**Assessment Procedure Guidance**

This document provides guidance on assessment procedures to test the conformance of an actor to statistical assumptions underlying the Claim, and assessment procedures to test the composite performance of a site (e.g. to compare against the performance described in the Claim itself).

Profile Claims usually involve underlying statistical assumptions. For example, the claim may assume that the wCV (within-subject coefficient of variation) of a given measurement by an Actor is 10%. If an Actor's performance does not meet that assumption, it can invalidate the Claim even if the Actor satisfies all the other procedural requirements in the Profile. So it is important that the Profile include requirements to test the conformance of Actors to those statistical assumptions.

For example, a vendor of an image analysis workstation needs to assess the precision of the analysis software and confirm that it satisfies the assumption about precision used in the claim. If the claim assumes that the wCV is 10%, then the vendor needs to confirm that its wCV is <10% with 95% confidence. A statistical procedure must be described to test the hypothesis that the Actor’s wCV meets the Profile requirement at a specified type I error rate (usually 5%). It is not sufficient to show that the observed wCV is <10% for only a sample of cases.

Conformance with statistical assumptions is required by the QIBA process with increasing rigor at each QIBA Profile Stage. Specifically:

* At the Public Comment Stage (Stage 1), the assumptions must be clearly stated in the Profile.
* At the Consensus Stage (Stage 2), the procedures for assessing the statistical assumptions must be described in detail.
* At the Technically Confirmed Stage (Stage 3), the statistical assumption assessment procedures must have been performed and found to be reasonable.
* At the Claim Confirmed Stage (Stage 4), the actors must pass all requirements using the assessment procedures and show that the site meets the composite performance requirements of the Claim.

This guidance describes:

1. The statistical assumptions underlying different types of Claims so that authors of the Profiles know which assumptions need to be assessed.
2. Procedures appropriate for assessing the composite performance of a site. Testing of sites appears in the Profile in two places:
   1. The requirements (in Section 3 of the Profile) for the site to satisfy the assumptions.
   2. The procedure (in Section 4 of the Profile) for testing the metric that underlies the assumptions
3. The procedures appropriate for testing individual actors
   1. The requirements for each actor to satisfy the assumptions (in Section 3 of the Profile)
   2. Procedures to assess the metric that underlies the requirement (in Section 4 of the Profile)

# 1. Statistical Assumptions Underlying Claims

The statistical assumptions depend on the type of claim (see Table 1). For example, a cross-sectional claim assumes a within-subject precision and bias of an Actor. A longitudinal claim makes assumptions about the within-subject precision, property of linearity, and regression slope. If different imaging methods are allowed at each longitudinal time point, a constant bias is assumed. All these assumptions must be assessed and validated.

**Table 1: Statistical Assumptions for different Types of Claims**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Within-subject Precision**  **(See 2.1)** | **Bias**  **(See 2.2)** | **Property of Linearity**  **(See 2.3)** | **Regression Slope**  **(See 2.4)** |
| **Cross-sectional Claim** | X | X |  |  |
| **Longitudinal Claim**  (same imaging methods at both time points) | X |  | X | X |
| **Longitudinal Claim** (different imaging methods allowed at each time point) | X | X | X | X |

- Assessing a Technical Performance Claim (as stated in Section 2)

(e.g. Site is measuring … with a wCV … )

Note, we do not (yet?) have guidance for assessing longitudinal and cross-sectional claims (generally 95% confidence intervals) but we can assess the assumptions.

- Assessing individual Actor performance related to the assumptions underlying the Claim (Requirements in Section 3)

(e.g. Radiologist has a repeatability of X on test data)

# 2. Assessing Sites

This section provides guidance on procedures for assessing Sites, meaning the composite performance of the site in generating the biomarker measurements that are the subject of the Profile. Separate guidance for assessing individual Actors is provided in Section 3. An important distinction is that the guidance in this section will focus on the biomarker measurement which may be produced by the last Actor in the measurement "production chain" but the assessment is not of the performance of that Actor, but rather the performance of the entire chain.

This guidance will focus on assessing Technical Performance, in terms of things like wCV. The guidance does not currently address assessing 95% Confidence Intervals (which are used in longitudinal and cross-sectional claims) due to challenges in obtaining ground truth, performing retests involving radiation or contrast on patients, etc.

The following subsections will address each assumption in Table 1. Note that not every claim requires all of these assessment procedures.

Assessment procedures in Profiles will generally need to describe obtaining an assessment dataset, performing the statistical procedures, and determining when the requirement is satisfied.

Need to do the whole chain for Claim Confirmed.

## 2.1 Within-subject Precision Assessment Procedure:

Note that the Within-subject Precision of the Site is a measure of the composite performance of the entire system. Each of the Actors in the system may contribute imprecision to the measurement but for the Site Assessment Procedure it doesn’t really matter where the source of imprecision is as long as the total performance stays within bounds specified in the Profile. In contrast, in order to assess individual actors in the chain, the total imprecision will need to be "allocated" appropriately to each actor (e.g. scanner, radiologist, software, etc) in the form of Profile requirements, and assessed based on the guidelines in Section 3 of this document. <To prove/test the Site Technical Performance, would we like estimates of each actors wCV to help us design the study/assessment procedure?>

### 2.1.1 Test Dataset Guidance

Authors of QIBA Profiles have estimated the within-subject precision for their claims by either performing a meta-analysis or conducting groundwork studies. These studies were performed to populate the claim statements with realistic estimates of the precision.

Assessing clinical site composite performance needs a different kind of dataset. Ideally, a single sequestered dataset should be used, often designed from DROs or phantoms. If DROs/phantoms cannot be used, then it may be possible for each site to generate its own sample of patients’ test-retest images.

Desirable properties of a dataset for assessing composite precision:

* Has not been used for training algorithms
* Meets the requirements of the Profile, e.g. slice thickness, etc.
* Spans the scope of the Profile, i.e. represents the range of variability permitted in the Profile (e.g. severity, spectrum, patient comorbidities, tumor sizes)
* Easily accessible (i.e. located on QIDW)
* Replicate measurements can be ethically obtained (i.e. ?radiation/contrast considerations of test-retest??? )

Some examples from QIBA follow:

* In the amyloid profile, …..
* In the US SWS profile, ….
* In the CT volumetry profile, ….

Details about the conformance dataset and where to find it should be given in Section 4 of the Profile.

### 2.1.2 Procedure Guidance

Based on groundwork studies or the literature, you should have a good understanding of the characteristics of the precision of your biomarker (i.e. the biomarker’s precision profile). For example, you may know that the within-subject standard deviation (wSD) is pretty constant over the relevant range of the biomarker, or that the within-subject coefficient of variation (wCV) is pretty constant, or even that the wCV is pretty constant only in small ranges. Knowing the precision profile of your biomarker allowed you to decide how to formulate your claims, i.e. whether you needed a single or multiple claims, and whether you used the wSD or wCV.

Sites need to use the conformance dataset (described in Section 2.1.1) to construct a precision profile. In Section 4 of the Profile you will want to instruct the sites how to generate a precision profile so that you can evaluate the site’s precision relative to the assumptions you have made about the precision in the profile. You will need to use your expert opinion about what characteristics you want to stratify on and the metrics you want to use. Make sure you have sufficient sample size in each stratum (i.e. at least 5 cases). Here are some examples of specifications for the precision profile from various QIBA profiles:

* In the CT Volumetry Profile, sites must estimate the wCV separately for a group of 15 small and 16 large tumors, and also separately for lesions of different shapes.
* For the US SWS profile, sites must estimate the wCV for …

In Section 4 of the Profile you also need to describe the statistical method for estimating a site’s precision. This should include a description of what to measure (usually wSD or wCV), as well as the formulae for calculating precision. Since most claims characterize precision using the metric within-subject coefficient of variation (wCV) and/or the repeatability coefficient (RC), boiler-plate language is given here.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For each case, calculate the *<name of QIB here>* at time point 1 (denoted Yi1) and at time point 2 (Yi2) where *i* denotes the *i*-th case. For each case, calculate: . Calculate: . Estimate the % Repeatability Coefficient as .

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### 2.1.3 Calculate the maximum allowable variability:

In section 3 of the Profile you must specify the maximum allowable within-subject variability, in other words, the maximum wCV that the site can have for the conformance dataset. This is the maximum test-retest variability that a site can have and still satisfy the claim with 95% confidence. This is not simply the wCV used in the claim statements because we need 95% confidence that the site meets the claim. Therefore, the site must have a wCV estimate that is actually lower than the wCV used in the claim.

The maximum test-retest variability depends on several factors:

1. The number of subjects in the conformance dataset (described in section 2.1.1), and
2. The estimate of precision used in the Profile claim.

For example, in the CT Volumetry Profile, the conformance dataset has N=31 cases with test-retest data. In the Profile, a Repeatability Coefficient (RC) of 21% is claimed. Given the sample size and the RC from the claim, it can be determined that a site’s estimated RC must be <16.5% in order to be 95% confident that the precision requirement is met. Thus, 16.5% is the maximum allowable wCV for a site and is specified in section 3 of the Profile.

Calculation of this maximum allowable variability is described in Appendix A; you can also consult a statistician for calculating this value. Note that when you have a large conformance dataset, the maximum allowable variance will be just slightly smaller than the wCV used in the claim statements; in contrast, when the conformance dataset is small, the maximum allowable variance will necessarily be much smaller than the wCV used in the claim statements in order to achieve 95% confidence. Profile authors will need to strike a balance between the size of the conformance dataset and the maximum allowable variance in order for the sample size to be of a practical size yet the maximum allowable variance to be sufficiently large.

In addition, in Section 3 of the Profile you should also specify the maximum allowable within-subject variability for each of the strata specified in the precision profile (e.g. group of small nodules and group of large nodules). Profile authors should use their discretion in deciding on the maximum allowable variability for each stratum because usually the sample size in each stratum is small and not amendable to statistical constraints. For example, in the CT Volumetry Profile, must be < 21% for each size subgroup in order for the conformance requirement to be met.

## 2.2 Bias Assessment Procedure

The following procedures are recommended for assessing the bias.

***Step 1 - Procedure for testing the assumption:*** First, identify a test dataset for evaluating actors’ bias. A phantom study is ideal for assessing bias because ground truth is known. Measurements should be taken at multiple values over the relevant range of the true value. Ideally, 10 nearly equally-spaced values should be chosen. For example, in the CT Volumetry Profile, the previously designed FDA Lungman phantom is described. Lungman phantom has 42 distinct target tumors. The Profile specifies the number and range of lesion characteristics to be measured (sizes, densities, shapes).

Second, specify the methods for generating a bias profile. A bias profile is a description of the bias at different magnitudes of the measurand. For example, in the CT Volumetry Profile, actors must stratify the cases by shape. For each stratum actors estimate the population bias.

***Step 2 - Boilerplate statistical language:*** Describe the method for estimating an actor’s bias. This should include a description of how and what to measure (the measurand), as well as the formulae for calculating bias and its 95% CI.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For each case, calculate the value of the measurand*<name of QIB here>* (denoted Yi), where *i* denotes the *i*-th case. Calculate the % bias: , where Xi is the true value of the measurand. Over N cases estimate the population bias: . The estimate of variance of the bias is . The 95% CI for the bias is , where is from the Student’s t-distribution with =0.025 and (N-1) degrees of freedom.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Step 3 – Requirement for satisfying the assumption:*** Specify the number of cases needed to measure the bias in order to construct tight Confidence Intervals (CIs) on the bias. For example, in the CT Volumetry Profile, it was decided that each tumor in the FDA Lungman phantom would be measured twice (N=82) in order to put a tight (+1%) CI around the bias. An actor’s CI must lie completely in the interval -5% to +5% for the conformance requirement to be met. (See Appendix B to determine the sample size needed for various widths of CIs.)

For the bias profile, the conformance requirements might be looser (unless there is a sufficient sample size for each subgroup). For example, in the CT Volumetry Profile, the estimated *popbias* (not the lower and upper bounds of a CI) must be between -5% and +5% for each stratum in order for the conformance requirement to be met.

## 2.3 Linearity Assessment Procedure

The following procedures are recommended for assessing the property of linearity.

***Step 1 - Procedure for testing the assumption:*** Identify a test dataset for evaluating the property of linearity. A phantom study is ideal for assessing linearity because ground truth is known, or at least multiples of ground truth can be formulated. Measurements should be taken at multiple values over the relevant range of the true value. Ideally, 5-10 nearly equally-spaced measurand values should be chosen with 5-10 observations per measurand value (a total of 50 measurements is recommended).

***Step 2 - Boilerplate statistical language:*** Describe the method for assessing the property of linearity. This should include a description of how and what to measure.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For each case, calculate the *<name of QIB here>* (denoted Yi), where *i* denotes the *i*-th case. Let Xi denote the true value for the i-th case. Fit an ordinary least squares (OLS) regression of the Yi’s on Xi’s. A quadratic term is first included in the model to rule out non-linear relationships: . If then a linear model should be fit: , and R2 estimated.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Step 3 – Requirement for satisfying the assumption:*** The estimate of should be <0.50 and R-squared (R2) should be >0.90.

## 2.4 Regression Slope Assessment Procedure

The following procedures are recommended for estimating the regression slope.

***Step 1 - Procedure for testing the assumption:*** Identify a test dataset for evaluating the property of linearity. A phantom study is ideal for estimating the slope because ground truth is known, or at least multiples of ground truth can be formulated. Measurements should be taken at multiple values over the relevant range of the true value. Ideally, 5-10 nearly equally-spaced measurand values should be chosen with 5-10 observations per measurand value (a total of 50 measurements is recommended).

***Step 2 - Boilerplate statistical language:*** Describe the method for estimating the slope. This should include a description of how and what to measure.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For each case, calculate the *<name of QIB here>* (denoted Yi), where *i* denotes the *i*-th case. Let Xi denote the true value for the i-th case. Fit an ordinary least squares (OLS) regression of the Yi’s on Xi’s: . Let denote the estimated slope. Calculate its variance as , where is the fitted value of Yi from the regression line and is the mean of the true values. The 95% CI for the slope is .

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Step 3 – Requirement for satisfying the assumption:*** For most Profiles it is assumed that the regression slope equals one. Then the 95% CI for the slope should be completely contained in the interval 0.95 to 1.05.

# 3. Assessing Actors

<Start by trying to determine which actors/activities are the major contributors of variance>

First, identify a test dataset for evaluating actors’ precision. For example, in the CT Volumetry Profile, a published test-retest dataset of 31 subjects with lung lesions, is described in the Profile, along with directions for obtaining the data.

Having test-retest is ideal (to have known ground-truth for zero change). This can be difficult when dealing with radiation or administered contrast>

<Cite the test-retest MR Paper>

<What "qualifies" as test-retest? Judgement call on whether change in the biomarker could be expected, e.g. lesion size. Generally minutes/hours/some days>

<But we know you may need to make do with what is available. Consult statistician about what you are using to flag any potential limitations due to using that dataset>

<Consider pointing to NCIA, QIDW, groundwork, etc. as possible sources of datasets>

<When testing for bias, it may be better to use phantoms due to the known groundtruth, but for precision it might be better to use human data even if the ground truth is less/unknown. If you have to use phantoms for precision you should probably set a tighter performance target since you would expect better performance on phantoms.>

If a clinical test-retest dataset is not available, another option is to generate DRO data to simulate clinical test-retest variability. <Then the hybrid of Duke's synthetic lesions in human scan data>

Still another option might be to require vendors to design their own test-retest study, recruit patients for the study, and then measure precision. For example, the MRE profile is considering this approach.

# References:

[1] Obuchowski NA, Buckler A, Kinahan P, 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] Obuchowski NA, Bullen J. Quantitative Imaging Biomarkers: Coverage of Confidence Intervals for Individual Subjects. *Under review at SMMR*.

[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.

# Appendix A:

Let the RC in the claim statement be denoted δ. Let θ denote the actor’s unknown precision. We test the following hypotheses:

versus .

The test statistic is: . Conformance is shown if , where is the α-th percentile of a chi square distribution with N dfs (α = 0.05). So, to get the maximum allowable RC (step 3), first look up the critical value of the test statistic, in a table of chi square values. Then solve for in the equation:

.

For example, in the CT Volumetry Profile, N=31 and δ=21%. = 19.3 from http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm. Then, solving for , we get the maximum allowable RC of 16.5%. Thus, an actor’s estimated RC from the Sloan Kettering dataset must be <16.5%.

# Appendix B:

Different Profiles will have different requirements for the bias. Some Profiles assume there is no bias, in which case the 95% CI for an actor’s bias should be totally contained within the interval of -5% and +5%. Other Profiles may allow actors to have some bias, so the Profile will specify an upper limit on the bias. In these Profiles, the 95% CI for an actor’s bias should be less than the upper limit on the bias.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Width of 95% CI for Bias** | | | | |
|  | **+ 1%** | **+ 2%** | **+ 3%** | **+ 4%** | **+ 5%** |
| **Varb**\***=5%** | 22 | 8 | <5 | <5 | <5 |
| **Varb=10%** | 42 | 13 | 7 | <5 | <5 |
| **Varb=15%** | 61 | 17 | 9 | 7 | <5 |
| **Varb=20%** | 80 | 22 | 12 | 8 | 6 |
| **Varb=25%** | 99 | 27 | 14 | 9 | 7 |

\*The variance is represented here as the between-subject variance divided by the bias.

For example, for a tight CI of +1%, the sample size requirements vary from 22 to 99 depending on the between-subject variability. If the between-subject variability is unknown, it is wise to consider larger values. When the variance between cases is 20%, 80 cases are needed for a tight +1% CI around the bias.