**Assessment Procedure Guidance**

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

**Rationale**

Profile Claims usually involve underlying statistical assumptions. For example, the claim may assume that the wCV (within-subject coefficient of variation) of a given Actor measurement is 10%. If an Actor 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 measure its software’s precision 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 test that its wCV is <10%.

Conformance with these statistical assumptions is required with increasing rigor at each QIBA Profile Stage. Specifically, at the Consensus Stage (Stage 2), the procedures for testing the statistical assumptions must be described in detail in the Profile. 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 demonstrate successful results from the statistical assumption assessment procedures, <and the site must measure the composite performance and see that it is consistent with 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; and
2. The type of procedures appropriate for testing each assumption. Testing appears in the Profile in two places:
   1. The requirement (in Section 3 of the Profile) for an actor to satisfy the assumption.
   2. The procedure (in Section 4 of the Profile) for testing the metric that underlies the assumption
3. <The type of procedures appropriate for testing the composite performance of a site which can then be compared to the target in the Claim>

# 1. Statistical Assumptions Underlying Claims

The statistical assumptions depend on the type of claim. For example, for a cross-sectional claim an assessment of actors’ within-subject precision and bias must be performed.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Within-subject precision** | **Bias** | **Property of Linearity** | **Regression slope** |
| **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 |

# 2. Statistical Assumption Assessment Procedures

Each of the assessment procedures described here will generally

## 2.1 Within-subject Precision Assessment Procedure:

The following procedures should be used for assessing the within-subject precision. (If they do not seem applicable, consult the QIBA Metrology subject matter experts).

<Do we need to provide guidance to Profile authors about "apportioning" the wCV contributions of each actor and then decide which actors need to have their contribution assessed? The Claim represents the wCV of the entire system composed of all actors. How do you apportion? How do you "prioritize" which actors should be assessed? (It might be impractical or unacceptable workload to assess certain actors and it's easier to make a generous assumption and not test it) It depends in part which are the biggest contributors.>

<All of the profiles have looked at wCV with a meta-analysis or groundwork to get the estimate of the system. Doesn’t really matter where the source of imprecision is as long as the total stays within bounds. What do you do

<Need explanation of how we're not measuring the entire system wCV (but will need to some day), but we've assumed the analysis software is the main focus and so we feed it known data that likely has "typical variance" for the scanner, and we measure and test the analysis software result.>

<US did phantom measurements and likely operator variability will be the main source in patients (and eliciting patient compliance with breath hold)>

<pull in SMEs (Anne Singer – amyloid – divided up imprecision)(Mike Boss – MR) to describe how they decided>

***Step 1 – Prepare a Test Dataset***

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.

<Ideally you don't want the algorithms to have been trained on the dataset>

<The dataset should meet the requirements of the profile, e.g. slice thickness etc.>

<Should be representative of the scope of the profile. Ideally it should exercise the range of variability permitted in the profile, e.g. if 10-120mm lesions, have that range in the test dataset. The most significant sources of variability should be present.

The cases in the test-retest study should match the target cases in terms of distribution of cases expected to be encountered by users of the profile. Spectrum/severity/comorbidities>

<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.

***Step 2 – Specify a Test Procedure***

Second, specify the methods for generating a precision profile. A precision profile is a description of the precision at different magnitudes of the measurand. For example, in the CT Volumetry Profile, actors must estimate the RC using the data from all 31 subjects, and also separately for the 15 smallest tumors and for the 16 largest.

<Based on groundwork you may sometimes be able to determine that your precision profile is relatively flat over the range of interest in the profile, so you might be able to skip doing a full profile, but usually it will be useful to do one anyway. How many "points" on the profile you evaluate is a judgement call. Note that having more points expands the needed dataset size and may increase the testing work (decreasing the practicality).>

***Step 2 - Boilerplate statistical language:*** Describe the method for estimating an actor’s precision. This should include a description of how and what to measure, as well as the formulae for calculating precision. Since most claims characterize precision using the metric within-subject coefficient of variation (wCV), the formulae for this metric are 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 .

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

***Step 3 – Requirement for satisfying the assumption:*** Specify the maximum allowable within-subject variability. This is the maximum test-retest variability that an actor can have and still satisfy the claim. The maximum test-retest variability depends on the number of subjects in the test dataset, the estimate of precision used in the Profile claim, and the actor’s (unknown) precision when following the Profile. For example, in the CT Volumetry Profile, the Sloan Kettering 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 an actor’s estimated RC must be <16.5% in order to be 95% confident that the precision requirement is met. (See Appendix A for how to calculate the maximum allowable variability.)

For the precision profile, the conformance requirements might be looser (unless there is a sufficient sample size for each subgroup). In the CT Volumetry Profile, must be < 21% for each size subgroup in order for this 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.

**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.

**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.