QIBA Metrology Committee: Works-In-Progress

Terminology Working Group – Technical Performance Working Group – Algorithm Comparison Working Group
Meta Analysis Working Group – Case Example Working Group

Quantitative Imaging Biomarkers Alliance RSNA

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.

The Emerging Science of Quantitative Imaging Biomarkers Terminology and Definitions for Scientific Studies and for Regulatory Submissions

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

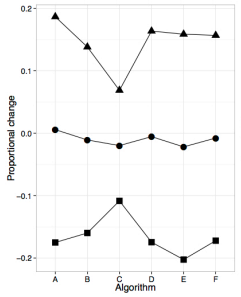
The **QIBA Metrology Terminology Working Group** is preparing a document that will present the metrological terminology associated with measuring quantitative imaging biomarkers (QIBs) and changes in QIBs. To drive consistency in the field, the document will ultimately serve as input for, e.g., study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

ABSTRACT: The development and successful marketing of quantitative imaging biomarkers (QIBs) has been hampered by the inconsistent and often incorrect use of terminology related to these markers. Sponsored by the Radiological Society of North America (RSNA), an interdisciplinary group of radiologists, statisticians, physicists, and other researchers worked to develop a comprehensive terminology that serve as a foundation for QIB claims. Where possible, this workgroup used existing national or international standards rather than invent new definitions for these terms. This terminology also serves as a foundation for the design of studies that evaluate the technical performance of QIBs and for studies that compare the algorithms that generate the QIBs from imaging devices that produce digital information. This paper provides examples of research studies and QIB claims that use terminology consistent with these definitions as well as examples of the rampant confusion in this emerging field. We also provide recommendations concerning additional science necessary to refine this field, and recommendations for appropriate terminological concepts. It is hoped that this document will assist reviewers of QIBs and inform guidance generated in the regulatory setting. More consistent and correct use of terminology could advance regulatory science, improve clinical research, and provide better care for patients who undergo imaging studies.

Statistical Issues in the Comparison of Quantitative Imaging Biomarker Algorithms using Pulmonary Nodule Volume as an Example

Nancy A Obuchowski, Huiman X Barnhard, Andrew J Buckler, Gene Pennello, Xiao-Feng Wang, Jayashree Kalpathy-Cramer, Hyun J Grace Kim, Anthony P Reeves

ABSTRACT: Quantitative imaging biomarkers (QIBs) are being used increasingly in medicine to diagnose and monitor patients' disease. The computer algorithms that measure QIBs have different technical performance characteristics. In this paper we illustrate the appropriate statistical methods for assessing and comparing the bias, precision, and agreement of computer algorithms. We use data from three studies of pulmonary nodules. The first study is a small phantom study used to illustrate metrics for assessing repeatability. The second study is a large phantom study allowing assessment of four algorithms' bias and reproducibility for measuring tumor volume and the change in tumor volume. The third study is a small clinical study of patients whose tumors were measured on two occasions. This study allows a direct assessment of six algorithms' performance for measuring tumor change. With these three examples we compare and contrast study designs and performance metrics, and we illustrate the advantages and limitations of various common statistical methods for QIB studies.



Example: Comparison of Quantitative Imaging Biomarker Algorithms using Pulmonary Nodule Volume

For the purposes of this paper, we have 9 cases in which the amount of actual change is unknown and 9 cases from a test-retest design where the subjects were re-measured after an interval of a few minutes and therefore there is no real change in the nodule size.

In the challenge instructions, participants were invited to download the CT images and to complete a spreadsheet in which the change in size measurement for the nodule in each image pair was recorded.

The limits of agreement (LOA) are illustrated in the figure on the left for six algorithms. Interestingly, Algorithm C, with the second highest population bias but smallest between-patient variability, has the narrowest LOA.

A Review of Statistical Methods for Technical Performance Assessment

David L Raunig, Paul L Carson, Patricia E Cole, Brian Garra, Constantine Gatsonis, Mithat Gonen, Marina Kondratovich, Brenda F Kurland, Lisa M McShane, Kevin O'Donnell, Gene Pennello, Nicholas Petrick, Adam J Schwarz, Daniel Sullivan, 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.

ABSTRACT: Technological developments and increased rigor in the quantitative measurement of biological features in medical images has given rise to an increased interest in using these quantitative imaging biomarkers (QIBs) to measure changes in these features. Critical to the performance of the QIB in a clinical setting 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 QIB for clinical use. It is therefore, difficult or not possible to integrate results from different studies. Sponsored by the Radiological Society of North America (RSNA), technical, radiological and statistical experts developed a set of technical performance analysis methods, metrics and study designs that will provide terminology, metrics and methods consistent with metrological standards. It is the hope of the authors that this document will provide a consistent framework that facilitates all QIB studies to be appropriately evaluated to facilitate the use of multiple studies to compare, contrast or combine results.

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.

Meta-Analysis of the Technical Performance of an Imaging Assay: Guidelines and Statistical Methodology

Erich P Huang, Xiao-Feng Wang, Kingshuk Roy Choudhury, Lisa M McShane, Mithat Gonen, Jinging Ye, Andrew J Buckler, Paul Kinahan, Anthony Reeves, Edward F Jackson, Alexandre R Guimaraes, Gudrun Zahlmann

ABSTRACT: Medical imaging serves many roles in the clinic, including surveillance, treatment response assessment, and evaluation for disease recurrence. Before an imaging assay, namely the process of acquiring an image and analyzing it to extract a quantitative imaging biomarker (QIB), an image feature relevant to the biomarker or underlying anatomic or biochemical quantity of interest, is put into routine clinical use, it needs to undergo evaluation to establish acceptable technical performance. Technical performance refers to the imaging assay's ability to measure the biomarker in terms of its bias, variability, and related performance metrics. Ideally, such an evaluation will encompass results from multiple studies to overcome the limitations of the typically small to moderate sample sizes of technical performance studies and/or to include a wider range of clinical settings and patient populations. We describe meta-analysis procedures to quantitatively summarize results from imaging assay technical performance studies, specifically steps to identify suitable studies, statistical methodology for carrying out the meta-analysis, and elements to include when reporting the results. Such a meta-analysis often presents some unique challenges. First, due to the smaller sample size of a typical imaging performance study and the form of many technical performance metrics, assumptions of approximate normality underlying standard meta-analysis techniques are not satisfied. Second, studies explicitly assessing an imaging assay's technical performance may be uncommon, making the accumulation of studies for the meta-analysis difficult. We also describe modifications to standard meta-analysis techniques to address these difficulties.

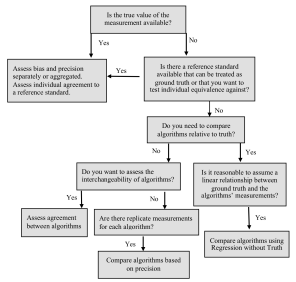
A Review of Statistical Methods for Computer Algorithm Comparisons

Nancy A Obuchowski, Anthony P Reeves, Erich Huang, Xiaofeng Wang, Andrew J Buckler, Hyun J Grace Kim, Huiman X Barnhart, Edward F Jackson, Maryellen L Giger, Gene Pennello, Alicia Y Toledano, Jayashree Kalpathy-Cramer, Tatiyana V Apanasovich, Paul E Kinahan, Kyle Myers, Dmitry B Goldgof, Daniel P Barboriak, Robert J Gillies, Lawrence H Schwartz, Daniel C Sullivan

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.

ABSTRACT: Quantitative biomarkers from medical images are becoming important tools for clinical diagnosis, staging, monitoring, treatment planning, and development of new therapies. While there is a rich history of the development of quantitative imaging biomarker (QIB) techniques, little attention has been paid to the validation and comparison of the computer algorithms that implement the QIB measurements. In this paper we provide a framework for QIB algorithm comparisons. We first review and compare various study designs, including designs with the true value (e.g. phantoms, digital reference images, and zero-change studies), designs with a reference standard (e.g. studies testing equivalence with a reference standard), and designs without a reference standard (e.g. agreement studies and studies of algorithm precision). The statistical methods for comparing QIB algorithms are then presented for various study types using both aggregate and disaggregate approaches. We propose a series of steps for establishing the performance of a QIB algorithm, identify limitations in the current statistical literature, and suggest future directions for research.

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.