QIBA Claim Guidance

Introduction:

This document provides guidance on how to develop and present the technical content of QIBA Profile Claims. The QIBA Profile Template document defines the location and format for such Claims.

QIBA Claims are summary statements of the technical performance of the quantitative imaging biomarker (QIB) being profiled. QIBA has adopted two kinds of claims:

- A **cross-sectional claim** describes the ability to measure the QIB at one time point
- A **longitudinal claim** describes the ability to measure change in the QIB over multiple time points.

QIBA Claim language is typically patient-centric rather than population centric. The performance describes the quantitative interpretation of a particular measurement of a feature in an individual patient (such as the size of a tumor or the stiffness of the liver or the aggregate tumor burden).

The **technical performance** of a QIB measurement is defined in terms of statistical metrics such as within-case standard deviation (**wSD**), within-case coefficient of variation (**wCV**), repeatability coefficient (**RC**) or reproducibility coefficient (**RDC**). In some cases, a claim can be written that states the technical performance of the QIB in simple terms, particularly its precision. These technical performance claims are particularly useful for researchers planning clinical trials. To express performance to clinicians in a clinically useful way, QIBA has currently settled on the 95% confidence interval (**CI**). See Glossary for definitions and considerations.

QIBA has not yet adopted **discriminatory claims**, which describe the ability of a QIB to distinguish groups of subjects (e.g. those with vs. without a particular disease, or those at different stages of disease). Such claims describe the clinical performance of a QIB by identifying one or more values of the QIB (i.e., cut-points) that discriminate the groups clinically and provides estimates of the sensitivity and specificity associated with each cut-point. Discriminatory claims are an area of active discussion within QIBA. They are potentially practical and appealing to QIBA's clinical audience; however they expand the scope of QIBA beyond the technical performance of biomarkers into clinical performance and might be significantly harder to prove. Although Profiles do not claim specific clinical performance, some do describe in the Discussion part of Section 2 discriminatory usage of biomarker values based on cut-points and performance assumptions made by users.

Steps in Developing a Claim:

Note that some amount of iteration over these claim development steps is to be expected. Groundwork findings, collected datasets and attempts to devise Profile requirements all lead to a greater understanding of the practical use of the biomarker and the associated Claims.

The recommended steps for developing a QIBA Claim statement are as follows [1]:

Step 0: Summarize Clinical Context / Use Case.

Summarize the primary intended Clinical Use Case(s) for the biomarker. A biomarker should inform one or more clinical decisions. The original proposal to form your biomarker committee will have relevant information you can use. This step is about refining that into statements that will drive development of a good claim.

- Decide: What clinical decision will the user of the biomarker make? What decisions are currently difficult due to the "fuzziness" of the finding?
- Know: What information is needed to make the decision?
- Measure: What do you need to measure to get this information? What is the imaging surrogate/finding that would drive a clinical decision? How will you determine that the measurement performance is adequate to make your decision? When would you change your decision/treatment/management?
- Method: How will you use the measurement to make the decision?

Example Summaries:

Amyloid PET Profile: The biomarker will measure beta amyloid deposition in the brain as a ratio of the tracer activity per tissue volume in several target regions compared to a reference region (SUVr) and is intended to be used to:

- Assess the efficacy of a therapeutic intervention as distinct from biologic age-relevant change, by comparing to a threshold change value.

CT Volumetry Profile: The biomarker will measure tumor volume and volume change (presence of growth, the amount of growth) of individual tumors and is intended to be used to:

- Interpret response, or lack thereof, to treatment.
- Quantify the amount of progression.

US SWS Profile: The biomarker will measure shearwave speed in liver tissue and is intended to be used to:

- Distinguish between mild and moderate fibrosis of the liver, which would drive the decision to initiate (expensive) antiviral therapy for Hep-C based on whether there is a good chance for the treatment to be effective. If severe, treatment is probably too late to be useful.
- Quantify the amount of progression, which would drive the decision on whether or how frequently to perform follow-up liver biopsies.

Step 1: Determine Type of Claim(s).

Based on the understanding described in Step 0, determine whether you need one or more of the following:

- Cross-sectional Claim
- Longitudinal Claim

A cross-sectional claim is represented by a confidence interval for the true value of a biomarker at a single time point. The true value is unknown, so the measured value and the uncertainty in the measurement are used to construct the confidence interval for the true value.

A longitudinal claim is represented by a confidence interval for the true change in the biomarker's value between two time points. The true change is unknown, so the measured value at the two time points and the uncertainty in the measurements are used to construct a confidence interval for the true change.

A Profile often has multiple claims, e.g., both a cross-sectional and a longitudinal claim for a single biomarker, or separate claims for different subpopulations when the performance of the biomarker would differ (See Step 3).

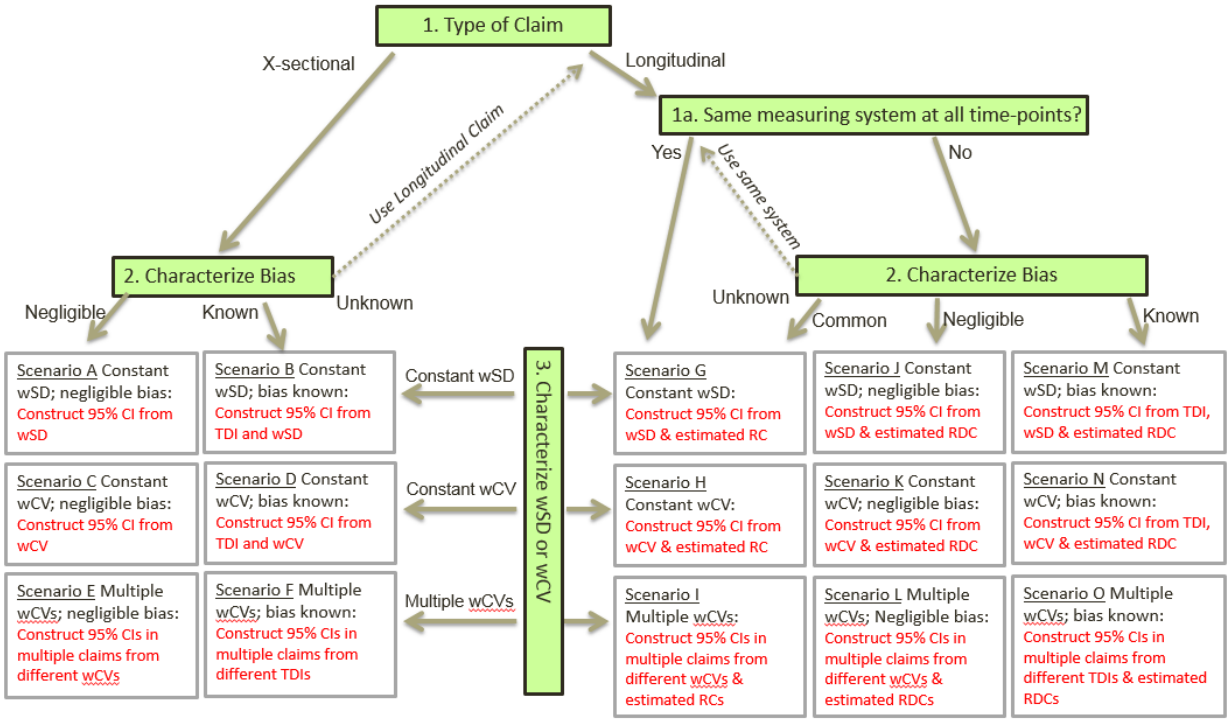
Step 2: Choose Metrics.

For each claim, the confidence interval (CI) needs to be constructed from one or more appropriate statistical metrics that quantify the uncertainty in the biomarker measurements. The choice of statistical metrics depends on:

- the type of claim
- whether the measurements tend to be biased or unbiased (i.e., do the measurements tend to systematically over-estimate or under-estimate the true value; see Glossary)
- whether the measurement uncertainty is constant or varies with the magnitude of the measurement.

Use the flowchart in Figure 1 to determine the appropriate statistical metrics. The characterizations described in Figure 1 (e.g., is there bias? is wCV constant?) will likely require carrying out QIBA groundwork studies, or referring to external studies if available. See [3,1] for guidance on designing and conducting such studies.

Figure 1: Selecting Metrics to Construct the 95% CI



Footnotes (see Glossary for terms and definitions):

- For some QIBs such as tumor volume, performance is characterized by the RC, estimated from a test-retest study performed over a very short period of time so that the tumor does not change. For other QIBs, such as SUVr to measure amyloid burden, performance is characterized by the RDC, estimated from a reproducibility study of healthy subjects' change in SUVr over several weeks or months.
- Characterizing precision with the wCV is only appropriate when the QIB is a ratio variable; it is not appropriate for interval variables.
- In the cross-sectional claim, negligible bias is average bias <5%. When the bias exceeds 5%, an estimate of the bias is needed for the claim (i.e., "known bias scenario").
- In the longitudinal claim, when different imaging equipment is used at the two time-points, the bias must be estimated. Sometimes the magnitude of the bias may be the same for the different imaging equipment ("common bias"); sometimes the bias is negligible (i.e., average bias <5%) for the different imaging equipment; and sometimes the bias of the imaging equipment differs but has been estimated (i.e., "Known bias scenario").
- The measuring system may incorporate multiple actor components (e.g., in CT Volumetry the variability of the measuring system is affected by the specific acquisition modality, radiologist and image analysis software). Therefore, changing one component (e.g., using different image analysis software) is effectively using a different measuring system. Further, characterizing the bias of a multi-component measuring system can get complex. Refer to the Section 2 Discussion of the CT Volumetry Profile for further details. Corresponding material may be added to this guidance in the future.

Step 3: Consider Subpopulations.

Technical performance (i.e., bias and/or precision) may vary depending on certain patient or feature characteristics. For example:

- Patients with head movement will have greater measurement variability for center of mass (in fMRI measurements).
- Spiculated tumors may be more difficult to measure (i.e., result in greater variability) than spherical tumors.
- Different organs (e.g., prostate, breast, liver) may display different technical performance for the same measurands.
- Different stages of disease may lead to different technical performance

If such characteristics are prevalent in the general population, you will need to consider one of following three approaches:

- Reflect the higher variability associated with the population variation in a single performance estimate and claim for the entire population
- Make separate performance estimates and claims for each subpopulation
- Exclude certain subpopulations from the Profile with appropriate bullets in the "holds when" text underneath the claim

If your groundwork data does not include adequate representation of a subpopulation, it will not be reflected in the performance estimate for a whole population claim, and neither will you have data to estimate performance for a separate subpopulation claim, so you will have to take the approach of excluding the subpopulation. Depending on the characteristic that defines the subpopulation it may be necessary to collect additional ground truth (which may or may not be available).

If a high level of performance is needed in order to be clinically useful (See Step 5), but is too difficult to achieve in the general population, it may make sense to start by limiting the

Profile to an identifiable "well-behaved" subpopulation for which the performance can be achieved. If a simple "universal" tool is more important and the clinically useful performance is not too high, it may make sense to have one claim for the broad population with a correspondingly lower technical performance that incorporates the broader variability.

The population(s) covered by a claim should be addressed in the "Holds when" part of the template. Also, consider adding corresponding checks in the QA Activity or the Patient Selection Activity sections of the Profile to confirm that those subpopulations are excluded.

Also consider the possibility of making separate Profiles for different subpopulations, which gives you the freedom to make different actor requirements that are appropriate or necessary for one subpopulation but not another. For example, CT tumor volumetry for screening (small nodules) has different requirements than tumor volumetry for advanced disease.

All of the above will depend on defining the subpopulations as clearly and unambiguously as possible.

Step 4: Estimate the Current Technical Performance.

Data from published papers and/or groundwork projects are used to estimate the current technical performance at typical sites (e.g. "current good practice") and perhaps the performance that would be reasonably achievable with the kind of improved practices the Profile could require. In order to get a reliable estimate of the QIB's precision, these published papers and/or groundwork projects should include at least 35 subjects [5]. In order to get a reliable estimate of the QIB's bias and assess its linearity property, a phantom study with at least 65 observations is needed [5].

This performance will be compared to the clinically useful performance values in the next step to understand if current practice is sufficient and just needs to be formalized, or whether improvements are needed to become clinically meaningful and, if so, how much improvement. It's even possible that current practice exceeds the needs and we might choose to either aspire to more advanced clinical usage or relax the practices.

The performance estimates will also inform the study design for groundwork projects, the appropriate sample sizes for conformance testing and whether to accept certain studies for use in meta-analysis.

Current performance might be expressed as a 95% confidence interval (CI) from a meta-analysis of published studies [6]. Alternatively, a range of values based on results from groundwork projects in QIBA or conducted by another outside group may be used to inform the claim. For example, for the Perc 15 Profile Claim for COPD, a meta-analysis was performed based on a synthesis of existing test-retest literature. From the meta-analysis a summary measure of the repeatability coefficient (RC) (i.e., a weighted average of the published studies on RC) was calculated and a 95% CI constructed for the summary measure. As another example, for the CT Volumetry Profile, multiple groundwork algorithm challenge projects were performed where various actors were invited to participate in studies involving a common set of images. The reproducibility coefficient (RDC) and bias were estimated from these studies under various scenarios (e.g., different lesion shapes, different subsets of actors) and the results were used to identify sets of plausible performance values [1].

Step 5: Determine the clinically useful performance values

The primary purpose of QIBs is to inform clinical decisions. What is the threshold of technical performance for the QIB to be clinically useful?

For example, ask: How small does tumor perfusion change need to be before medication is changed? How precise does the volume of a lung nodule need to be measured so you can discriminate suspicious nodules which might need to be biopsied from stable nodules which might need to be followed?

In some cases the performance that would be clinically useful might be based on informed judgment by experts. Surveying treating physicians to find what level of performance would make a difference to them may sometimes be possible. There is likely to be some interplay between the variability of the current measurements and identifying a definitive threshold for what is clinically significant. There may also be challenges with current clinicians not really using the quantitative measure yet. Some iteration should be expected. In other words, if the selected value does not produce the expected improvements in the quality and/or confidence of the clinical decisions, the value will be reexamined and revised.

Comparing the clinical requirements and the current technical performance gives a sense of how much work the committee is facing to achieve a viable biomarker. For example in the Perc 15 Profile Claim, the weighted average of the RC from published studies was 11 HU (and the 95% CI was from 4.5 HU to 18.4 HU). It was noted, however, that 11 HU represents a very small percent change in lung density. Clinical experts in the field advised that a value somewhat larger than 11 HU would be acceptable in the Profile claim statement [1]. For example, a value of 18 HU would be clinically useful and would fall within the 95% CI.

The clinical need is the ultimate driver: if the need allows for a low performance target, then set the requirements to be as inclusive as possible. If the need is much higher than current good practice, then that's what it is and the Profile should clearly set the bar that sites need to aspire to get that clinical utility.

Note that even if the current technical performance falls short of the desired clinical utility, it may still make sense to proceed with the Profile to clearly quantify the current state of the art and serve as a comparison for more advanced technologies or methods in the future.

Step 6: Consider Sample Sizes for Conformance Testing.

Whereas many of the requirements documented in the Profile are declaratory in nature, a subset of the requirements, and the assumptions underlying the claim itself, need an assessment procedure to demonstrate conformance.

For example, an image analysis workstation may be required to estimate the precision of its measurements and confirm they meet a certain target. For cross-sectional claims, the bias of the actors' measurements must be compared against the assumptions used in the claim statement. For longitudinal claims, the assumption of linearity must be assessed, along with estimates of the slope of a regression line of the measured vs. true biomarker values.

When the performance of an actor device can be expected to be much better than the required performance value, then a small sample size may be adequate to properly power the study to verify that the actor's imaging device conforms ~~with~~to the requirement. If an actor's imaging device has precision very close to the required performance value, then larger studies would be needed to reach adequate confidence that the actor meets the requirement.

For example, if groundwork studies have shown that the RC for most actors is about 7% and if the performance requirement in the Profile is set at 10%, then a study with 30 subjects is needed to test that the actor meets the profile requirements [1]. Alternatively, if the performance requirement in the Profile was set at 8%, then a study with nearly 200 subjects would be needed to show conformance of such actors.

So while a claim of better performance is appealing, it may come at the cost of more effort from each actor and site that must demonstrate conformance. It may make sense to set the performance claim slightly worse (as long as it is still adequate for the clinical utility) if it reduces the cost of assessing conformance.

Note that passing these assessment procedures is not itself sufficient to conform to the Profile. Actors must also conform to the other requirements in the specification tables. Of course if an actor can meet the assessment targets while violating specifications, then perhaps the Profile authors need to revisit those specifications.

For further details about what statistical assumptions need to be assessed to establish conformance and for standardized language, see "Guidance For Testing Actors Conformance With Statistical Assumptions Underlying The Claim".

Step 7: Choose Performance Value.

From the plausible range of technical performance in step 4, and taking into consideration the clinical needs in Step 5 and sample size requirements for testing conformance in step 6, experts from the fields of imaging physics and medicine now choose a reasonable performance value for each of the Claims.

For example, for the Perc 15 Profile Claim a change of 18 HU was chosen based on the fact that the clinical requirements do not demand detection of very small changes in lung density; furthermore, if most actors can show a RC near 11, then the sample size requirements for testing conformance are quite reasonable (i.e., a test-retest study of <17 cases is needed) [1].

Step 8: Construct Claim Text.

Claims should be kept reasonably brief, clear, statistically accurate and, ideally, be "parsable" by the clinicians and other stakeholders who will be using the Profile. Given the challenge of meeting all those goals, the exact wording of QIBA claims is still evolving, but the following examples are a good place to start.

Profiles will typically have several claims beginning with one stating the technical performance of the biomarker measurements as shown below. Additional claims about cross-sectional and longitudinal measurement confidence intervals may follow. It is

important to keep these as separate claims since the underlying statistical assumptions, which depend on the nature of each claim, can differ.

The examples below correspond to some of the scenarios identified above in Figure 1, which describe for each scenario the appropriate performance metric. The examples also show how the performance metric is used to construct the 95% confidence interval.

The examples also highlight some key issues to be mentioned in the Profile Discussion section that follows the Claims. See the Profile Template example text and Guidance comments for more information on the Claim Discussion section. Commonly the Discussion will describe the statistical metric, any statistical assumptions underlying the claim, how the claim might be applied to clinical interpretation, some realistic examples, a brief description of how the numerical values were estimated, etc. If the claim depends on things like the same imaging system being used at both time points, that should be stated as a requirement in the appropriate activity.

Cross-sectional and Longitudinal claims are generally preceded by the claim of Technical Performance on which they are based.

Technical Performance claims can use the following style:

"*Claim 1: A* *<QIB measurement (Y)>* ***has a within-****<subject>* *<performance metric>* ***of*** *<performance value>****."***

- Example of Scenario A – Constant wSD:
  Claim 1: An ADC measurement (Y) has a within-tumor standard deviation (wSD) of 2.55×10^{-4} mm^2/s.
  Holds when:
  - *measured in solid tumors greater than 1 cm in diameter or twice the slice thickness (whichever is greater)*

  Discussion*:*
  Claim 1 assumes that the wSD is constant over the range of relevant ADC values.

- Example of Scenario C – Constant wCV
  Claim 1: A lung tumor volume measurement (Y) has a within-tumor coefficient of variation (wCV) of 14%.
  Holds when:
  - *the longest in-plane diameter of the tumor is initially 10-34mm*

  Discussion*:*
  Claim 1 assumes that the wCV is constant over the range of relevant tumor volumes. Note that wCV is wSD/Y.

- Example of Scenario E – Multiple wCV:
  (Note: this can also be expressed as three claims in the form of Scenario C with different "Holds when" conditions)

Claim 1: A lung tumor volume measurement (Y) has a within-tumor coefficient of variation (wCV) that depends on the longest in-plane diameter category (see Table 2.1). Holds when:

- *the longest in-plane diameter of the tumor is 10-100mm*

Table 2.1 - wCV by Longest In-plane Diameter Category

Diameter	*10-34mm*	*35-49mm*	*50-100mm*
wCV	*0.141*	*0.103*	*0.085*

Discussion:
Claim 1 assumes that the estimated wCV is constant for tumors in each specified size range.

Cross-sectional claims can use the following style:

"Claim 2: A 95% confidence interval for the true *<QIB>* ***value is Y ±*** *<precision value>****."***

- Example of Scenario A – Constant SD:
  Claim 2: A 95% confidence interval for the true ADC value is Y ± 1.96×2.55×10^{-4} mm^2/s. Holds when:
  - *measured in solid tumors greater than 1 cm in diameter or twice the slice thickness (whichever is greater)*

  Discussion*:*
  Claim 2 assumes that there is no bias, the wSD is constant over the range of relevant ADC values, and replicate measurements are normally distributed.

- Example of Scenario C – Constant wCV:
  Claim 2: A 95% confidence interval for the true volume is Y ± (1.96×Y×0.14) mm^3
  Holds when:
  - *the longest in-plane diameter of the tumor is initially 10-34mm*

  Discussion*:*
  Claim 2 assumes that there is no bias, the wCV is constant over the range of relevant tumor volumes, and replicate measurements are normally distributed. Note that wCV is wSD/Y.

- Example of Scenario E – Multiple wCV:
  Claim 2: A 95% confidence interval for the true volume is Y ± (1.96×Y×wCV) mm^3. (See Table 2.1 for wCV)

  Discussion:
  Claim 2 assumes that there is negligible bias (i.e. <5%), the estimated wCV is constant for tumors in each specified size range, and replicate measurements are normally distributed.

Longitudinal claims can use the following styles:

"Claim 3: A true change (>0%) has occurred with 95% confidence if the measured change is Δ or larger."

"Claim 4: A 95% confidence interval for the true change is $(Y_2 - Y_1) \pm$ *<precision value>.*

- Example of Scenario G – Constant wSD:
  Claim 3: A true increase in the extent of emphysema has occurred with 95% confidence if the measured decrease in Perc15 without volume adjustment is 18 HU or more."

  Claim 4: A 95% confidence interval for the true change is $\Delta \pm (1.96 \times \sqrt{2} \times wSD)$*, i.e. [Δ-18 HU, Δ+18 HU].*

  Discussion:
  Claim 3 assumes that the wSD (within-subject) is constant over the range of relevant Perc15 values and replicate measurements are normally distributed.
  Claim 4 assumes that the wSD (within-subject) is constant over the range of relevant Perc15 values, the measurements possess the property of linearity, the regression slope of the measurements on the true values is nearly one, and replicate measurements are normally distributed. Note that for the wSD of 6.5, the repeatability coefficient (RC) is $(1.96 \times \sqrt{2} \times 6.5)$ *= 18HU.*

- Example of Scenario H – Constant wCV:
  Claim 3: A true change has occurred with 95% confidence if a measured increase/decrease is 39% or more.
  Holds when:
  - *the longest in-plane diameter of the tumor is initially 10-34mm*

  Claim 4: A 95% confidence interval for the true change is $(Y_2 - Y_1) \pm 1.96 \times \sqrt{(Y_1 \times 0.14)^2 + (Y_2 \times 0.14)^2}$.

  Discussion:
  Claim 3 assumes that the wCV is constant over the range of relevant tumor volumes and replicate measurements are normally distributed.
  Claim 4 assumes that the wCV is constant over the range of relevant tumor volumes, the measurements possess the property of linearity, the regression slope of the measurements on the true values is nearly one, and replicate measurements are normally distributed. Note that for the wCV of 0.14, the repeatability coefficient (RC) is (2.77 × 0.14 × 100) = 39%.

- Example of Scenario I – Multiple wCV:
  (Note: these can also be expressed as three claims in the form of Scenario H with different "Holds when" conditions)

Claim 3: A true change in volume has occurred with 95% confidence if a measured increase/decrease is more than the %RC (See Table 2.2), based on the diameter of the tumor at baseline.

Holds when:

- *the longest in-plane diameter of the tumor is 10-100mm at both timepoints*

Table 2.2 %RC by Longest In-plane Diameter Category

Diameter	*10-34mm*	*35-49mm*	*50-100mm*
%RC	*39%*	*29%*	*24%*

Claim 4: A 95% confidence interval for the true change is $(Y_2 - Y_1) \pm 1.96 \times \sqrt{(Y_1 \times wCV_1)^2 + (Y_2 \times wCV_2)^2}$*, where the wCV values at baseline and follow-up are given in Table 2.1.*

Discussion:

Claim 3 assumes that the wCV is constant within the ranges specified in the table and replicate measurements are normally distributed.

Claim 4 assumes that the wCV is constant within the ranges specified in the table, the measurements possess the property of linearity, the regression slope of the measurements on the true values is nearly one, and replicate measurements are normally distributed.

Step 9: Confirming Validity of Claim Statistical Assumptions.

Having settled on the nature of the biomarker and the type of claim selected, the corresponding statistical assumptions will have become clear, e.g., that the measurements have no bias, that the wSD is constant over the range of relevant measurement values, and that replicate measurements are normally distributed.

The validity of the Claim depends in part on the validity of those assumptions. The committee should plan on re-confirming the validity of those statistical assumptions. The necessary data for validation has likely already been collected during Profile groundwork or associated literature searches and meta-analysis.

In drafting the Profile, the committee should also add appropriate requirements and assessment procedures to the Profile for each site/actor to confirm the relevant assumptions as well (e.g., demonstrate linearity, estimate precision).

Further guidance on this topic will be published in the "Guidance For Testing Actors Conformance With Statistical Assumptions Underlying The Claim".

References:

[1] Obuchowski NA, Buckler A, Kinahan PE, Chen-Mayer H, Petrick N, Barboriak DP, Bullen J, Barnhart H, Sullivan DC. Statistical Issues in Testing Conformance with the Quantitative Imaging Biomarker Alliance (QIBA) Profile Claims. Academic Radiology 2016; 23: 496-506.

[2] Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers: terminology and definitions for scientific studies and for regulatory submissions. SMMR 2015; 24: 9-26.

[3] Raunig D, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. SMMR 2015; 24: 27-67.

[4] Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative Imaging Biomarkers: A Review of Statistical Methods for Computer Algorithm Comparisons. SMMR 2015; 24: 68-106.

[5] Obuchowski NA, Bullen J. Quantitative Imaging Biomarkers: Effect of sample size and bias on confidence interval coverage. SMMR 2017; *in press*.

[6] Huang EP, Wang XF, Choudhury K, McShane LM, Gonen M, Ye J, Buckler AJ, Kinahan PE, Reeves AP, Jackson EF, Guimaraes AR, Zahlmann G. Meta-analysis of the technical performance of an imaging procedure: Guidelines and statistical methodology. Meta-Analysis Working Group. Quantitative Imaging Biomarkers Alliance. Stat. Methods Med Res. 2014 May 28 (Epub). PMID 24872353.

***Glossary*:**

Bias: Bias is an estimate of systematic measurement error; it is the difference between the average (expected value) of measurements made on the same object and its true value. Percent Bias is Bias divided by the true value in percent. [2]

Interval variable: Measures for which the difference between two values is meaningful, but the ratio of the two values is not, are called interval variables. [2]

Precision: Precision is the closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar experimental units under specified conditions [2].

Quantitative Imaging Biomarker: (QIB) an objective characteristic derived from an in vivo image MEASURED on a ratio or interval scale as indicators of normal biological processes, pathogenic processes or a response to a therapeutic intervention.[2]

Ratio variable: A variable such that the difference between any two measures is meaningful and any two values have a meaningful ratio, making the operations of multiplication and division meaningful. A ratio variable possesses a meaningful (unique and non-arbitrary) zero value. [2]

Repeatability: Repeatability represents the measurement precision under a set of repeatability conditions of measurement. [2]

Repeatability condition of measurement: The repeatability condition of measurement is derived from a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same physical location, and replicate measurements on the same or similar experimental units over a short period of time [2].

Repeatability coefficient (RC): The least significant difference between two repeated measurements taken under identical conditions at a two-sided significance of α=0.05:

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

where s_w^2 is an estimate of σ_w^2, the within-subject variance. [3]

Reproducibility: Reproducibility is measurement precision under reproducibility conditions of measurement [2].

Reproducibility condition of measurement: The reproducibility condition of measurement is derived from a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects. [2]

Reproducibility coefficient (RDC): The least significant difference between two repeated measurements taken under different 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. [3]

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

Total deviation index (TDI): The difference, TDI_{π_0} satisfying the equation $\pi_0 = \Pr(|Y - X| < TDI_{\pi_0})$, where Y is the measurement of the QIB and X is the corresponding true value measurement. We usually set Π_o equal to 0.95. [4]

Within-subject coefficient of variation (*wCV*):

$wCV = \frac{\sigma_w}{\mu}$ where σ_w is the square root of the within-subject variance and μ is the mean of the measurements. [3]

Within-subject variance, σ_w^2: 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 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. [3]