**Guidance for QIBA Profile Authors
Drafting Actor and Composite Performance Assessment Procedures**

This document is one of several guidance documents related to the QIBA Profile Development Process. Readers are recommended to review the GUIDANCE comments in the QIBA Profile Template, and the contents of the Claim Guidance document since that material provides relevant background which is not re-stated here.

This document provides guidance on drafting assessment procedures to test:

* the conformance of an actor, and
* the composite performance of a site (e.g. to compare against the performance described in the Claim itself).

Assessing composite performance is central to achieving Stage 4 (Claim Confirmed) and Stage 5 (Clinically Confirmed), but may or may not be part of site conformance to a given Profile.

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. Thus, it is important that the Profile include requirements to test the conformance of Actors to those statistical assumptions. Furthermore, conformance to the statistical assumptions should be shown at the 95% confidence level. For example, a vendor of an algorithm (implemented in an image analysis workstation) needs to assess the precision of the algorithm 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 Actor (the algorithm) has a wCV <10% with 95% confidence. An assessment 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 those procedures have been found to be reasonable.
* At the Claim Confirmed Stage (Stage 4), the actors must pass all requirements using the assessment procedures, the composite performance assessment procedures must be described in detail and sites must meet 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. (See Section 1 of this document)
2. Procedures, and requirement pass-fail thresholds, appropriate for assessing the composite performance of a site (See Section 2 of this document).
3. Procedures, and requirement pass-fail thresholds, appropriate for testing individual actors (See Section 3 of this document).

The Profile is affected by (2) and (3) in two places:

1. Pass-fail thresholds, for each actor to satisfy the assumptions, appear as requirements in Section 3 of the Profile
2. Assessment procedures, to generate the metric values that will be compared against the thresholds, appear in Section 4 of the Profile

When related Profiles use the same performance metric, assessment procedures can often be re-used, albeit with different pass-fail thresholds depending on the needs of the Profile.

This guidance uses several specific terms:

* Measurand –the quantity being measured, e.g. the volume of a tumor
* Measurement – an assessed value, which often includes variability and/or bias, for a measurand
* Ground Truth – the true value of a measurand for a given measurement
* Algorithm – the image analysis that generates a measurement; typically implemented in software as part of an analysis workstation

# 1. Statistical Assumptions Underlying Claims

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  **Type of Claim** | **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 |

**Figure 1** illustrates the process of making a biomarker measurement, from prepping the subject to recording the final biomarker measurement. The process typically involves multiple actors (e.g. scanner, image processing, reader, algorithm). ***Composite performance*** refers to the total precision and total bias of the final biomarker measurements, regardless of the source of the imprecision and bias. We discuss the assessment of composite performance in Section 2. In contrast, ***actor performance*** refers to the precision and bias of a particular actor in a particular activity (shown as a gray box in the diagram) in the biomarker production process. We discuss the assessment of actors’ performance in Section 3. Also in Section 3 we discuss how to determine actors’ portion of the total variability and total bias. We discuss two general approaches: In a ***top-down*** ***approach*** we would take the known composite performance and partition the total variability and total bias into the contributing sources (e.g. individual actors). In a ***bottom-up*** ***approach***, we would measure the performance of individual actors in the biomarker production chain and then pool the actors’ variability and bias to estimate the composite performance.



**Figure 1: Actor/Activity Performance vs Composite Performance**

# 2. Assessing Sites (Composite Performance)

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 is produced by the last Actor in the measurement "production chain" but the assessment is not of the performance of that Actor alone, but rather the performance of the entire chain up to and including that Actor.

This guidance will focus on assessing technical Claims, as well as the assumptions underlying the cross-sectional and longitudinal Claims, for example the claimed precision in terms of the wCV.

The guidance does not currently address whether the 95% Confidence Intervals (which are used in longitudinal and cross-sectional Claims) are performing at the nominal level of 95%. Such assessments face 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.

**Site Assessment Datasets**

Assessing a clinical site’s composite performance needs a different kind of dataset. Ideally, patient datasets acquired at the site should be used. In principle, the dataset can be obtained from patient scans, phantom scans or digital reference objects (DROs). Patient studies face a challenge in getting reliable ground truth. Since phantom studies bypass the influence of variability and bias in the patient and patient handling, it would be good to first determine that those early activities are not expected to be a source of bias before considering phantom data for site assessment. Finally, DROs bypass the variability and bias in the modality acquisition process, but if it is not possible to come up with a reasonable phantom, then a DRO may be a reasonable alternative, although one should consider allowing for some bias in the activities that have been bypassed in the test data. In addition, it is conceivable that different types of data (patient, phantom, DRO) may be used for testing different characteristics as part of an overall site assessment. For example, phantom data may be use for bias and linearity assessment while patient scan are used as part of precision assessment.

Ultimately the choice between these different approaches will be a judgement call for the Biomarker Committee developing the Profile. The Biomarker Committee is encouraged to record the basis for their judgement in the Profile.

A phantom study is often well suited to assessment of Site Bias, Linearity and Regression Slope due to being able to control the range of ground truth values and the ability to get multiple measurements at the same ground truth value.

Some biomarkers may have a corresponding "reference standard" that is reasonable to use for patient scans. For example, in the case of MR Proton Density Fat Fraction, in-vivo spectroscopy is not a gold standard but is an acceptable reference measure of lipid fraction in liver, especially for bias assessment. That said, the MR Committee has recently moved away from in-vivo spectroscopy and back to a phantom-based assessment.

## 2.1 Assessment Procedure: Site Within-subject Precision

The Within-subject Precision of the biomarker measurement at a 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 the bounds specified in the Profile. In contrast, Assessment Procedures for individual actors (which will be discussed in Section 3 of this document), will require that the total imprecision eventually is "allocated" appropriately to each actor (e.g. scanner, radiologist, algorithm, etc.) and that required performance documented in the form of a Profile requirement for that actor.

### 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. Ideally, a dataset of test-retest subjects scanned at the site with measurements made by readers at the site using algorithms available at the site should be used. See also the Site Assessment Dataset discussion in section 2.

Desirable properties of a dataset for assessing composite precision are that it:

* Has not been used for training the algorithm being tested
* 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, tumor morphology
* Is easily accessible, e.g. located on QIDW
* Allows for replicate measurements to be ethically obtained, e.g., radiation/contrast considerations of test-retest

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

### 2.1.2 Procedure Guidance

Based on groundwork studies or the literature, a good understanding of the characteristics of the precision of their biomarker (i.e. the biomarker’s precision profile) should be obtained. For example, one 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 a biomarker allowed one to decide how to formulate Claims, i.e. whether a single or multiple Claims are needed, and whether the wSD or wCV is used.

Sites need to use the conformance dataset (described in Section 2.1.1) to construct a precision profile. In Section 4 of the Profile one will want to instruct the sites how to generate a precision profile so that the site’s precision can be evaluated relative to the assumptions made about the precision in the Profile. One will need to use expert opinion about what characteristics to be stratified on and the metrics to be used. Make sure sufficient sample size is obtained in each stratum (i.e. at least 5 cases). For example, 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.

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

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

For each case, the assessor shall calculate the *<name of QIB here>* for the first replicate measurement (denoted Yi1) and for the second replicate measurement (Yi2) where *i* denotes the *i*-th case. For each case, the assessor shall calculate the within-subject Coefficient of Variation and % Repeatability Coefficient as shown:

.

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

### 2.1.3 Performance Guidance:

In Section 3 of the Profile, 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 such that the 95% confidence limits of the site wCV fall below the Claim value.

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, suppose a site's conformance dataset has N=30 cases with test-retest data. In the Profile, wCV of 8.5%, 10.3%, and 14.1% are claimed for lesions 50-100mm, 35-49mm and 10-34mm in diameter, respectively. The conformance dataset represents a mixture of lesions sizes; the authors determined that for this mixture of cases, the average wCV should be <9.0% (since most of the lesions were large). Given the sample size of N=30, a site’s estimated wCV must be <7.1% for the conformance dataset in order to be 95% confident that the precision requirement of <9% is met. Thus, 7.1% is the maximum allowable wCV for a site and would be specified in section 3 of the Profile.

Calculation of this maximum allowable variability is described in Appendix A; a statistician can also be consulted for calculating this value. Note that when there is 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, 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) should be specified. 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 Assessment Procedure: Site Bias

The bias of the biomarker at a Site is a measure of the composite performance of the entire system at the Site. Each of the Actors in the system may contribute bias to the measurement but for the Site Assessment Procedure it doesn’t really matter where the source of bias is as long as the total performance stays within the bounds specified in the Profile.

In contrast, in order to assess individual actors in the chain, the total bias will need to be "allocated" appropriately to each actor (e.g. scanner, radiologist, algorithm, etc.) in the form of Profile requirements, and assessed based on the guidelines in Section 3 of this document.

Figure 2 illustrates the study design of a phantom study to assess bias. Lesions of known truth values can be measured multiple times, from which the bias can be estimated. By choosing at least 5 different truth values, the same study design can be used to assess linearity (see Section 2.3) and estimate the regression slope (see Section 2.4).



**Figure 2**: Bias can be estimated by using a phantom with several (in this case 5) lesions with known truth values, and imaging and measuring them multiple times. The mean of the differences between the measurements for a given truth value is an estimate of the bias. The bias should be estimated at multiple truth values, as illustrated here, to construct a bias profile. These same data can be used to assess the property of linearity (see Section 2.3) and to estimate the regression slope (see Section 2.4).

### 2.2.1 Test Dataset Guidance

Meta-analyses of published literature or groundwork studies are often used by QIBA authors to understand the bias of their biomarker. These studies have been performed to populate the Claim statements with realistic estimates of the bias. See also the Site Assessment Dataset discussion in section 2.

Desirable properties of a dataset for assessing composite bias follow are that it:

* Has known ground truth or a reliable reference standard is available.
* Has not been used for training the algorithm being tested
* 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. location of disease, severity of disease, spectrum of disease characteristics (diffuse vs focal), confounding factors (artifacts, degraded signal from patient weight), tumor sizes
* Is easily accessible, e.g. located on QIDW

For example, in the advanced disease CT Volumetry Profile, the previously designed FDA Lungman phantom is being provided to sites on the QIDW. The Lungman phantom has 42 distinct target tumors. The Profile specifies the number and range of lesion characteristics to be measured (sizes, densities, shapes).

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

A key characteristic of the test dataset is the sample size. Since this interacts strongly with the decisions about the desired performance target and confidence, it is discussed below in section 2.2.3.

### 2.2.2 Procedure Guidance

Based on groundwork studies or the literature, one should have a good understanding of the characteristics of the bias of the biomarker and how it varies over different measurement targets (i.e. the biomarker’s bias profile). For example, it may be known that the bias is pretty constant over the relevant range of the biomarker, or that the %bias (i.e. bias / true value) is pretty constant, or even that the %bias is pretty constant only in small ranges. Knowing the bias profile of the biomarker helps guide decisions on how to formulate Claims. See Figure 2 for an illustration of a bias profile.

Sites need to use the conformance dataset (described in Section 2.2.1) to construct a bias profile. In Section 4 of the Profile, instruct the sites how to generate a bias profile so that the site’s bias relative to the assumptions made about the bias in the Profile can be evaluated. Expert opinion will likely be the basis for about what characteristics to stratify on and the metrics to use. Make sure sufficient sample size is obtained in each stratum (i.e. at least 5 cases). For example, in the CT Volumetry Profile, sites must stratify the lesions by shape. For each stratum actors estimate the population bias.

In Section 4 of the Profile, describe the statistical method for estimating a site’s bias. This should include a description of what to measure, as well as the formulae for calculating bias and its 95% CI. Sample language for estimating the % bias is given here:

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

For each case, the assessor shall calculate the value of the *<name of QIB here>* (denoted Yi), where *i* denotes the *i*-th case. The assessor shall 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 assessor shall calculate the 95% CI for the bias as , where is from the Student’s t-distribution with =0.025 and (N-1) degrees of freedom.

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

### 2.2.3 Performance Guidance:

In Section 3 of the Profile, specify the maximum allowable bias, meaning the maximum bias that the site can have for the conformance dataset. For most current Profiles, assumptions about the bias take on one of two forms:

1. The bias is negligible, or
2. The bias is less than a certain threshold.

For situation i.), the maximum allowable bias is +5%; for situation ii.), the maximum allowable bias is + the threshold specified in the Profile. For both of these situations, sites need to estimate their bias and construct a 95% CI for the bias. In situation i.), the upper bound of the CI should be less than 5% and the lower bound should be greater than -5%. In situation ii.), the upper and lower bounds of the CI should be less than the specified threshold.

The sample size for testing the bias depends on several factors:

1. The variability in bias between subjects (This is the between-case differences in bias. If the magnitude of the bias is pretty constant for all cases, then the sample size requirement will be smaller (because the between-subject variance is small). If the magnitude of the bias varies greatly between cases, then the sample size requirement will be larger (because the between-subject variance is large.)), and
2. The desired width of the 95% CI for bias. (If sites are expected to have little bias, then a sample size can be chosen that will give wider CIs because of confidence that sites will still have CIs below the maximum allowable bias. If sites are expected to have bias near the maximum allowable bias, then a sample size that will give tighter CIs should be chosen.)

For example, in the CT Volumetry Profile, which specifies that the bias is negligible (situation i), 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. The Profile authors believed that sites’ bias could be as large as 4%, so in order to be 95% confident that the bias was <5%, they chose a sample size that would provide a very tight CI of +1%. A site’s CI must lie completely in the interval -5% to +5% for the conformance requirement to be met.

Calculation of the sample size is described in Appendix B; a statistician can also be consulted for calculating this value.

In addition, in Section 3 of the Profile, the maximum allowable bias for each of the strata specified in the bias profile (e.g. nodules grouped by shape) should be specified. Profile authors should use their discretion in deciding on the maximum allowable bias 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, the estimate*d population bias* (not the lower and upper bounds of the CI) must be between -5% and +5% for each stratum in order for the conformance requirement to be met.

## 2.3 Assessment Procedure: Site Linearity

Longitudinal Claims that provide a 95% CI for the true change in the biomarker rely on the property of linearity. In this section we discuss the procedures for Sites to assess the linearity of their measurements. Note that each of the Actors in the system may play a role in linearity but for the Site Assessment Procedure it doesn’t really matter which actor(s) is responsible.

### 2.3.1 Test Dataset Guidance

Ideally, a single dataset that has been sequestered (i.e., not accessed before by the algorithm, e.g. for training) should be used to assess linearity. The dataset may be generated from phantoms which has the advantage of knowing ground truth or clinical data if there is a reliable reference standard. (Note that since phantom studies bypass the influence of variability and bias in the patient handling, it would be good to first determine that those early activities are not expected to be a source of non-linearity.) See also the Site Assessment Dataset discussion in section 2.

Desirable properties of a dataset for assessing linearity are that it:

* Has not been used for training the algorithm being tested
* 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, e.g. located on QIDW
* Has ground truth, or a reliable reference standard to serve as truth, known or at least multiples of ground truth can be formulated (the precise relationship between several measurements is known even if the exact values are not i.e. X, 2X, 3X, etc.)
* 5-10 nearly equally-spaced ground truth values are available (See Figure 2)
	+ During the testing procedure the system will make 5-10 observations per ground truth value (a total of 50 measurements is recommended).

As an example, in the advanced disease CT Volumetry Profile, the previously designed FDA Lungman phantom is being provided to sites on the QIDW to assess linearity.

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

### 2.3.2 Procedure Guidance

In Section 4 of the Profile, describe the statistical method for assessing linearity. This should include a description of what to measure, as well as the formulae for making the calculations. Sample language is given here:

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

For each case, the assessor shall 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. The assessor shall 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 the assessor shall fit a linear model: , and estimate R2.

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

### 2.3.3 Performance Guidance:

Specify the maximum allowable and, potentially, the minimum R-squared (R2). For example, the absolute value of the estimate of should be <0.5 and R-squared (R2) should be >0.90.

## 2.4 Assessment Procedure: Regression Slope

Longitudinal Claims that provide a 95% CI for the true change in the biomarker rely on the assumption that the slope of the regression of the biomarker on the true value is known. For most Claims, it is assumed that the regression slope equals one. In this section we discuss the procedures for Sites to estimate the slope.

### 2.4.1 Test Dataset Guidance

Ideally, a single sequestered dataset should be used to estimate the slope, often designed from phantoms. See also the Site Assessment Dataset discussion in section 2.

Desirable properties of a dataset for estimating the slope are that it:

* Has not been used for training the algorithm being tested
* 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
* Is easily accessible, e.g. located on QIDW
* Has ground truth or multiples of ground truth can be formulated (i.e. X, 2X, 3X, etc.) , or a reliable reference standard is available to serve as truth
* Has 5-10 nearly equally-spaced ground truth values with 5-10 observations per ground truth value (a total of 50 measurements is recommended) (See Figure 2).

As an example, in the advanced disease CT Volumetry Profile, the previously designed FDA Lungman phantom is being provided to sites on the QIDW to estimate the slope.

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

### 2.4.2 Procedure Guidance

In Section 4 of the Profile you will need to describe the statistical method for estimating the slope. This should include a description of what to measure, as well as the formulae for making the calculations. Sample language is given here:

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

For each case, the assessor shall 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. The assessor shall fit an ordinary least squares (OLS) regression of the Yi’s on Xi’s: . Let denote the estimated slope. The assessor shall calculate its variance as , where is the fitted value of Yi from the regression line and is the mean of the true values. The assessor shall calculate the 95% CI for the slope as .

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

### 2.4.3 Performance Guidance:

Specify the allowable range for the slope. 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. These thresholds should be specified in Section 3 of the Profile.

# 3. Assessing Individual Actors (Actor Performance)

In section 2 of this document we described how to assess the conformance of a site to the Profile Claims. For example, in section 2.1 where we discussed precision, we assumed that it was reasonable to obtain an estimate of the precision of the entire biomarker production chain at a particular site. This is accomplished by performing a test-retest study at the site, where subjects are scanned two or more times and measurements are made by one or more readers using algorithms available at the site.

In this section we describe how to assess conformance of individual actors in the biomarker production chain. Each QIBA biomarker committee will need to consider both the necessity and practicality of assessments at the site-level and the actor-level. If an actor (e.g. technician) is known to contribute negligibly to the bias and/or imprecision (e.g., it has been demonstrated in groundwork results, or the specific actor has been assessed), then its bias and precision do not need to be tested. It is important to focus on the actors expected to contribute the most to the bias and imprecision when testing conformance to the statistical assumptions underlying the Claims. Similarly, if the Profile requirements on a given actor are entirely procedural ("shall orient the patient on the scanner table in a supine position") rather than performance ("shall achieve a noise standard deviation of less than 50HU"), the requirement checklist may not need any data-oriented assessment procedures.

Note that some actors in the biomarker production chain generate intermediate data. For example, the scanner generates the image from which the biomarker is measured. For these actors, intermediate outcomes can be assessed for conformance. For example, image noise and resolution could be assessed for the scanner. In order to determine the acceptable levels of these intermediate outcomes, published studies may be available that have looked at the effect of intermediate outcomes on the technical performance of the biomarker measurements. Otherwise, groundwork projects may be needed to study the relationships between the intermediate outcomes and the bias and precision of the biomarker. As a temporary means while groundwork studies are performed, expert opinion may be used.

In assessing the technical performance of individual actors, the first thing to consider is the difference between assessing the performance of the site and assessing the performance of individual actors. For assessing the site, both the performance metric and the performance target are taken directly from the Profile Claims; for example, the within-subject Coefficient of Variation (wCV) must be below a certain threshold for the biomarker measurement. In contrast, for assessing the performance of individual actors, we need to determine the portion of the bias and variability attributable to that actor. We suggest two options below for determining reasonable expectations for individual actors.

* In the ***Bottom-Up Approach***, we design studies that isolate an actor so that we can estimate this actor’s technical performance (i.e. start from "what's reasonably achievable" for each actor.) With this bottom-up approach, we must be careful not to accept too much bias and/or variability from any one actor because the composite performance of the actors in the biomarker production chain still needs to meet the Claim.
* In the ***Top-Down Approach***, we take the composite performance and use statistical modeling (and expert judgment) to determine the allocation of the composite measurement error to the individual actors. This approach requires data from a test-retest study.

Each of these approaches is described below.

## 3.1 Bottom-Up Threshold Selection Approach:

In the bottom-up approach, studies are performed that isolate an actor so that this actor’s technical performance can be estimated. These types of studies provide data on reasonable targets of performance of individual actors, and thus might provide reasonable thresholds for actor requirements in the Profile. Table 2 describes several study designs to assess the precision of individual actors. For completeness, the last two rows of the table also include study designs for assessing the variability of the whole biomarker production chain.

**Table 2: Estimating Actor Precision**

|  |  |  |
| --- | --- | --- |
| **What you want to Estimate:** | **Actors Assessed** | **Study Design** |
| Intra-algorithm, Intra-reader variance (Algorithm/Reader Repeatability) | Algorithm/ Reader combination | Multiple subjects are each scanned once, all using the same scanner. From each scan, the same reader makes multiple measurements using the same algorithm.  |
| Intra-algorithm, Inter-reader variance (Reader Reproducibility) | Reader | Multiple subjects are each scanned once, all using the same scanner. From each scan, multiple readers each make multiple measurements, all using the same algorithm.  |
| Inter-algorithm, Intra-reader variance (Algorithm Reproducibility) | Algorithm | Multiple subjects are each scanned once, all using the same scanner. From each scan, the same reader makes multiple measurements using each different algorithm.  |
| Intra-scanner, Intra-algorithm, Intra-reader variance (Variance associated with the entire biomarker production chain) | Scanner/ Algorithm/ Reader combination | Multiple subjects are each scanned two or more times, all using the same scanner. From each scan, the same reader makes multiple measurements, all using the same algorithm. |
| Inter-scanner, Intra-algorithm, Intra-reader(Variance associated with the entire biomarker production chain)  | Scanners | Multiple subjects are each scanned two or more times on different scanners. From each scan, the same reader makes multiple measurements, all using the same algorithm. |

Table 3 describes several study designs based on phantoms or DROs that can be used to estimate individual actors’ bias. For completeness, the last two rows of the table also include study designs for assessing the bias of the whole biomarker production chain.

**Table 3: Estimating Actor Bias**

|  |  |  |
| --- | --- | --- |
| **What you want to Estimate:** | **Actors Assessed** | **Study Design** |
| Intra-algorithm, Intra-reader bias  | Algorithm/ Reader combination | A phantom with known ground truth value is scanned multiple times, using the same scanner. From each scan, the same reader makes multiple measurements using the same algorithm. (Alternatively, DROs can be used where the same reader makes multiple measurements using the same algorithm from the same DRO) |
| Intra-algorithm, Inter-reader bias | Readers | A phantom with known ground truth value is scanned. Multiple readers, using the same algorithm, make measurements from the same image. (DROs may also be used). |
| Inter-algorithm, Intra-reader bias | Algorithm | A phantom with known ground truth is scanned. A reader, using different algorithm, makes measurements from the same image. (DROs may also be used). |
| Intra-scanner, Intra-algorithm, Intra-reader bias (Bias of the entire biomarker production chain) | All Actors (intra-scanner, algorithm, reader) in combination | A phantom with known ground truth is scanned multiple times with the same scanner. A reader, using the same algorithm, measures the biomarker value from each scan.  |
| Inter-scanner, Intra-algorithm, Intra-reader bias (Bias of the entire biomarker production chain) | Scanners | A phantom with known ground truth is scanned multiple times with different scanners. A reader, using the same algorithm, measures the biomarker value from each scan.  |

Consider the following example. In the CT volumetry Profile, an algorithm/reader repeatability challenge study was performed with multiple algorithms. From this study, the Profile authors estimated the intra-reader intra-algorithm variance of multiple algorithms. The algorithms were separated into three groups based on their wCV estimates: poor-, middle-, and high-precision. It was decided by the Profile writers that the threshold for algorithm precision be set at the wCV of middle-precision performers. In their Profile, the general methods described in Section 2 are provided so that the individual algorithm actor’s repeatability can be estimated and tested to determine if it is less than or equal to the value corresponding to the middle performers, with 95% confidence.

## 3.2 Top-Down Threshold Selection Approach:

A top-down approach is to start from a level of site performance that is considered to be clinically valuable, or a level of performance that has been shown empirically to be achievable, and then allocate a proportion of the corresponding overall bias and overall imprecision to each actor.

To estimate the proportion of overall bias and overall imprecision attributable to individual actors, one can use well-controlled clinical or phantom studies and statistical modeling to parse out the individual sources of variance and bias. Once these proportions are estimated, the general methods described above in Section 2 should be applied with the goal that the individual actor’s bias and/or imprecision is less than or equal to the proportion of bias/precision attributable to their component in the imaging process.

Consider the following example. A test-retest study is performed at a site on a sample of N subjects. The same scanner (make, model and protocol) is used at two time points. The same algorithm is used multiple times by the same reader to generate K biomarker measurements from each of the N x 2 acquisitions. The N x 2 x K observations are used to estimate the composite wCV. A random effects statistical model can then be built where the variability attributable to intra-acquisition and the variability attributable to intra-algorithm can be estimated. The intra-acquisition variability also includes intra-subject variability. (Note that the study could also be performed on phantoms to estimate the proportion of the total bias attributable to individual actors.) A similar approach was used in the QIBA PET Amyloid Profile (<http://qibawiki.rsna.org/index.php/Profiles>). From their meta-analysis, the composite wCV was known. Instead of a statistical modeling approach, the authors used expert opinion to partition the composite wCV into that attributable to the scanner and that attributable to the algorithm (including the reader using the algorithm).

# 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. *Statistical Methods in Medical Research* 2017. Jan 1: [Epub ahead of print]

[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.

Table B-1: Sample size requirements for various different bias/variance combinations.

|  |  |
| --- | --- |
|  | **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 variance value in determining your sample size. When the variance between cases is 20%, 80 cases are needed for a tight +1% CI around the bias.