Claims Guidance*

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

Claims are summary statements of the technical performance of the Quantitative Imaging Biomarker (QIB) being profiled. There are two kinds of claims: cross-sectional and longitudinal. A **cross-sectional claim** describes the imaging procedure's ability to measure the QIB at one time point, while a **longitudinal claim** describes the ability to measure change in the QIB over multiple time points.

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 stiffness of the liver) or an aggregate feature (such as tumor burden).

The **technical performance** of QIB measurements are 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**). See Glossary for definitions. QIBA has currently settled on the 95% confidence interval (**CI**) as an effective way to express performance to clinicians.

The steps for choosing technical performance values for the claim statements are as follows [1]:

Step 1: Choose Metric.
The choice of statistical metrics (See Figure 1) depends on:

- whether the claim is cross-sectional or longitudinal
- whether the imaging biomarker 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 QIB measurement variability is constant or varies with the magnitude of the measurement.

Step 2: Consider Variability Sources.
When technical performance is affected by patient or feature characteristics, and if these characteristics are prevalent in the general population, then the technical performance value used in the claim statement is often limited to apply only to the appropriate subpopulations. For example, Center of mass may be measured with greater variability in patients with head movement. For another example, spiculated tumors may be more difficult to measure (i.e. result in greater variability) than spherical tumors. If spiculated tumors are relatively common in the population, then the higher variability associated with measuring these tumors should be reflected in the claim. In some cases multiple claim statements may be needed to appropriately reflect different performance levels of the QIB depending on the patient/feature

Comment [OK1]: TODO Should we consider drilling down to describe the kind of Groundwork Study you might do to determine biased/unbiased.

Nancy: I think we should refer them to reference 1.
Kevin will add a note to that effect.

characteristics. The population assumed by the claim statement should be stated in the "Holds when" part of the template.

Step 3: Estimate the Range of Values of the Technical Performance.
Data from published papers and/or groundwork projects are used to estimate a range of technical performance values. This range might be the 95% confidence interval (CI) of the performance from a meta-analysis of published studies. Alternatively, this range might be based on results from groundwork projects in QIBA or conducted by another outside group. For example, for the Perc 15 Profile 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. For the CT Volumetry Profile, multiple groundwork 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 4: Consider Clinical Requirements.
After considering the estimated technical performance from Step 3, the clinical needs for the QIB performance are considered. For example, we 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 estimated so suspicious nodules are appropriately biopsied and stable nodules are followed? Comparing the clinical requirements and the estimated technical performance gives a sense of how much work the committee is facing to achieve a viable biomarker. When possible, these clinical needs are considered in determining the performance value for the claim. For example in the Perc 15 profile, the weighted average of the RC from published studies was 11 HU (and the 95% CI range 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.
Note that even if the estimated technical performance falls short of the clinical requirements, it may still make sense to proceed with the Profile based on the estimated performance to clearly quantify the current state of the art and serve as a comparison for more advanced technologies or methods in the future.

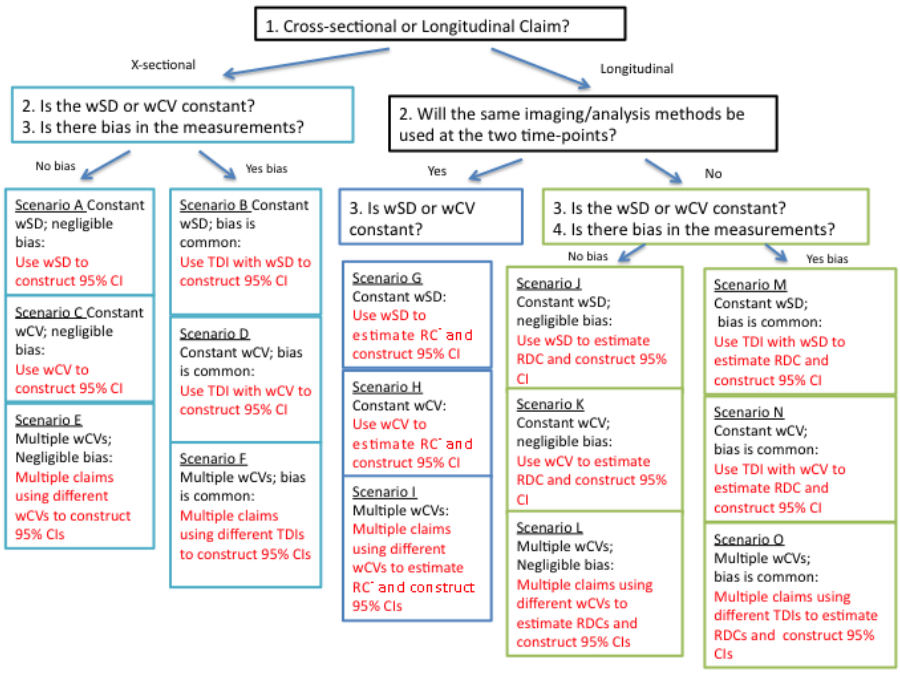
Step 5: Consider Sample Size for Conformance Test.
Whereas many of the requirements documented in the Profile are declaratory in nature, a subset of the requirements need to be demonstrated by a given actor which seeks to indicate that they conform. If an actor's imaging device has precision very close to the required performance value, then very large studies are needed to verify that the actor's imaging device conforms with the requirement. If an actor's imaging device has performance much better than the required performance value, then

Comment [OK2]: Add text clarifying what goes in as a profile requirement and what is part of the Holds When detail. Population vs Technique. The latter should be solvable by some "improved technique" the former can only be resolved by rejecting the data. Highlight the tradeoff of broadening/narrowing your performance vs population.
And when do you make multiple profiles. By prostate vs breast vs liver, and stage of disease.

smaller studies could be adequate. 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.

Step 6: Choose Performance Value.
From the plausible range in step 3, and taking into consideration the clinical needs and sample size requirements for testing conformance in steps 4-5, experts from the fields of imaging physics and medicine choose a reasonable performance value for the Profile. For example, for the Perc 15 profile a HU of 18 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].

Figure 1: Selecting a Performance Metric



Footnote:
See Glossary for terms and definitions.

Comment [OK3]: TODO
How do we express the Same/Same/Same issue?
At it's root it seems to be whether two passes of the same activity are done by the same actor or different actors.

- 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.

Cross-sectional claims should use the following style:

"*For a QIB measurement of Y units, a 95% confidence interval for the true QIB value is Y ± precision value*."

- Example 1 (Constant SD – Scenario A): *"For an ADC measurement of X mm^2/s in solid tumors greater than 1 cm in diameter or twice the slice thickness (whichever is greater), a 95% confidence interval for the true ADC value is X ± 5 $\times 10^{-10} mm^2/s$."*
  Note that "*5 $\times 10^{-10} mm^2/s$"* is equal to (1.96 × wSD), where wSD is the within-tumor standard deviation (2.55×10^{-10} here) and 1.96 is the 95% confidence factor. It is assumed that the wSD is constant over the range of relevant ADC values.

- Example 2 (Constant wCV – Scenario C): *"For a measured lung tumor volume of Y mm^3, a 95% confidence interval for the true volume is Y ± (1.96 × Y × 0.14)."*
  For some QIB measurements, such as tumor volumes, the precision varies with the magnitude of the measurement. In these cases, precision is often quantified by the wCV (wSD/Y). In this example the wCV=0.14 (or 14%). It is assumed that wCV is constant over the range of relevant tumor volumes.

- Example 3 (Look-up Table for wCV – Scenario E): *"For a measured lung nodule volume of Y mm^3, a 95% confidence interval for the true volume is Y ± (1.96 × Y × wCV)."* For some QIB measurements, such as tumor nodules, not only does the precision vary with the magnitude of the measurement, but we cannot assume that the wCV is constant. In these situations a look-up table is provided in the Profile which lists the wCV for various ranges of the measured QIB. The user must use the table to determine which wCV should be used based on the measured Y.

- Following each claim statement, there should be footnotes which describe
  - the statistical metric used in the claim,
  - the statistical assumptions underlying the claim, and
  - realistic examples illustrating use of the claim.
  - For example, one might say, "These claims are based on estimates of the within-tumor coefficient of variation (wCV) for nodules in this size range. In the claim statement the CI is expressed as Y ± 1.96 × Y × wCV. The

claim is based on the assumption that the wCV is constant for tumors in the specified size range and that there is negligible bias in the measurements (i.e. bias < 5%).

Longitudinal claims should use the following two-part style:

"A measured change in the QIB of Δ or larger indicates that a true change has occurred with 95% confidence"

and

"For a measured change of Δ, a 95% confidence interval for the true change is Δ ± precision value."

- Example 1 (Constant RC – Scenario G): *"A measured decrease in Perc15 of 18 HU or more without volume adjustment indicates that a true increase in the extent of emphysema has occurred with 95% confidence. For a measured change of Δ HU in Perc15 without volume adjustment, a 95% confidence interval for the true change is [Δ -18 HU, Δ +18 HU]."* Note that "18" is the Repeatability Coefficient, or (1.96 × √(2) × wSD). It is assumed that the wSD is constant over the range of relevant Perc15 values.

- Example 2 (Constant wCV – Scenario H): *"A measured change in the tumor's volume of Δ% indicates that a true change has occurred with 95% confidence if Δ% is larger than 38%"* and *"If Y_1 and Y_2 are tumor volume measurements at the two time points, a 95% confidence interval for the true change is* $(Y_2 - Y_1) \pm 1.96 \times \sqrt{(Y_1 \times 0.14)^2 + (Y_2 \times 0.14)^2}$. For some QIB measurements, such as tumor volumes, the precision varies with the magnitude of the measurement. In these cases, precision is often quantified by the wCV (wSD/Y). In this example, the wCV=0.14 (or 14%). Then the RC is (2.77 × wCV × 100) = 38%. It is assumed that wCV is constant over the range of relevant tumor volumes.

- Example 3 (Look-up Table for wCV – Scenario I): *"A measured change in the lung nodule's volume of Δ% indicates that a true change has occurred with 95% confidence if Δ% is larger than (2.77 × wCV × 100)"* and *"If Y_1 and Y_2 are the nodule volume measurements at the two time points, a 95% confidence interval for the true change is* $(Y_2 - Y_1) \pm 1.96 \times \sqrt{(Y_1 \times wCV)^2 + (Y_2 \times wCV)^2}$." For some QIB measurements, such as tumor nodules, not only does the precision vary with the magnitude of the measurement, but we cannot assume that the wCV is constant. In these situations a look-up table is provided in the Profile which lists the wCV for various ranges of the measured QIB. The user must use the table to determine which wCVs should be used based on the measured Y_1 and Y_2.

- Following each claim statement, there should be footnotes which describe
  - the statistical metric used in the claim,
  - the statistical assumptions underlying the claim,
  - the imaging methods used at the two time points, and
  - realistic examples illustrating use of the claim.
  - For example, one might say, "These claims are based on estimates of the within-nodule coefficient of variation (wCV) for nodules in this size range. For estimating the critical % change, the % Repeatability Coefficient (%RC) is used: $2.77 \times wCV \times 100$. The claim is based on the assumptions that the same imaging methods will be used at the two time points, the wCV is constant for nodules in the specified size range, and that the measurements follow the linearity property with slope equal to one (i.e. slope differs from unity by < 5%).

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 *in press*.

[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.

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 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 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 out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same physical location, and replicate measurements on the same or similar experimental units over a short period of time [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]