QIBA Metrology Committee: Work-In-Progress

Terminology Working Group – Technical Performance Working Group – Algorithm Comparison Working Group

Rationale for the QIBA Metrology Committee

Medical imaging is an effective tool for clinical diagnosis, staging, monitoring, treatment planning, and assessing response to therapy. In addition it is a powerful tool in the development of new therapies. The measurements of anatomical, physiological, and biochemical states of the body through medical imaging, so called quantitative imaging biomarkers (QIBs), are becoming increasingly used for clinical decision making and therapeutic development.

The mission of QIBA is to improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, patients and time, i.e., build "measuring devices" rather than "imaging devices". As "measuring devices" is it important to incorporate into our studies, metrology, which is the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology.

A biomarker is defined generally as an objectively measured indicator of a biological/ pathobiological process or pharmacologic response to treatment. We focus on quantitative imaging biomarkers, defined as imaging biomarkers that consist only of a measurand (variable of interest), or a measurand and other factors that may be held constant, and if the difference between two values of the measurand is meaningful. In some cases, the following additional requirement is considered -- there is a clear definition of zero such that the ratio of two values of the measurand is meaningful.

Each QIB requires a pre-defined computation algorithm, which may be simple or highly complex. In these working groups, members focus on QIBs generated from computer algorithms that may or may not require human involvement.

Terminology Working Group

Co-chairs: Larry G Kessler and Marina V Kondratovich; Members: Huiman X Barnhart, Michael Boss, Mary C Brady, Gregory Campbell, Kingshuk Roy Choudhury, Ross Filice, James J Filliben, Alexander R Guimaraes, Philip Judy, Mark Rosen, Daniel V Samarov, Alicia Toledano, Jingjing Ye, Zheng Zhang

The **QIBA Metrology Terminology Working Group** is preparing a document that will present the metrological terminology associated with measuring quantitative imaging biomarkers (QIBs) and changes in QIBs. To drive consistency in the field, the document will ultimately serve as input for, e.g., study design documents driving groundwork projects undertaken by QIBA.

The working group has drawn definitions from several sources, including documents from the International Organization for Standardization (ISO), the Clinical and Laboratory Standards Institute (CLSI), and the National Institute of Standards and Technology (NIST), and is relating them to QIBs. When there are discrepancies between sources, the working group aims to describe them and make recommendations. When providing examples, they will use International System of Units (SI units) and their abbreviations.

Consider, for example, the measurement of tumor volume, which may be of interest as a biomarker to describe response to therapy. Tumor volume itself is a quantity. From a process perspective, the working group will describe a measurement that will give a quantity value for the item of interest (i.e., tumor volume), referred to as a **measurand**. They propose using the International Vocabulary of Metrology (VIM) definitions:

- Measurement: The process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.
- Measurand: The quantity intended to be measured. All of the above falls under the umbrella of metrology.
- Metrology: The science of measurement and its application. As noted in VIM, this includes all aspects of measurement, whatever the measurement uncertainty and field of application.

Quantitative Imaging Biomarker (QIB): An imaging biomarker is quantitative if it consists only of a measurand (variable of interest) or if it consists of a measurand and other factors where all factors that are used to obtain the value of the imaging biomarker other than the measurand may be held constant, and if the difference between two values of the measurand is meaningful. In some cases, the following additional requirement is considered -- there is a clear definition of zero such that the ratio of two values of the measurand is meaningful. For example, tumor volume is a QIB because if one tumor has a volume of 0.5 cm^3 and another tumor has a volume of 1.5 cm^3, the following statements have real meaning: 1) the larger tumor is 1.0 cm^3 bigger than the smaller tumor; and 2) the larger tumor is 3 times the size of the smaller tumor. In another example, PET SUV is a QIB because all factors for obtaining its value (i.e., injected dose, body weight, and time of measurement, t) other than the measurand (i.e., concentration of radioactivity at time t) can be held constant, and the measurand is a ratio variable. Consider two tumors receiving the same ratio of injected dose to body weight. If one looks at SUV at time t and the tissue radioactivity concentration at that time (the measurand) is 5 MBq/kg for one tumor and 10 MBq/kg for the other tumor, then the following statements have real meaning: 1) the second tumor has 5 MBq/kg more radioactivity than the first tumor and 2) the second tumor has 2 times as much radioactivity as the first tumor.

The working group will not address non-quantitative imaging biomarkers, i.e., measures for which values have a magnitude, but neither the difference between two values nor the ratio of two values is meaningful, or measures for which values have no magnitude, and neither the difference between two values nor the ratio of two values is meaningful. Examples of such non-quantitative imaging biomarkers in mammography are the BI-RADS Assessment Categories 1 (Negative) through 5 (Highly suggestive of malignancy or BI-RADS features for mass margin (circumscribed, microlobulated, obscured, indistinct, and speculated) or shape (round, oval, lobular and irregular).

The working group is reviewing the various aspects and definitions involved in measuring a QIB. In metrology, it is said that no measured value is complete without an indication as to its **uncertainty**, which can be defined as a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand. Uncertainty combines many components. It may derive from the technical performance characteristics of the measure and/or the applicability of the measure to the clinical context for use. Some components of uncertainty arise from systematic effects, e.g., **bias**. Other components of uncertainty arise from random effects, e.g., **precision**. For these reasons, the QIBA Terminology Working Group recommends evaluating closeness of measurements made on the same experimental unit either via aggregated approaches or disaggregated approaches. Aggregated approaches use one parameter to summarize the uncertainty where disaggregated approaches use more than one parameter to summarize the uncertainty. Significant sources of uncertainty should be identified, and the parameter measuring any of these sources should be stated explicitly. It is not sufficient to say, e.g., "Uncertainty is 10%." It is better to say, e.g., "In this group of patients, the average coefficient of variation in PET SUV was 10%."

Such review, assessment, and recommendations on other terms as they relate to QIBs are also being conducted by the Terminology Working Group. These include **variability**, which can be defined as the tendency of the measurement process to produce slightly different measurements on the same test item, where conditions of measurement are either stable or vary over time. The working group therefore recommends evaluating separately the components of interest contributing to variability, explicitly identifying each source and stating the parameter being used to describe it (e.g., standard deviation, coefficient of variation).

The Terminology Working Group is also discussing the terms accuracy and trueness, and how they relate to **precision** and **bias**. Here, the working group recommends using **bias, and not accuracy**. Measuring bias requires knowing the truth. When the true value may or may not be known, there is a broader term called **agreement**, with a narrower term **reliability** being often used in practice. Usually an experiment for assessing the technical performance characteristics of a QIB is difficult (or impossible) to perform with multiple measurements of the same experimental unit; therefore, different experimental units at different time points are measured in the experiment and this experiment includes technical, and within-subject and between-subjects biological components of variance.

With QIBA's emphasis being on building "measuring devices" rather than "imaging devices", it is important that the QIB has the property of **linearity**. The working group is considering **limit of quantitation**, **measuring intervals**, and **measuring change**. Assessing the clinical performance of a biomarker, which is typically measured by its ability to predict a clinical outcome will be discussed in terms of **sensitivity** and **specificity**. To obtain a comprehensive picture of the diagnostic ability of the quantitative biomarker across the range of possible decision thresholds, one can use a **Receiver Operating Characteristic (ROC) curve**.

Technical Performance Working Group

Co-chairs: David L Raunig and Constantine Gatsonis; Members: Paul L Carson, David A Clunie, Patricia E Cole, Lori Dodd, Brian Garra, Mithat Gonen, Brenda F Kurland, Libero Marzella, Lisa M McShane, Kevin O'Donnell, Mary S Pastel, Gene Pennello, Nicholas Petrick, Adam J Schwarz, James T Voyvodic, Richard L Wahl, Gudrun Zahlmann

The **QIBA Metrology Technical Performance Working Group** is preparing a review of the metrological development of technical performance metrics and methods for quantitative imaging biomarkers. Quantitative biomarkers acquired from an imaging feature are comprised of continuous interval or ratio data that measure a specific feature of the image within a region of interest. These biomarkers differ from psychometric scales developed for imaging systems in that the final outcome is a measurement of the actual image and not an assessment of the disease severity by an experienced rater.

The objectives of the Technical Performance Group are to arrive at a reasonable consensus among clinical, technology and statistical imaging experts to establish the following:

- Performance metrics needed to measure and report technical performance -- Metrics and methodologies for performance assessment will be confined to those that are defined and accepted in metrology literature or widely accepted by the imaging community. While most if not all metrics and methodologies have their critics and champions, novel methods to measure performance will not be considered here for metrology use.
- Methodologies to arrive at those metrics -- Technical performance will not explicitly consider the clinical diagnostic or prognostic abilities of the biomarker. However, it has become evident that clinical performance cannot be completely ignored when assessing the biomarker technical performance in a clinical setting. Clinically relevant biomarker properties that are directly applicable to technical performance assessment are within the scope of this work. However, the association, or correlation, of the biomarker to clinical outcome will not be assessed for this effort.
- Study designs and considerations to arrive at meaningful and interpretable assessment of technical performance -- Study designs to assess quantitative biomarkers will be confined to acquiring the information and will not explicitly address design issues that may be more appropriate for a clinical drug trial. However, study designs will necessarily include many of the components that would follow the imaging modality into a clinical trial setting.

Technical performance is defined here as it applies to the ability to quantitatively measure a biological feature of the image as a single measure or as a measure of quantitative change. The ability to assess the performance of a quantitative biomarker is critical to ensure the consistent and reliable quality of that biomarker when used to measure a disease feature such as, for example, size, biological activity, pharmacodynamic parameter estimation or physiological function. The technical performance of each biomarker to measure its intended feature addresses three metrology areas that apply most directly to the use of these biomarkers to consistently measure the biological feature of interest.

Areas of Technical Performance	
Linearity - The strength of the linear relationship of the biomarker to a known or related standard reference	General consensus from this working group noted that a biomarker need not have a linear relationship with the measurand but the function that defines the biomarker measurement in terms of the measurand should be monotonic; and a calibration function should exist such that there is a unique one-to-one mapping of the measurand and biomarker over the biomarker range of interest and so can be repeated within the specified tolerance. As a consequence of no universally agreed upon definition of linearity, there is also no consensus on a single measure of linearity. For calibrated biomarker measurements that are calibrated to an accepted standard reference, there is some expectation for a direct linear response such that the biomarker vs. reference relationship is linear with slope=1 and intercept=0. The working group is considering performance metrics for linearity of quantitative biomarkers for calibration to a standard reference ("truth"), to an imperfect reference, and to a nonlinear/nontransformable function.
Repeatability - The measure of the biomarker performance to repeat the quantitative measurement on the same experimental unit	The working group is considering various aspects of repeatability, such as the "Test-Retest", along with repeatability metrics and study designs. Test-Retest, sometimes referred to as the "coffee-break test", is the replication of a quantitative measurement with all other effects, such as a morphological change in a lesion, being negligible with the goal of estimating the variance of a measurement that would be expected in the use claim. Test-retest results require that the measurement conditions be defined and may include the variability associated with other factors such as contrast, time, position and other variances that comprise the measurement variance of the biomarker. The working group is reviewing various methods to assess the repeatability of a QIB, including intra-subject variance, inter-subject variance, limits of agreement, repeatability coefficient (RC), intraclass correlation coefficient (ICC), concordance correlation coefficient (CCC), coefficient of variation (CV), and others.
Reproducibility - The measure of the biomarker performance to consistently measure image features in predetermined different clinical conditions	In the coming months, the working group will also be reviewing various methods to assess the reproducibility of a QIB, including the concordance correlation coefficient (CCC), variance components, and confidence bounds for variance, and summarizing their relevance and role in the technical assessment of QIBs.

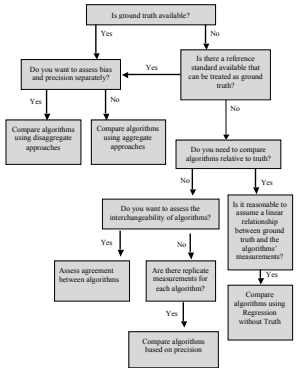
Algorithm Comparison Working Group

Co-chairs: Nancy A Obuchowski and Anthony P Reeves; Members: Tatiyana V Apanasovich, Daniel P Barboriak, Huiman X Barnhart, Andrew J Buckler, Alden A Dima, Brandon D Gallas, Maryellen L Giger, Robert J Gillies, Dmitry B Goldgof, Erich Huang, Edward F Jackson, Jayashree Kalpathy-Cramer, Hyun J Grace Kim, Paul E Kinahan, Qin Li, Kyle Myers, Lawrence H Schwartz, Daniel C Sullivan, Xiaofeng Wang

The **QIBA Metrology Algorithm Comparison Working Group** is preparing a paper on the validation and comparison of the algorithms used to produce the QIB results. Estimation errors in algorithm output can arise from several sources during both image formation and the algorithmic estimation of the QIB. These errors combine (additively or non-additively) with the inherent underlying biological variation of the biomarker. Studies are thus needed to evaluate the biomarker assay with respect to **bias**, defined as the difference between the average value of the measured biomarker and the true value, and **precision**, defined as the closeness of agreement between values of the measured biomarker on the same experimental unit.

There are several challenges in the evaluation and adoption of QIB algorithms. A recurring issue is the lack of reported estimation errors associated with the output of the QIB. One glaring example is the routine clinical reporting of PET SUVs with no confidence intervals. If patient disease progression versus response to therapy is determined based on changes of SUV ± 30%, then the need to state the SUV measurement uncertainties for each scan becomes apparent. Another challenge is the inappropriate choice of biomarker metrics, e.g. the use of tumor volume doubling time instead of tumor growth rate. Confidence intervals, or some variant thereof, are needed for a valid metrology standard; however, many studies inappropriately use tests of significance, e.g. p-values, in place of appropriate metrics. In addition, there is often a disconnect between what might be a superior metric statistically versus what is clinically accepted and what is considered clinically relevant. Finally, when potentially improved algorithms are developed, data from previous studies are often not in a form that allows new algorithms to be tested against the original data. Public image databases are being developed to provide a resource of documented images that may be used for computer algorithm evaluation and comparison.

Flow of analysis in comparing algorithms for QIBs



Types of QIBs -- When designing a study it is important to evaluate and report the correct measurement type. For example, in measuring lesion size there are at least three different measurement types: absolute size, a change in size, and growth rate. Each of these has a different measurand and associated uncertainty; characterizing one type does not mean that other types are characterized. A related issue is the suitability of a measurand for statistical analysis.

Measurement type	Parameters
Extent (e.g., volume)	Single image
Geometric form (e.g., set of locations comprising an object)	Single or multiple images
Geometric location (e.g., distance)	Single or multiple images
Proportional change (e.g., fractional change in area)	Two or more repeat images
Growth rate (e.g., proportional change per unit time in volume)	Two or more repeat images and time intervals
Morphological and texture features (e.g., circularity)	Single or multiple images
Kinetic response (e.g., K^{trans})	Two or more repeat images during the same session
Multiple acquisition protocols (e.g., ADC)	Two or more repeat images based on different protocols during same session

Studies on QIBs currently face two challenges compared to most other quantitative biomarkers: **human intervention** and a **lack of ground truth**. For many QIBs, human involvement in making the actual measurement is often permitted or required. In some cases fully automated measurement is possible; therefore, both approaches need to be considered in study designs. In patient studies ground truth is often not available even when histology or pathology tests are acquired. Even in the latter case there are well-known concerns with sampling errors relative to tissue heterogeneity and the non-quantitative nature of histopathology tests.

Phantoms and digital reference images will be simpler to measure than real images, and there is then ground truth. Testing with phantoms can establish a necessary minimum but cannot establish a sufficient performance level. A method will not be expected to perform better on real images than it does on phantoms. Zero-change sets may be able to characterize the bias and precision for the case when the change is zero. Again this establishes a minimum performance indication; bias may be higher and precision may be lower in the presence of a real change. Finally, it may be possible to use experienced markings in exceptional cases where computer assisted methods make obvious "errors" such as including a part of a vessel with a lesion. Further, phantoms may not represent all important imaging issues associated with real biological images.