or it might be piecewise linear and restricted to one or more subintervals. Determination of the linear range can be somewhat subjective, and it may be necessary to iterate between tentative specification of the linear range and formal testing for linearity within the specified range. A cross-validation of the results should be conducted for a final linearity assessment in an independent experiment to avoid the possibility of exaggerating the adherence to linearity within a data-derived linear range.

5.2.2 Linear relationship to truth

Many studies to assess linearity that are phantom based will compare the linear slope to the identity slope of 1. Since the comparison is to the truth, nonlinearities or slopes other than 1 are an indication of QIB discrepancies and should be investigated further for calibration errors. When comparing the QIB to phantom or truth data, 9–11 levels of the true value are consistent with recommendations that have been proposed for linearity assessment for laboratory assays[19] but the optimum number will ultimately depend on the specific characteristics of the QIB. If linearity is being established de novo, more levels should be considered, whereas around nine levels might be sufficient to confirm linearity of an imaging measurement system for which linearity had been previously established but measurement conditions changed such as updates, site change, etc. The measurand levels should typically be roughly equally spaced over the measuring interval but it may be necessary to concentrate additional observations 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.

An example of replicate measurements of linearity compared to a known or precisely determined reference is derived from the QIBA 3A challenge[31] and found in Section 10.1. There are five replicates for each measurand and a total of 31 different measurand levels. The results are plotted and summarized here. In this example, there is no evidence of bias since the intercept is not significantly different from 0. Also, it is important to note that the hypothesis test for the slope, shown as the term ''measurand'' is for a comparison of the slope to 0 and not 1.

5.2.3 Linear relationship to a reference

Establishing a linear relationship to a measurand will depend on the relationship of the reference value to the truth, which is not often known. The linear relationship to the reference is primarily concerned with deviations from linearity (piecewise linearity or curvature) and should not typically compare the slope(s) to a superiority threshold. While slopes relative to the reference can be an indication of increased sensitivity to changes in the QIB, the same can result if the QIB is simply a scalar multiple of the reference.

5.2.4 Linear estimations methods

Testing for curvature. A formal assessment of linearity using sequential tests of polynomial fits to the plot of mean QIBs versus reference can also be conducted. This is an approach recommended in CLSI EP06-A for testing linearity for laboratory assays.[19] The assessment begins with fitting a third

degree polynomial to the mean QIBs as a function of reference values, followed by testing the coefficient of the third-order term to determine if it is significantly different from zero. If this test result is statistically significant, then the hypothesis of linearity is rejected in favor of evidence for curvature. If the test fails to reject, a second–order polynomial is then fit to the data and the coefficient on the second-order term is tested. If the coefficient on the second-order term is statistically significantly different from zero, the hypothesis of linearity is rejected. If the test fails to reject, it is concluded that the hypothesis of linearity cannot be rejected. A finding of a significant polynomial degree of two or greater does not imply that the true underlying relationship is best described by a polynomial but only as evidence for local curvature. Conversely, the failure to reject curvature by the higher order coefficients may be due to the small number of samples and the low power to declare significance. In any test of bias and linearity, the sample size and measurand distribution must also be included in the description of the results.

A test of the slope in the linear regression would then follow to establish whether there is a linear trend or no trend in the data. Finally, there must be an assessment of the variability of the measurements around the line to establish whether the imaging system produces results that are reliably within an acceptable tolerance of a linear relationship between true and measured values with acceptability standards defined by the profile. It is not necessary, or even advisable, to fit a regression line with intercept forced to zero. If there is any curvature near the lower end of the range, a better overall approximation to the relationship would be obtained allowing estimation of a nonzero intercept. Even when the true intercept is zero, a single degree of freedom is used to include the intercept with only a small effect on precision of the regression parameter estimates.

Choosing the number, range, and distribution of measurands. The recommendation to collapse the replicates at each reference value to a single mean rather than use the replicates to obtain an estimate of pure error in the regression analysis is often made and debated. This recommendation is motivated by the desire to separate the impact of precision from the underlying functional relationship and may be important when the presence of outliers may have too much weight in the estimation of the parameters. While the linear relationship may be strengthened by summarizing replicates by the mean or median, it will be important to assess linearity in the presence of replicate variability. Additionally, for well-behaved normally distributed replicates, the loss of error degrees of freedom when using collapsed data will result in confidence intervals of the parameter estimates that will be very nearly identical to those from the raw, un-collapsed analysis. Therefore, all of the replicates at each measurand should be used as the default. Collapsing or summarizing replicates may also be done as an additional analysis. Analyses for replicate data that are not normally distributed are covered in the next section on methods.

Linearity methods. Other modifications to the regression approach might need to be considered. When nonlinear relationships are identified, it should be investigated whether data transformations can be applied to achieve approximate linearity. Additionally, it may be useful to apply transformations to stabilize variance, or it may be necessary to use weighted regression techniques to properly account for unequal variance. Use of these approaches may require substantial statistical expertise and specialized software.[29]

Least squares regression is based on the premise that the predictor variable (regressor or x-variable) is known or determined without error. Examples are phantom measurements where the values are known or measured with a high degree of precision (i.e. very low variance) and measurands that are determined prior to the QIB measurement). Therefore, when errors exist for both the reference and measurement, then the result of simple regression may be a biased

underestimation of both the slope and residual error.[27] Several alternatives to simple linear regression are available to address this issue and should be considered for their applicability to the specific QIB. Included in these methods are schemes that allow the imperfect reference to be considered as a ''target value'' where the QIB is then measured and the method of least squares may be used.[26,32] This method may have particular appeal when using patient-measured reference data and may simplify the analysis under those assumptions. A random slopes/random intercept model may be considered when different measurand values are acquired within the same experimental unit. Other ''error in variables'' methods provide unbiased estimates of the regression coefficients; however, these methods require some knowledge of the variance of both independent and dependent variables which are not always possible. An overall review of these methods, including Deming, Passing/Bablok, orthogonal regression and of ordinary least squares methods is provided by Haekel et al.[33]

Lack of a standard reference can be a challenge for linearity assessment, as for bias assessment, but linearity can sometimes still be assessed by direct evaluation of proportionality. The idea is analogous to use of dilution series in laboratory assays. If reference standards can be constructed to satisfy proportionality requirements, or if another imaging measurement method is available for which preservation of proportionality has already been established, then these can be used as the reference standard against which proportionality of the new imaging measurement method can be compared. If an appropriate proportionality reference series can be constructed, the proportionality of the new imaging assessments would be compared directly to the known proportions (perhaps on a log-transformed scale to allow for assessment of linear relationship). It is important in such evaluations to include proportions that cover the entire measuring interval. The variance of replicate measures, and hence the variances of the proportions might also depend on the absolute levels of the primary measurements. All of these factors must be considered and stated in the statistical analysis of the proportionality series.

6 Repeatability

Repeatability and reproducibility are commonly confused concepts with one term often substituted, incorrectly, for the other. In this section, we restate the metrology definitions for repeatability more completely defined by Kessler[7] and the conditions necessary to achieve the quantitative measurements required to assess QIB repeatability.

In short, repeatability studies encompass test–retest studies that have been applied to phantoms and patients to assess within-patient variability as a proportion of total variability.

Repeatability is the magnitude of measurement error under a set of repeatable conditions.

Repeatable conditions involve the same measurement procedure, same measuring system, same operators, same operating conditions, same location, and (most importantly) the same or equivalent experimental units over a reasonably short interval.

The requirements of the imaging system will define the repeatability conditions and the minimum time interval between repeats. Shorter time intervals minimize the effects of other variance components; however, factors such as scanning period, radiation, contrast washout, radioactive half-life, subject fatigue, etc. place restrictions on the conditions of repeat scans. For example, sequential repeats within the same scanning session capture effects due to scanner adjustments and image noise that defines the minimum detectable signal above a base level of noise. Additionally, if the subject is repositioned within the scanner, additional variability due to even slight differences in subject positioning will also be captured. Natural biological decorrelation as time progresses (autoregressive correlation, separate from disease progression) will further add to

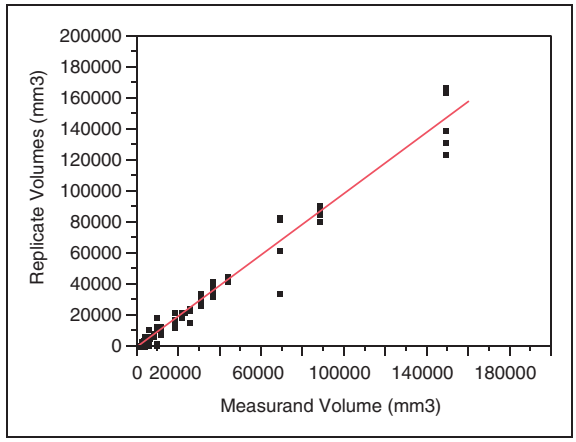

Figure 6. Linearity example derived from the QIBA 3A challenge.

variation, as may scanner performance drift and disease variability. All of these sources of variability are relevant to assessing repeatability of a QIB.

6.1 Repeatability model definition

Before we address statistical issues related to repeatability we need to translate the definition we have been working with so far into suitable notation. Suppose two measurements under repeatability conditions are indexed as y_{ij}, where i denotes the subject ($i = 1, \ldots n$) and j indexes the measurements under the repeatability conditions ($j = 1, 2 .. k$). There are two sources of variability in y_{ij}. Within-subject variability provides an estimate of the variability for all k observations nested within each ith patient for all n patients and typically under the assumption of equal within-subject variance. Between-subject variability represents how different one patient is from another in their average measurement. Repeatability is concerned with within-subject variability since between-subject variability stems from the natural variability within the population from which the subjects are sampled and it is not inherent to technical performance of the QIB. A common method to obtain separate estimates of these two sources of variability is the following one-way random effects model

$$Y_{ij} = \mu + u_i + \varepsilon_{ij} \tag{7}$$

where μ is the overall mean, u_i is the random contribution to the intercept from the ith subject and $\sim N(0, \sigma_b^2)$, and ε_{ij} is the error term for each observation and $\sim N(0, \sigma_w^2)$.

6.2 Plots for repeatability analyses

When repeated measurements are made on a known measurand (truth or reference value is known), the measurements should be plotted against the measurand. Raw values as well as box–whisker plots provide the reader with information on both the bias to the measurand and the repeated variability on all of the subjects. Figure 5 is a simulation of the assessment of measurement variability for a QIB. QIB variability at low true values may be dominated by noise or physical limitations whereas QIB variability measured at high truth values may be restricted by the physical limitations of the

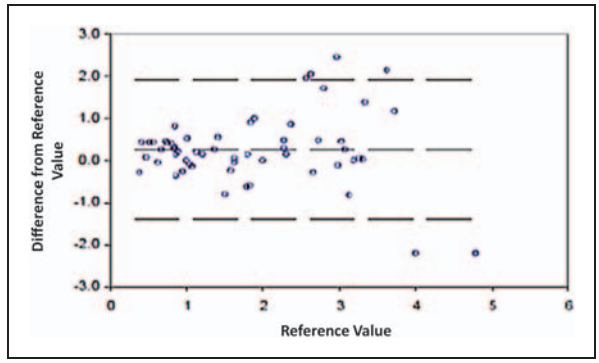

Figure 7. Bland–Altman-like plot example of agreement when a reference is available.

biology, imaging, or both. Therefore, it is important when assessing repeatability performance of a QIB to estimate measurement error for the entire measurement interval.

Studies to estimate repeatability may not have access to either the gold standard or even a reference value. In these cases, repeatability estimates measurement error at the expected value of the QIB over a range of QIB measurements, likely defined by patient status. Plotting the difference against the mean of two repeated measurements, commonly referred to as a Bland–Altman plots,[34] can show any trends in the variability of the QIB measurements over the measuring interval. An example of a Bland–Altman plot is shown in Figure 7. Here the variance increases with the mean but there does not appear to be any systematic bias between the two measurements. Bland–Altman plots help to illustrate the bias–variance relationship (if scatter does not show the same pattern over the range of average values) and limits of agreement (LOA) (discussed below). For measurements made against a reference value, the difference from the reference is compared to the reference value (see Figure 7) to visually assess any dependencies of the measurement to the measurand. Measurements made when reference values are not known use two or more repeated measurements of the QIB to compare to the mean measurement (see Figure 8). Test–retest designs typically have no a priori hierarchy between scans and differences may be performed in any order under the assumption that measurements are randomly distributed about a mean. Systematic differences between test and retest scans may exist due to unavoidable biological changes and a predefined order may demonstrate that problem but the absolute difference would not. Therefore absolute differences should also include a predefined ordered difference. For data with more than two repeated measures, the SD can be plotted against the mean to examine the mean–variance relationship.

6.3 Repeatability statistical metrics

The basis for estimates of repeatability is the within-subject variance. It is assumed that all other factors have been controlled through experimental design. Technical performance assessment of repeatability of a QIB should include the following metrics:

Within-subject variance

The within-subject variance, σ_w^2, is simply the estimated variance of repeated measurements from a single experimental unit, measured over replicates. All replicates are assumed to have the same intra-subject variance for the same measurand. Within-subject variance may include biological or

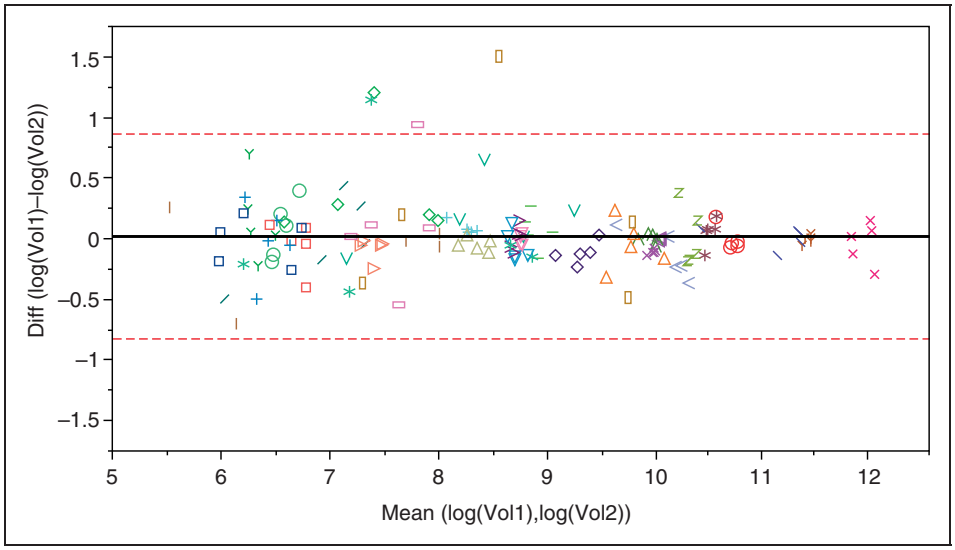

Figure 8. Bland–Altman-like plot for analysis of repeatability using QIBA publicly available data[35] when reference values are not available (see Section 10.2.1 for data). All original volumes are in mm^3.

physiological variability, which may more appropriately describe the technical performance of the QIB than a more controlled assessment of only instrument variability. For example, both patient repositioning and scanner calibrations may contribute to within-subject variance. The assumption in test–retest studies to assess repeatability is that the variance is homogeneous within the scan periods and the estimates for σ_w^2 can be obtained by pooling the within-scan period variances. This assumption will hold for most studies and in the case of unavoidable longitudinal drift, pooling the variances will at least partially adjust for differences in the scan-period means.

Repeatability coefficient (RC) and LOA

Given the model for observed measurements above in equation (7) and the within-patient variance of σ_w^2, the variance for the difference of two independent measurements of the same measurand is $\text{var}(\varepsilon_{ij} - \varepsilon_{ij'}) = 2\sigma_w^2$. The RC may be defined as the least significant difference between two repeated measurements taken under identical conditions at a two-sided significance of $\alpha = 0.05$

$$RC = 1.96\sqrt{2s_w^2} = 2.77s_w \tag{8}$$

where, for normally distributed residuals, $Z_{0.975} = 1.96$, s_w^2 is an estimate of σ_w^2 and may be obtained by either analysis of variance or likelihood-based methods.[36,37] The corresponding 95% CI ($\alpha = 0.05$) is

$$2.77\sqrt{s_w^2}\left(1/\left(\sqrt{df_\varepsilon}\chi^2_{1-\alpha/2}\right), 1/\left(\sqrt{df_\varepsilon}\chi^2_{\alpha/2}\right)\right)$$

where $df_\varepsilon = n(K-1)$.

The Analysis of Variance (ANOVA) results table can be used to estimate the within-subject variance, s_w^2, from the following identity

$$s_w^2 \equiv M_\varepsilon = SS_\varepsilon / df_\varepsilon$$

The LOA is defined as the interval where the difference between two measurements under repeatability conditions for a randomly selected subject ($Y_{i1}-Y_{i2}$) is expected to be 95% of the time and is expressed as the interval

$$LOA = [-RC, RC] \tag{9}$$

The original use of LOA stemmed from Bland–Altman analysis of agreement between two different methods and is not particularly suited for use as a measure of agreement or reliability.[38] However, the LOA can be a convenient means of assessing the ability of the QIB to meet the particular needs of a study.

The width of the RC does not depend on the sample size of a repeatability experiment, but on the precision of measurement of σ_w^2 and thus RC does depend on the number of subjects and replicates. It is important to remember that its precision must also be estimated and methods to determine the confidence limits of the RC are detailed by Barnhart and Barboriak.[30]

Intraclass correlation coefficient (ICC)

The ICC is a measure of repeated measures consistency relative to the total variability in the population.[39,40] If the entire measurement system is defined as the instrument, the patient and any factor inherent in the acquisition and measurement of an image region of interest (e.g. positioning) then ICC is the proportion of total error that is not associated with measurement error. The ICC is widely used and accepted in many QIB disciplines as an aggregate measure of repeatability. The working definition of ICC for repeatability is

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{10}$$

where σ_b^2 is the between-subject variance and σ_w^2 is the measured within-subject variance. ICC, though a measure of relative variance, may be overly high (i.e. approaches 1.0) when σ_b^2 is much greater than σ_w^2. ICC values for a very heterogeneous subject sample may yield very nearly perfect correlation based solely on the between-subject variance and ICC values that are very close to 1.0 should be cautiously interpreted. Therefore, intra- and inter-subject variance should also be evaluated when interpreting ICC as a measure of repeatability.

Some references are made in the literature to the use of ICC(j,k) terminology where j is the number of repeated measurements (e.g. scan/rescan for each subject case) and k indicates the number of observations used in the response (e.g. averaging). The ICC described in equation (10) represents ICC(2,1) and is the appropriate form of ICC for most repeatability studies. Other forms of ICC have different interpretations and a full description of the different types of ICC for different values of j and k is found in Barnhart and Barboriak.[30]

Coefficient of variation (wCV). The within-subject coefficient of variation (wCV) is often reported for repeatability studies to assess repeatability in test–retest designs. When the QIB is normally distributed with constant variance over the linearly measureable range, the value for wCV will be a function of the mean and a single estimate of RC has limited usefulness in describing the LOA and may incorrectly demonstrate a difference in repeatability due entirely to different mean QIB measurements.

Very often, however, the QIB values can be adequately characterized as lognormally distributed where the SD of the QIB measurements varies proportionally with the mean and wCV is constant.

There is a direct functional relationship of the variance of the lognormal data to wCV. The wCV is defined as in equation (11) and the relationship to the variance of the logtransformed QIB values shown in equation (12)[41]

$$wCV = \frac{\sigma_w}{\mu} \tag{11}$$

$$wCV = \sqrt{e^{w\sigma_{lt}^2} - 1} \tag{12}$$

where $w\sigma_{lt}^2$ is the within-subject variance of the logtransformed QIB. A first-order approximation of equation (12) yields a convenient approximation for wCV

$$wCV \approx w\sigma_{lt} \tag{13}$$

that hold for small values of σ_{lt}, typically much less than 1.0. Confidence intervals of wCV when QIB data are lognormally distributed can be determined on the logtransformed data as described earlier and then back transformed to the QIB original scale. While wCV is not recommended as a repeatability metric for normally distributed data, it may be appropriately used if the measurable and linear range is small and the mean QIB adequately represents the entire population described in the context of use. When not using logtransformed data to estimate wCV, confidence intervals may be approximated using methods proposed by Vangel[42] or the bootstrap which is easily implemented on repeatability studies.

6.4 Repeatability study design and analysis

6.4.1 Longitudinal changes

Repeatability studies designed for image measurement repeatability performance are influenced by many biological and technical factors and an assessment of repeatability should consider how those factors will fit into the QIB profile. For example, if repeated measurements on a patient are limited by contrast administration constraints, then there may be biological changes (e.g. tumor growth, decrease in function) and the within-patient variance for the QIB would also include the difference between the means. Therefore, any model or calculation that is used to estimate the within-patient variance should also consider including into the design an adjustment for any possible longitudinal change in the means over the scan periods.

6.4.2 Subject limitations

Repeatability study design is primarily conducted with each test performed at a single clinical site with a specific scanner to be investigated according to the procedure as described in the respective QIBA Profile. The measurements may be conducted on phantoms, animals, or human subjects, often patients. Phantom scans can be repeated several times in a sequence or with defined time period in between. Patient scan repetition has limitations due to radiation exposure for Computed Tomography (CT) or Positron Emission Tomography (PET) procedures, use of contrast media or tracers that have their own kinetic behavior (washout period needs to be considered before a rescan is possible), and patient consent. Therefore, typically these repeatability tests are limited to two performed as test–retest (e.g. Dynamic Contrast Enhanced MRI (DCE-MRI) at least 48 h apart) or sometimes known as "coffee-break" experiments with only a short break (e.g. on the order of 15 min) between scans. In practice, repeatability studies can be embedded within a larger study design that also includes multiple scanner/site reproducibility testing, discussed more fully in Section 7.

6.4.3 Sample size considerations

Repeatability studies are often designed to estimate repeated measures within-subject variance and no hypothesis tests are typically planned. Estimation studies would be more interested in the precision of the variance estimates and not in sizing for an effect size. Therefore, the precision of the within-subject variance estimate is the primary interest and is a function of both k and n. Most QIB values will be either normally or lognormally distributed and normality is a reasonable assumption. Assuming normality and sufficient degrees of freedom, the standard error of the repeatability variance ($SEVw$) for normally distributed QIB values is a function of the population variance

$$SEVw = SE\left(S_w^2\right) = \sqrt{\frac{2\sigma_w^4}{n(k-1)}}. \tag{14}$$

Equation (14) may be expressed as a ratio of the sampled to the population variance ($rSEVw$) to illustrate the relationship to the study design parameters k and n as shown in Figure 9. Though the assumptions do not rigidly hold for very small sample sizes, the results nonetheless illustrate the instability of the variance estimate under these conditions.

Most repeatability studies are configured as test–retest (i.e. $k = 2$) but different configurations may be more practical and an equivalent $rSEVw$ may be found with more repeated measurements

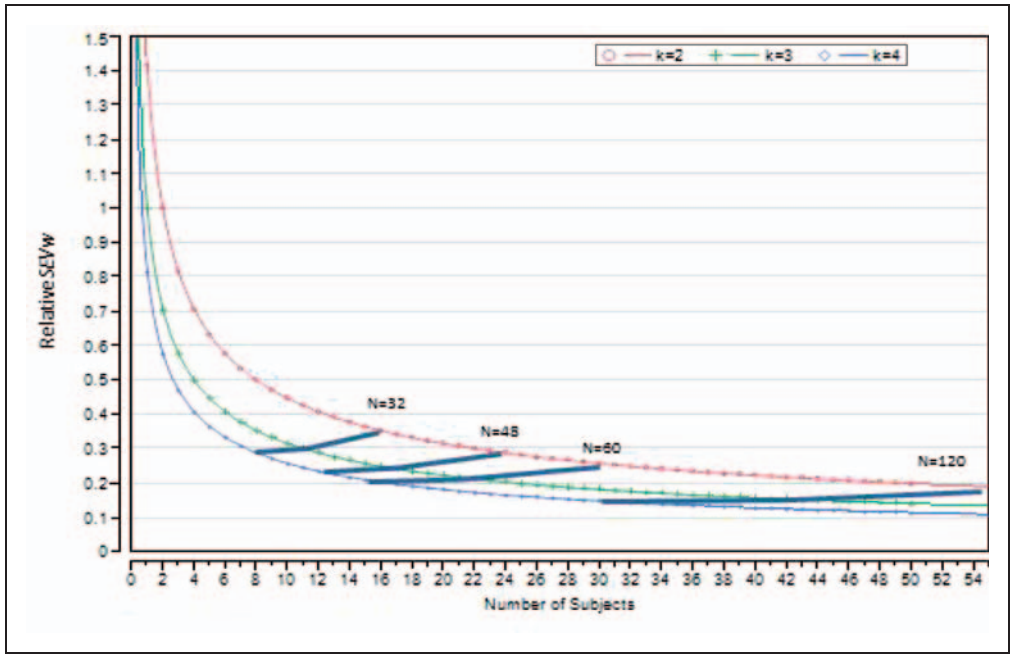

Figure 9. Relationship of sample size to $rSEVw$ for $k = 2,3,4$. $rSEVw$ values as a function of the total number of N measurements are also shown for $N = 32, 48, 60, 120$.

Table 2. Examples of repeatability study design sample size considerations.

Example 1.	Number of QIB measurements limited to $n = 60$	k = 2, $n = 30$	rel. SEVw = 0.26
		k = 3, $n = 20$	rel. SEVw = 0.22
		k = 4, $n = 15$	rel. SEVw = 0.21
Example 2.	Desired relative SEVw is 25%	k = 2, $n = 31$	
		k = 3, $n = 16$	
		k = 4, $n = 11$	

and less subjects. Often the study design is constrained by budget or logistical considerations and the total number of evaluations may be limited to a fixed value, $N = nk$. The sensitivities of $rSEVw$ to changes in n and k are shown by taking the partial derivatives and are as follows

$$\frac{\mathrm{d}(SEVw)}{dn} \propto \frac{-1}{(N-n)n}$$

and

$$\frac{\mathrm{d}(SEVw)}{dk} \propto \frac{-1}{(N-n)(k-1)}$$

When n is large enough to adequately represent the population and $k \ll n$, increasing k has a larger effect on decreasing $rSEVw$ than the proportional change in n. Two examples of designs that consider possible study constraints are presented in Table 2.

6.4.4 Analysis model considerations

Analysis methods to estimate the variances and means as described include repeated measures (ANOVA) and maximum likelihood estimation (MLE) methods. The appropriate model should be selected based on the study design assumptions understanding that MLE methods are generally superior for small sample sizes due to the asymptotic assumption of method of moments (MOM) methods. Study design should be as parsimonious as possible to reduce problems with the interpretation of the results in the presence of confounding factors or significant strata. Due to the difficulty in obtaining a homogeneous set of recruited patients in some QIB measurement studies, the use of subject data without a truth standard should be assessed for subject or patient homogeneity prior to analysis for consistency with the Profile.

7 Reproducibility

Reproducibility demonstrates the ability of a QIB to obtain the same measurement when made on the same experimental unit under different experimental conditions. It is similar to repeatability in the sense that repeated measurements are made on the same subject; however, the measurement of reproducibility includes the sum of both the within-subject and the between-condition variances

$$\sigma^2_{reproducibility} = \sigma^2_{repeatability} + \sigma^2_{between-factors} \tag{15}$$

While repeated scans on the same subject provide information on the inherent ability of the QIB to repeat the same measurement under identical or near identical conditions, repeated use of the QIB technology under different conditions stated within the QIBA Profile assesses the ability of that QIB to provide reliable data for an analysis that may include data collected under a diverse set of random or fixed conditions. Some examples of reproducibility conditions include different scanners, sites, countries or regions, and measuring systems but could also include other factors more specific to an indication. An example of a reproducibility study would compare organ volumes using one scanner type (CT) to the same organ volumes obtained on the same subject using a different scanner type (MRI) done under experimentally equivalent conditions. Another example may assess the performance of a biomarker to be reliably acquired in different geographical regions (e.g. United States/European Union/Asia) in order to determine the applicability of the biomarker within a large clinical trial. While reproducibility studies ideally evaluate different QIB conditions repeated on the same experimental unit, this design is not always possible and those study designs are addressed within this section.

The metrology definitions for reproducibility and the necessary conditions to measure reproducibility are defined in detail in Kessler.[7]

7.1 Reproducibility model definition

Reproducibility studies are designed with the goal of evaluating different factors that may affect the QIB measurement. Consider a study to evaluate reproducibility between sites randomly chosen from a large set of available sites. For n cases, $J \geq 2$ repeated measurements are taken per site for $S \geq 2$ sites. In this study design, the condition being varied is site, and when the experimental unit is available at different sites such as when phantoms are the experimental units, sites are crossed with cases. For jth repeated measurement y_{isj} on site s for case i, $i = 1, 2, \ldots n$, $s = 1, 2, \ldots S$, and $j = 1, 2, \ldots J$, consider the model

$$y_{isj} = \mu + \gamma_i + \delta_s + (\gamma\delta)_{is} + \varepsilon_{isj} \tag{16}$$

with random effects $\gamma_i \sim N(0, \sigma_\gamma^2)$ for cases, $\delta_s \sim N\left(0, \sigma_\delta^2\right)$ for site, $(\gamma\delta)_{is} \sim N(0, \sigma_{\gamma\delta}^2)$ for case by site interactions, and $\varepsilon_{isk} \sim N\left(0, \sigma_\varepsilon^2\right)$ for replicates within site and case. This model will be used for the remainder of the discussion on reproducibility. More complex designs will be relatively similar for metrics and design

7.2 Reproducibility metrics

As with an assessment of repeatability acquired by a test–retest study design, repeated measurements of a subject under different measurement conditions may also be described by similar metrics. The RC is defined for repeatability and describes an interval where the range of differences of two identical measurements obtained on the same experimental unit may be expected to occur 95% of the time. Very often the term ''repeatability coefficient'' is used to describe reproducibility and, to avoid confusion or inappropriate interpretation of RC, a similar term is defined for measurements acquired on the same experimental unit under two different conditions as the Reproducibility Coefficient (RDC)[43,44] and is defined in metrology guidelines as the absolute difference between two measurement conditions that should be exceeded by the measured differences only 5% of the time. Also described in this section is the concordance correlation coefficient (CCC) which is specifically tailored to an assessment of reproducibility with unpaired data.

It should be noted that the terms "reproducibility" and "reproducibility coefficient" are used interchangeably with "repeatability" and "repeatability coefficient" in the literature outside of metrology to define repeatability or to describe the predictive ability of a hierarchically supported item within a scale.[45] The definitions for reproducibility, including the RDC, apply to reproducibility as defined in Kessler.[7]

The reproducibility coefficient

Similar to *RC*, the reproducibility coefficient (*RDC*) may be defined as the least significant difference between two repeated measurements taken under different conditions. Using the model above as an example, the repeated measurements are taken at different sites but also could be designed to measure reproducibility across different instruments (e.g. scanners), readers/reviewers, algorithms, or any factor of interest to a clinical trial. This definition of *RDC* proposed here is consistent with metrology standards set out in ISO 5725:1994[43,44,46] and described as the "reproducibility index" by Kimothi and Kimothi[47] that extends the notion of reproducibility SD, the SD of measurement results obtained under most reproducibility conditions.[22,44,48–50]

RDC is defined here under the assumption of normality as 1.96 times the SD of a difference between two measurements y_{isj} and $y_{is'j'}$ taken on the same case i but at different site s and s'. The SD is equal to square root of two times the sum of all the variance components except for σ^2_γ, the random case effects variance. Thus

$$RDC = 2.77\sqrt{V}, \quad V = \sigma^2_\delta + \sigma^2_{\gamma\delta} + \sigma^2_\varepsilon$$

An unbiased moments-based estimate of *RDC* may be calculated from the sums of squares output of the ANOVA model as

$$\tilde{RDC} = 2.77\sqrt{\tilde{V}}, \quad \tilde{V} = k_\delta M_\delta + k_{\gamma\delta} M_{\gamma\delta} + k_\varepsilon M_\varepsilon \tag{17}$$

where coefficients $k_\delta = 1/nS$, $k_{\gamma\delta} = (n-1)/nJ$, and $k_\varepsilon = (J-1)/J$, mean squares $M_\delta = SS_\delta/df_\delta$, and $M_{\gamma\delta} = SS_{\gamma\delta}/df_{\gamma\delta}$, and $M_\varepsilon = SS_\varepsilon/df_\varepsilon$, sums of squares $SS_\delta = nJ\sum_{s=1}^{S}(\bar{y}_{s\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$, $SS_{\gamma\delta} = J\sum_{i=1}^{n}\sum_{s=1}^{S}(\bar{y}_{is\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet s\bullet} + \bar{y}_{\bullet\bullet\bullet})^2$, and $SS_\varepsilon = \sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{j=1}^{J}(y_{isj} - \bar{y}_{is\bullet})^2$, and degrees of freedom $df_\delta = S-1$, $df_{\gamma\delta} = (n-1)(S-1)$, and $df_\varepsilon = nS(J-1)$.

An example of the above model can be found in Section 10.3.1. Assuming the measurements are normally distributed, a large sample 95% CI on *RDC* may be obtained by the method of Graybill and Wang[51] because *RDC* is a linear combination of mean squares with nonnegative coefficients. The 95% CI is

$$2.77\left(\tilde{V} - \sqrt{(p_\delta k_\delta M_\delta)^2 + (p_{\gamma\delta} k_{\gamma\delta} M_{\gamma\delta})^2 + (p_\varepsilon k_\varepsilon M_\varepsilon)^2},\right.$$
$$\left.\tilde{V} + \sqrt{(q_\delta k_\delta M_\delta)^2 + (q_{\gamma\delta} k_{\gamma\delta} M_{\gamma\delta})^2 + (q_\varepsilon k_\varepsilon M_\varepsilon)^2}\right)^{1/2}$$

where $p_a = 1 - 1/F_{1-\alpha/2}(df_a, \infty)$, $q_a = 1/F_{\alpha/2}(df_a, \infty) - 1$, and $F_\beta(f_n, f_d)$ is the 100 β th percentile of the F distribution with numerator and denominator degrees of freedom f_n and f_d. Note $F_\beta(df_a, \infty) = \chi^2_\beta(df_a)/df_a$.

Alternative estimators and confidence intervals on *RDC* could be constructed based on maximum likelihood, Satterthwaite approximation, an exact method, etc. For various estimators and confidence intervals of functions of variance components, see Searle et al.,[52] Milliken and

Johnson,[53] and Burdick and Graybill.[54] Conceivably, the bootstrap could be used to obtain a nonparametric CI on *RDC*. However, formulation of the bootstrap samples is unclear because of the multiple sources of variation involved.

Concordance Correlation Coefficient (CCC)

The CCC was proposed by Lawrence and Lin[55] as a more complete evaluation of agreement between multiple observations per case made without ANOVA assumptions. It describes the extent to which paired measurements diverge from perfect agreement, reflecting both systematic differences between repeated measurements and variability. In the case where two factors are being compared agreement, the CCC provides a metric of reproducibility that may be easily calculated for two classes, such as scanner type, chosen as fixed effects. The CCC is defined as

$$CCC = \frac{\sigma_1 \sigma_2 \rho_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \tag{18}$$

where σ_1^2 and σ_2^2 are the variances for each class, and μ_1 and μ_2 are the group means. Confidence intervals should be included in any reporting of CCC and methods are found in Chen and Barnhart.[40]

7.3 Plots for reproducibility analyses

Plots representing the measurements at each of the factors to be evaluated for reproducibility are an important tool to visually inspect the data summarized by the metrics. The following plots are recommended for any analysis of reproducibility

- Paired data
  - Scatter Plots: Method 1 versus Method 2 with corresponding fitted regression or robust nonparametric fitted lines (if applicable) when the data are paired with the same experimental unit.
  - Bland–Altman plots[34] are a widely used and valuable visualization tool in the analysis of QIB reproducibility and are especially valuable when a standard reference is not available.
- Box and whisker plots with reproducibility conditions as x-categories when data are not paired with outlier identification.
  - Individual points identified and jittered.
  - Outlier detection using interquartile range (IQR) criteria of $1.5 * \text{IQR}$.
  - Connecting median values between groups may unduly influence an analysis trends and is not generally recommended for these analyses.
- Distribution analyses
  - Histograms with normal or lognormal distribution fit (typical). Other distributions may also be shown (not typical).
  - Q–Q plots of model residuals.
  - Leverage plots for detection and analysis of high leverage points.

Data chosen at random from QIBA 3A project[31] for two different measurement groups for reproducibility were used in Figures 10 and 11 as an example of an analysis of CT volumetry measurements using two measurement methodologies. Several different phantoms of different sizes and configurations as well as different measurement methods were used. All phantoms were imaged at the same slice thickness. In Figure 10a and 10b, Group 01 and Group 02 were plotted

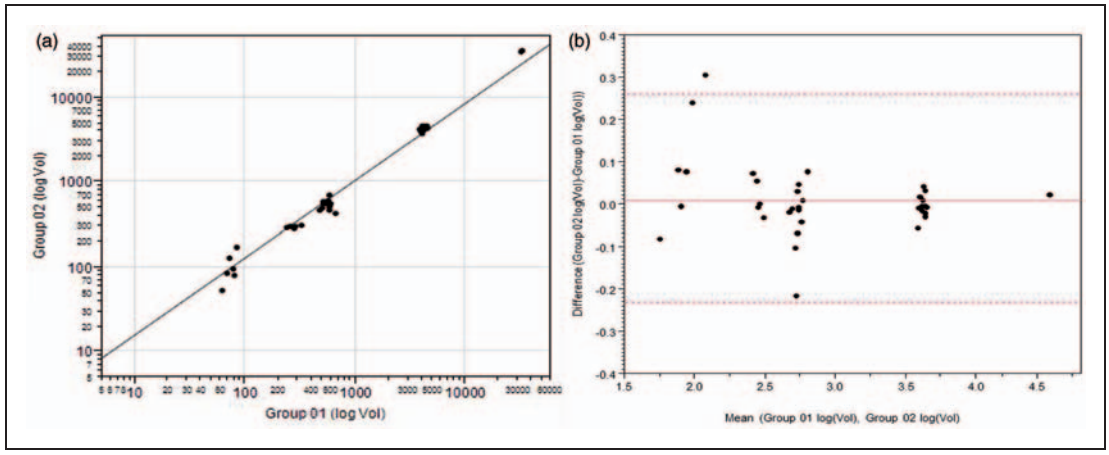

Figure 10. Phantom CT volumetry comparison of two methods plotted for logtransformed data (original units mm^3). (a) slope = 0.974, intercept = 0.086 and (b) shows the Bland–Altman plot with lower and upper agreement limits. Correlation = 0.99; RDC = 0.234.

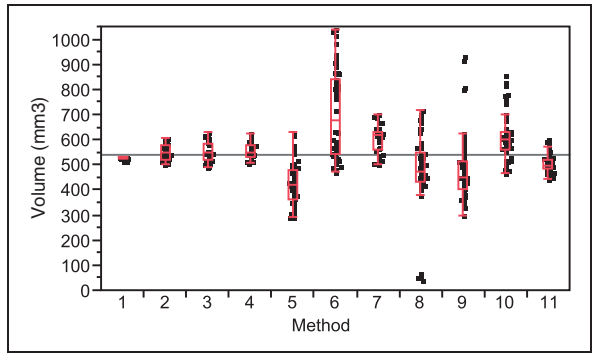

Figure 11. Box–whisker plot for multigroup reproducibility. Points are jittered for viewing. Whiskers indicate 1.5 * IQR outlier boundary.

against each other along with a fitted least squares regression line. An orthogonal regression also was run, though the fitted line is nearly identical and not distinguishable from the least squares line displayed. All analyses were conducted using JMP® 10.0.1 software (SAS Institute Inc. Cary, NC). There is essentially no bias (intercept is equivalent to a volume bias of approximately 1 mm^3) but a slight tendency for Group 02 to measure volumes smaller than Group 01. There is also an apparent inverse relationship of the variability relative to the mean. Alternately, Figure 11 displays box–whisker plots for 11 different methods for estimating volumes taken for a single phantom set at one slice thickness from the QIBA 3A challenge data.[31] This plot shows the point-wise and distribution statistics of each method about the overall mean, essentially equivalent to a Bland–Altman plot for

Table 3. Examples of reproducibility study design component examples.

Model component	Type (typically) Fixed or random	Example
Methodology	Fixed (typically)	Slice thickness
Scanner type	Fixed	MRI/CT/Ultrasound
Scanner manufacturer	Fixed	Siemens/Phillips/GE
Site	Fixed or random	Sites within US
Region	Fixed	EU/US/Asia
Population	Fixed	Consortia databases (ADNI versus Descripa)
Operator	Random	MRI tech
Radiological reviewer	Random (preferable) Fixed (rare)	– Oncology radiologist measurement of longest diameter – Experienced versus inexperienced

multiple groups. These plots are common and found in most statistical plotting tools. Other types of plots that display means and distributions in a different may not be commonly available or familiar.

7.4 Reproducibility study design and analysis

Study design

Designs for reproducibility studies fall under two main categories

(1) Repeated measurement design: reproducibility conditions are used to acquire repeated measurements on each experimental unit. These designs result in the total variance as shown in equation (15). Furthermore, measurements repeated for each condition allow for further estimation of each of the variance components. The inference is on the experimental unit.
(2) Cohort measurement design: measurements for different reproducibility components are acquired from different subjects. This is especially true when evaluating reproducibility between sites. The inference is on the cohort.

Examples of different reproducibility components are presented in Table 3 and are only limited by the ability to compare one component to another.

Analysis—Hypothesis versus Descriptive

The analysis of reproducibility data may be simply descriptive; however, a typical reason to conduct a reproducibility study would be to determine the ability to reliably conduct a study under different conditions or with different QIB methodologies with results statistically identical to a study conducted with no variation in conditions. There may also be an interest in contrasting one methodology to another for superiority. Therefore, analyses of reproducibility will typically involve hypothesis tests for equivalence, superiority, or noninferiority. Descriptions of these hypotheses for algorithm reproducibility are found in the companion paper in this series by Obuchowski[8] which can easily be modified for other reproducibility conditions.

Analysis of relative nonsystematic bias

The mixed effects model described by equation (14) may be used to test fixed effects, including interaction effects to test for a systematic bias between methods. Similarly, nonsystematic bias

between methods may be tested using the following model

$$y_{1i} = \mu + (1 + \beta)(y_{2i} - a_{2i}) + \varepsilon_{1i} \tag{19}$$

where y_{1i} is condition 1 and y_{2i} is condition 2; a_{2i} is the error term for the random variable y_{2i} and ε_{1i} is the error term for variable y_{1i}. The analysis tests for evidence of nonsystematic bias as

$$\begin{aligned} &\mathrm{H}_0(\text{no non} - \text{systematic bias}): \quad \beta = 0 \\ &\mathrm{H}_\mathrm{A}(\text{non} - \text{systematic bias}): \quad \beta \neq 0 \end{aligned}$$

Since the variance of both error terms are likely to be near equal, a least squares linear regression analysis is not appropriate and an alternative error in variables method must be used to estimate β. Some useful methods to determine unbiased estimates of β are discussed by Graybill[56] and Draper and Smith.[57]

Analysis of RDCs

There may be an interest in comparing reproducibility to repeatability since equivalence of the two reliability metrics would be evidence of equivalence in the absence of any nonsystematic bias. If the repeated measurements acquired for all reproducibility conditions (e.g. test–retest for both reproducibility conditions) and the measurements can be assumed to have been acquired without any change in the condition of the subject, then an estimate of RC is appropriate. Recall that the repeatability SD is σ_ε. The RC is $RC = 2.77\sigma_\varepsilon$. An estimate of RC is $\widetilde{RC} = 2.77\sqrt{M_\varepsilon}$. The corresponding 95% CI is

$$2.77\sqrt{df_\varepsilon M_\varepsilon}\left(1/\chi^2_{1-\alpha/2}(df_\varepsilon), 1/\chi^2_{\alpha/2}(df_\varepsilon)\right) \tag{20}$$

with $\alpha = 0.05$.

One may wish to test if reproducibility variation exceeds repeatability variation. The null and alternative hypotheses are

$$\begin{aligned} &H_0 : \sigma^2_\delta = \sigma^2_{\gamma\delta} = 0 \\ &H_\mathrm{A} : \sigma^2_\delta > 0 \quad \text{or} \quad \sigma^2_{\gamma\delta} > 0 \end{aligned}$$

Equivalently, one may test

$$\begin{aligned} &H'_0 : n\sigma^2_\delta + nJ\sigma^2_{\gamma\delta} = 0 \\ &H'_\mathrm{A} : nJ\sigma^2_\delta + J\sigma^2_{\gamma\delta} > 0 \end{aligned}$$

Under H'_0, the test statistic $F_\delta = M_\delta / M_\varepsilon \sim F(df_\delta, df_\varepsilon)$. Thus at level α, H_0 is rejected when $F_\delta > F_{1-\alpha}(df_\delta, df_\varepsilon)$.

The reproducibility study described earlier is relatively simple, with the levels of only one condition (site) being varied and crossed with cases. More complex reproducibility studies can be considered, in which variation in the levels of multiple conditions is studied simultaneously, with the conditions either nested or crossed with each other. It is important that the statistical analysis of the variance components reflects the study design.

A reproducibility study could also be designed to compare the *RDC*s of two or more algorithms. For the same design as described earlier but with repeated measurements on each of $m = 1, 2, \ldots M$

algorithms instead of just one, the model shown in equation (16) would be modified to include fixed effects for the algorithms, and random effects that depend on *j* for case, site, and site by case. In this mixed effects model, the *RDC*s for the algorithms could be estimated by extending the method of moments described to obtain estimators based on linear combinations of mean squares. Note, for a difference in *RDC* between two algorithms, some of the mean squares in the linear combination estimator will have negative coefficients, invalidating the method of Graybill and Wang for constructing a 95% CI.[51] Reproducibility of different algorithms may also be compared by the CCC.[58]

For studies in which it is desired to vary many conditions simultaneously, to compare many algorithms, or both, the number of repeated measurements necessary per case could be prohibitive. The number could be reduced by considering an incomplete block design, with each case being considered a block. In principle, designs of such studies could be constructed such that lower order variance components are estimable or least confounded with higher order variance components that are expected to be relatively small.

Anecdotally, reproducibility of structural measurements is reader- and algorithm-dependent and is amenable to analysis using these summary measures,[59] but the greater source of error for functional measurements such as the standardized uptake value (SUV) from fluorodeoxyglucose (FDG) PET imaging is outliers due to mistaken transfer of information such as patient weight and injection dose.[60] Specification of the random effects model is critical to the design, analysis, and interpretation of these studies. Additionally, note that published results of studies involving rater agreement should adhere to Guidelines for Reporting Reliability and Agreement Studies.[61]

RDC depends on the conditions being varied in any given study of measurement reproducibility. Ideally, the largest sources of variation for the QIB are known, and the reproducibility study includes all of these sources. The calculated *RDC* would then represent the total variation that could be expected in practice. If instead the reproducibility study includes only a subset of the important sources of variation, then the *RDC* is intermediate between the total *RDC* and the *RC* and may not even apply to a particular clinical setting. To illustrate, for FDG-PET, the largest sources of variation might be interscan, interobserver, intraobserver, and interday.[62] A reproducibility study could then be designed to include all of these sources of variation.

Reproducibility answers the question regarding reliability of measures across different clinical sites, different scanner models and manufacturers, different operators, technicians, radiologists, or standard procedures and any variable within a clinical trial that may affect the reliability of the study results. Ideally, all engaged sites in a reproducibility study should use equivalent phantom types and follow the QIBA Profile for the respective QIB in their respective clinical environment. The number of reproducible factors that are evaluated is determined by the QIBA Profile and the needs of the end-user. However, in a large clinical trial where the number of sites may be large, sites should be selected in a random or near-random manner and reflect the impact of a site on the number of subjects or patients to be included.

8 Summary

A general model for the overall technical performance of a QIB may use equation (16) with measures repeated as test–retest for each reproducibility factor as an estimate of measurement error. Linearity and bias assessment are included into the model by determining the measurand range to capture the full measurement interval which can use previous technical descriptions of QIB performances. Table 4 provides a general summarized description as a starting point for designing a QIB technical performance study.

Table 4. A general QIB technical performance study design and recommended design parameters. Actual study requirements may differ depending on the specific QIB and study requirements.

Performance element	Preferred reference	Measurand levels	Number of replicates at each level	Repeated measures	Statistical metrics
Bias/Linearity	• Phantom (preferred) • Reference QIB (typical)	• Equally spaced across the measuring interval • Recommended: 5–7 • Required: ≥ 3 • Curvature consideration for spacing	$l = 3$ (recommended)	Ideal but not required	• Intercept = 0 • Slope = 1 • Curvature=0 • Unimodality
Repeatability	• Subject • Phantom not required	Fully represent the measuring interval	Varies to get stable estimate of the variance[34]	$k \geq 2$	• Variance components • ICC(2,1) • RC/LOA
Reproducibility	• Phantom (preferred) • Same subject (recommended) • Same population (may be necessary)	Fully represent the measuring interval	Varies to get stable estimate of the variance[34]	Ideal to estimate variance separate components but not typically required	• Variance components • CCC • RDC / LOA
Combined	• Phantom (preferred) • Same subject (recommended) • Same population (may be necessary)	Fully represent the measuring interval	Varies to get stable estimate of the variance[34]	$k \geq 2$	• Variance components • ICC • CCC • RC/RDC • LOA

9 Discussion and future directions

This paper gives a basic introduction in designing and conducting a technical performance analysis study for a QIB and some of the key issues to consider. Clearly, we just touch the surface of what is involved in conducting this type of study. Practical concerns related to balancing QIB performance with available resources have to be addressed in any QIB study and often prove to be quite challenging. Likewise, many of the issues discussed earlier are not fully characterized and may require additional research efforts. This includes topics such as how to address linearity of the QIB if no standard reference is available or if only a partial standard reference exists. An example of this is not having a gold reference standard for FDG SUV in clinical cases. In this example, a complementary modality may be able to provide a partial standard but not a true gold standard. For some QIBs, evaluating linearity and dynamic range across lesion properties and imaging conditions may be achievable through phantom studies. For other QIBs, phantoms may not be available and an alternate approach for dealing with linearity would need to be developed. The lack of a standard reference would likely also impact possible claims and study designs for assessing the QIB. Another area that needs to be tailored to the application area and specific claim is the reproducibility evaluation. For QIBs, the patient prep, image acquisition, and image processing parameter space and other factors can quickly grow to an unmanageable size. Clearly, work is needed to develop systematic approaches for quickly identifying important parameters across which reproducibility testing must be conducted. This would likely involve the collection of preliminary data where the use of computation imaging models or phantom data may be able to play an important role.

Again, the focus of this paper is on technical performance analysis to address the question of how well a measurement can be made. The technical assessment is necessary but not sufficient for validating a QIB. The next step in the process is then to validate the clinical utility of making the measurements. A different class of studies is likely required to show that a QIB has utility in either clinical practice or clinical trials. For example, it would likely involve randomized clinical trials where the QIB is expected to be a surrogate endpoint for patient response. However, QIBs used for defining treatment populations or for safety may require different and perhaps less resource-intensive evaluations to prove effectiveness. Currently, systematic and efficient approaches for validating clinical utility of imaging biomarkers have not been fully developed. Consensus on how to design and conduct these clinical assessment studies is the next step in the process of developing a systematic framework for assessing QIBs and promulgating them into wider clinical practice and clinical trial use.

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.

References

1. Mueller DK, Kutscherenko A, Bartel H, et al. Phantom-less QCT BMD system as screening tool for osteoporosis without additional radiation. *Eur J Radiol* 2011; **79**: 375–381.
2. Habte F, Budhiraja S, Keren S, et al. In situ study of the impact of inter-and intra-reader variability on region of interest (ROI) analysis in preclinical molecular imaging. *Am J Nucl Med Mol Imaging* 2013; **3**: 175.
3. Bhavane R, Badea C, Ghaghada KB, et al. Dual-energy computed tomography imaging of atherosclerotic plaques in a mouse model using a liposomal-iodine nanoparticle contrast agent. *Circ Cardiovasc Imaging* 2013; **6**: 285–294.
4. Hamburg MA and Collins FS. The path to personalized medicine. *N Engl J Med* 2010; **363**: 301–304.
5. Atkinson AJ, Colburn WA, DeGruttola VG, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Therapeut* 2001; **69**: 89–95.
6. Atkinson AJ, Colburn WA, DeGruttola VG, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework∗. *Clin Pharmacol Therapeut* 2001; **69**: 89–95.

7. Kessler L. The emerging science of quantitative imaging biomarkers: Terminology and definitions for scientific studies and for regulatory submissions. *Stat Methods Med Res* 2014 (in press).
8. Obuchowski N. Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res* 2014 (in press).
9. Sullivan DBL, Seto B, Obuchowski N, et al. Introduction to metrology series. *Stat Methods Med Res* 2014 (in press).
10. McShane L. Meta-analysis of the technical performance of an imaging biomarker: Guidelines, statistical methodology, and examples. *Stat Methods Med Res* 2014 (in press).
11. http://qibawiki.rsna.org/index.php?title=Main_Page/ QIBA Main Page (accessed 21 December 2013).
12. ISO/IEC. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*. ISO/IEC Guide 99. Geneva: International Organization for Standardization, 2007.
13. http://qibawiki.rsna.org/index.php?title=What_Are_ Profiles_and_Protocols%3F. What are Profiles and Protocols (accessed 21 December 2013).
14. http://qibawiki.rsna.org/index.php?title=Main_Page. QIBA Wiki Main Page. (accessed 21 December 2013).
15. Barnhart HX, Haber MJ and Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Statist* 2007; **17**: 529–569.
16. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3): 983–997.
17. QI-Bench, 10 October 2013.
18. Natrella M. *NIST/SEMATECH e-handbook of statistical methods*. NIST/SEMATECH, 2010, http://www itl nist gov/div898/handbook
19. CLSI/NCCLS. *Evaluation of the linearity of quantitative measurement procedures: A statistical approach; approved guideline. CLSI/NCCLS document*. Wayne, PA: NCCLS, 2010.
20. Czichos H, Saito T and Smith LE. *Springer handbook of metrology and testing*. Berlin Heidelberg: Springer-Verlag, 2011.
21. ISO. *Statistical interpretation of data – Part 4: Detection and treatment of outliers ISO 16269-4*. 2010.
22. (2002) C. *Evaluation of precision performance of quantitative measurement methods; Approved guideline–second edition. CLSI document*. EP5-A2. 2002.
23. CLSI. *Measurement procedure comparison and bias estimation using patient samples: Approved guideline-third edition. CLSI Document EP09-A3*. Wayne, PA: Clinical and Laboratory Standards Institute, 2013.
24. Penny KI and Jolliffe IT. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *J R Stat Soc Ser D (Statist)* 2001; **50**: 295–307.
25. Rosner B. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 1983; **25**: 165–172.
26. Snedecor G and Cochran WG. *Statistical methods*. Iowa: Ames, 1989.
27. Montgomery DC, Peck EA and Vining GG. *Introduction to linear regression analysis*. Hoboken, NJ: Wiley, 2012.
28. Levene H. In: Olkin I (ed.) *Contributions to probability and statistics: Essays in honor of Harold hotelling*. Redwood City, CA: Stanford University Press, 1960.
29. Carroll RJ and Ruppert D. *Transformation and weighting in regression*. New York, NY: Chapman & Hall A International Thomson Publishing Company, 1988.
30. Barnhart HX and Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: A review of statistical analysis of repeat data sets. *Transl Oncol* 2009; **2**: 231.
31. http://qibawiki.rsna.org/index.php?title=VolCT_-_Group_3A. QIBA 3A study. (accessed 21 December 2013).
32. Berkson J. Are there two regressions? *J Am Stat Assoc* 1950; **45**: 164–180.
33. Haeckel R, Wosniok W and Klauke R. Comparison of ordinary linear regression, orthogonal regression, standardized principal component analysis, Deming and Passing-Bablok approach for method validation in laboratory medicine. *Laboratoriumsmedizin* 2013; **37**: 147–163.
34. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **327**: 307–310.
35. Demonstrator Q-BQ. s_rdg_Test-retest-reliability-study.csv. 2 January 2014.
36. Neter J, Kutner MH, Nachtsheim CJ, et al. *Applied linear regression models*. Chicago, USA: Irwin, 1996, p. 1050.
37. Fitzmaurice GM, Laird NM and Ware JH. *Applied longitudinal analysis*. Hoboken, NJ: Wiley, 2012.
38. Streiner DL. Learning how to differ: Agreement and reliability statistics in psychiatry. *Can J Psychiatry* 1995; **40**: 60–66.
39. Shrout PE and Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979; **86**: 420–428.
40. Chen C-C and Barnhart HX. Comparison of ICC and CCC for assessing agreement for data without and with replications. *Comput Statist Data Anal* 2008; **53**: 554–564.
41. Johnson NL, Kotz S and Balakrishnan N. Continuous univariate distributions, vol. 1. Hoboken, NJ: Wiley & Sons, 1994.
42. Vangel MG. Confidence intervals for a normal coefficient of variation. *Am Statist* 1996; **50**: 21–26.
43. I. *Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definitions. ISO 5725-1: Cor 1:1998*. Geneva, 1994.
44. I. *Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. ISO 5725–2* 1994.
45. Jobling D and Snell EJ. The use of the coefficient of reproducibility in attitude scaling. *Incorp Statist* 1961; **11**: 110–118.
46. (BIPM). BIdPeM. *International vocabulary of metrology – Basic and general concepts and associated terms (VIM, 3rd edition, JCGM 200:2008) and Corrigendum (May 2010)*.
47. Kimothi SK and Kimothi S. *The uncertainty of measurements: Physical and chemical metrology: Impact and analysis*. ASQ Press, Milwaukee, WI, 2002.
48. R D. Vocabulary for use in measurement procedures and description of reference materials in laboratory medicine. *Eur J Clin Chem Clin Biochem* 1997; **35**: 141–173.
49. ISO. *Statistics – Vocabulary and symbols – Part 2: Applied statistics. ISO-3534-2-2006*.
50. Linnet K and Boyd JC. Selection and analytical evaluation of methods with statistical techniques. In: Burtis CA, Ashwood ER and Bruns DE (eds) *Tietz textbook of clinical chemistry and molecular diagnostics*, 4th ed. St Louis: Elsevier Saunders, 2006, pp.7–47.
51. Graybill FA and Wang C-M. Confidence intervals on nonnegative linear combinations of variances. *J Am Statist Assoc* 1980; **75**: 869–873.
52. Searle S, Casella G and McCulloch C. *Variance components*. New York: John Wiley and Sons, 1992.
53. Milliken GA and Johnson DE. *Analysis of messy data volume 1: Designed experiments*. Chapman and Hall/CRC, Boca Raton, FL, 2009.
54. Burdick R and Graybill F. *Confidence intervals on variance components*. New York: Marcel Dekker Inc, 1992.
55. Lawrence I and Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 255–268.
56. Graybill FA. *An introduction to linear statistical models*. New York: McGraw-Hill, 1961.

57. Draper NP and Smith H. *Applied regression analysis*. Wiley, New York, 1998.
58. Crawford SB, Kosinski AS, Lin H-M, et al. Computer programs for the concordance correlation coefficient. *Comput Methods Prog Biomed* 2007; **88**: 62–74.
59. Jacene HA, Leboulleux S, Baba S, et al. Assessment of interobserver reproducibility in quantitative 18F-FDG PET and CT measurements of tumor response to therapy. *J Nucl Med* 2009; **50**: 1760–1769.
60. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. *J Nucl Med* 2009; **50**: 1646–1654.
61. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int J Nurs Stud* 2011; **48**: 661–671.
62. Rudd JH, Myers KS, Bansilal S, et al. Atherosclerosis inflammation imaging with 18F-FDG PET: Carotid, iliac, and femoral uptake reproducibility, quantification methods, and recommendations. *J Nucl Med* 2008; **49**: 871–878.

Appendix 1

10.1. Linearity example

The following data were derived from the QIBA 3A challenge[35] results to represent linearity of replicate measurements of a known reference value or measurand

Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)	Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)
1	43,568	1	41,762	17	1768	1	2867
1	43,568	2	45,493	17	1768	2	324
1	43,568	3	41,687	17	1768	3	286
1	43,568	4	44,925	17	1768	4	3034
1	43,568	5	43,156	17	1768	5	2166
2	624	1	590	18	7176	1	7914
2	624	2	724	18	7176	2	6956
2	624	3	611	18	7176	3	6817
2	624	4	433	18	7176	4	6797
2	624	5	739	18	7176	5	8625
3	1900	1	1341	19	1137	1	1655
3	1900	2	3147	19	1137	2	1488
3	1900	3	770	19	1137	3	324
3	1900	4	3003	19	1137	4	935
3	1900	5	3014	19	1137	5	1546
4	21,169	1	18,810	20	68,505	1	34,410
4	21,169	2	20,423	20	68,505	2	82,976
4	21,169	3	20,608	20	68,505	3	61,929
4	21,169	4	19,918	20	68,505	4	83,204
4	21,169	5	20,664	20	68,505	5	81,726
5	17,492	1	16,690	21	2092	1	2834
5	17,492	2	18,213	21	2092	2	1571
5	17,492	3	11,735	21	2092	3	1678
5	17,492	4	16,764	21	2092	4	3912
5	17,492	5	21,765	21	2092	5	1339
6	87,563	1	91,529	22	9039	1	2340
6	87,563	2	80,985	22	9039	2	18,890
6	87,563	3	86,002	22	9039	3	1226

(continued)

Continued

Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)	Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)
6	87,563	4	89,348	22	9039	4	10,982
6	87,563	5	90,034	22	9039	5	13,196
7	6315	1	6252	23	717	1	771
7	6315	2	6488	23	717	2	589
7	6315	3	6151	23	717	3	613
7	6315	4	6470	23	717	4	776
7	6315	5	6325	23	717	5	1008
8	29,821	1	32,159	24	3816	1	4114
8	29,821	2	29,541	24	3816	2	3995
8	29,821	3	26,684	24	3816	3	4003
8	29,821	4	34,306	24	3816	4	3472
8	29,821	5	27,027	24	3816	5	4331
9	3277	1	2827	25	10,986	1	11,309
9	3277	2	5551	25	10,986	2	13,070
9	3277	3	1043	25	10,986	3	8089
9	3277	4	446	25	10,986	4	9456
9	3277	5	6504	25	10,986	5	10,269
10	579	1	881	26	148,060	1	132,118
10	579	2	359	26	148,060	2	167,965
10	579	3	411	26	148,060	3	124,519
10	579	4	548	26	148,060	4	164,118
10	579	5	671	26	148060	5	139948
11	22,463	1	22,531	27	4197	1	4438
11	22,463	2	22,386	27	4197	2	3425
11	22,463	3	22,438	27	4197	3	3873
11	22,463	4	22,567	27	4197	4	4033
11	22,463	5	22,221	27	4197	5	4747
12	1615	1	1402	28	563	1	581
12	1615	2	1472	28	563	2	541
12	1615	3	1452	28	563	3	503
12	1615	4	1711	28	563	4	674
12	1615	5	1727	28	563	5	736
13	20,706	1	18,568	29	5992	1	6170
13	20,706	2	21,029	29	5992	2	6295
13	20,706	3	21,883	29	5992	3	5636
13	20,706	4	20,774	29	5992	4	5422
13	20,706	5	21,479	29	5992	5	5492
14	5274	1	1172	30	35818	1	31772
14	5274	2	11522	30	35818	2	41970
14	5274	3	6294	30	35818	3	38900
14	5274	4	6478	30	35818	4	35898
14	5274	5	3916	30	35818	5	36461
15	24,667	1	16,011	31	806	1	763
15	24,667	2	24,526	31	806	2	860
15	24,667	3	24,602	31	806	3	913

(continued)

Continued

Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)	Subject	Measurand (mm^3)	Replicate number	Measured volumes (mm^3)
15	24,667	4	23,365	31	806	4	667
15	24,667	5	24,867	31	806	5	712
16	6024	1	6226				
16	6024	2	5568				
16	6024	3	5549				
16	6024	4	6630				
16	6024	5	6344				

10.2. Repeatability example

10.2.1. Data

The following data were chosen as a subset of the QIBA 3A challenge[35] for repeated measures of phantom volumes over a range of volumes and phantom shapes. All measurements are natural log transformation of volumes with the original units as mm^3.

Subjid	Sample	Test	Retest	Subjid	Sample	Test	Retest
1	1	10.71	10.64	17	1	8.03	7.96
1	2	10.75	10.73	17	2	6.48	5.78
1	3	10.46	10.64	17	3	5.39	5.66
1	4	10.78	10.71	17	4	7.97	8.02
1	5	10.71	10.67	17	5	7.69	7.68
2	1	6.04	6.38	18	1	8.70	8.98
2	2	6.42	6.58	18	2	8.71	8.85
2	3	6.43	6.42	18	3	8.79	8.83
2	4	6.56	6.07	18	4	8.97	8.82
2	5	6.66	6.61	18	5	9.00	9.06
3	1	6.92	7.20	19	1	7.14	7.41
3	2	7.90	8.05	19	2	7.34	7.31
3	3	6.50	6.65	19	3	6.27	5.78
3	4	6.80	8.01	19	4	7.00	6.84
3	5	7.81	8.01	19	5	6.91	7.34
4	1	9.98	9.84	20	1	10.35	10.45
4	2	10.03	9.92	20	2	11.26	11.33
4	3	10.03	9.93	20	3	11.17	11.03
4	4	10.00	9.90	20	4	11.26	11.33
4	5	10.01	9.94	20	5	11.33	11.31
5	1	9.78	9.72	21	1	7.85	7.95
5	2	9.76	9.81	21	2	7.90	7.36
5	3	9.69	9.37	21	3	7.31	7.43
5	4	9.48	9.73	21	4	7.33	8.27
5	5	10.15	9.99	21	5	7.18	7.20
6	1	11.38	11.42	22	1	7.55	7.76

(continued)

Continued

Subjid	Sample	Test	Retest	Subjid	Sample	Test	Retest
6	2	11.35	11.30	22	2	9.70	9.85
6	3	11.35	11.36	22	3	7.47	7.11
6	4	11.41	11.40	22	4	7.79	9.30
6	5	11.40	11.41	22	5	9.97	9.49
7	1	8.78	8.74	23	1	6.55	6.65
7	2	8.73	8.78	23	2	6.57	6.38
7	3	8.77	8.72	23	3	6.55	6.42
7	4	8.73	8.77	23	4	6.45	6.65
7	5	8.73	8.75	23	5	6.52	6.92
8	1	10.00	10.38	24	1	8.26	8.32
8	2	10.42	10.29	24	2	8.21	8.29
8	3	10.37	10.19	24	3	8.23	8.29
8	4	10.28	10.44	24	4	7.97	8.15
8	5	10.37	10.20	24	5	8.31	8.37
9	1	6.80	7.95	25	1	9.44	9.33
9	2	8.68	8.62	25	2	9.44	9.48
9	3	7.39	6.95	25	3	9.13	9.00
9	4	6.31	6.10	25	4	9.38	9.15
9	5	8.92	8.78	25	5	9.35	9.24
10	1	6.69	6.78	26	1	11.77	11.79
10	2	6.06	5.88	26	2	11.87	12.03
10	3	5.96	6.02	26	3	11.85	11.73
10	4	6.09	6.31	26	4	11.93	12.01
10	5	6.76	6.51	26	5	12.14	11.85
11	1	10.01	10.02	27	1	8.51	8.40
11	2	10.02	10.02	27	2	8.19	8.14
11	3	10.03	10.02	27	3	8.23	8.26
11	4	10.03	10.02	27	4	8.38	8.30
11	5	10.02	10.01	27	5	8.47	8.47
12	1	7.30	7.25	28	1	6.11	6.36
12	2	7.30	7.29	28	2	6.23	6.29
12	3	7.52	7.28	28	3	6.44	6.22
12	4	7.49	7.44	28	4	6.47	6.51
12	5	7.49	7.45	28	5	5.90	6.60
13	1	9.83	9.83	29	1	8.59	8.73
13	2	10.01	9.95	29	2	8.88	8.75
13	3	9.99	9.99	29	3	8.62	8.64
13	4	9.89	9.94	29	4	8.77	8.60
13	5	9.94	9.97	29	5	8.76	8.61
14	1	7.23	7.07	30	1	10.50	10.37
14	2	9.11	9.35	30	2	10.45	10.64
14	3	8.09	8.75	30	3	10.49	10.57
14	4	8.75	8.78	30	4	10.41	10.49
14	5	8.11	8.27	30	5	10.42	10.50
15	1	9.57	9.68	31	1	6.65	6.64
15	2	10.46	10.11	31	2	6.80	6.76

(continued)

Continued

Subjid	Sample	Test	Retest	Subjid	Sample	Test	Retest
15	3	10.33	10.11	31	3	6.73	6.82
15	4	10.28	10.06	31	4	6.38	6.50
15	5	10.09	10.12	31	5	6.97	6.57
16	1	8.69	8.74				
16	2	8.74	8.62				
16	3	8.69	8.62				
16	4	8.65	8.80				
16	5	8.72	8.76				

10.3. Reproducibility example 1

10.3.1. Data 1

Group 01 Volumes	Group 02 Volumes	Group 01 Volumes	Group 02 Volumes	Group 01 Volumes	Group 02 Volumes	Group 01 Volumes	Group 02 Volumes
518.799	553.894	585.938	695.801	587.463	463.104	4191.82	4273.94
507.685	492.705	83.9233	169.373	33335.9	34961.7	72.4916	125.496
582.886	592.804	79.3457	94.6045	62.561	51.8799	70.0508	84.2336
606.538	550.082	3982.54	3882.6	241.089	283.813	284.353	278.581
561.523	542.45	679.016	413.513	4142.76	3637.7	4525.76	4209.9
524.902	582.886	588.989	501.251	32841.5	34493.3	262.451	296.021
4551.7	4465.48	4205.32	4511.26	290.943	291.033	3930.66	4074.46
4107.67	4506.68	4182.43	4118.35	325.114	301.288		
4412.84	4357.15	480.652	458.527	555.769	544.955		
4512.54	4295.18	4208.37	4059.6	80.3022	79.1063		