**Guidance for Profile Authors Drafting Statistical Assumption and Composite Performance Assessment Procedures**

This document provides guidance on drafting 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). Assessing composite performance is central to achieving Stage 4 (Claim Confirmed) and Stage 5 (Clinically Confirmed), but it may or may not be part of general 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. 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)

TODO Bring diagram up here to describe Top-Down vs Bottom-up and how that relates to assessing sites vs assessing actors. Start from Eric's sketch.

# 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 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.

Need to do the whole chain for Claim Confirmed.

## 2.1 Assessment Procedure: Site Within-subject Precision

The Within-subject Precision of the biomarker 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 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?>

*TODO Diagram here showing, say, a test-retest on humans, and how the "study design" would flow. TODO [Duplicate work items in each of 2.2, 2.3, 2.4] The later 2.2, 2.3, 2.4 are often better done as a phantom study (although in some BCs, e.g. PDFF, they may have a "reference standard" that is reasonable to use in humans – in vivo spectroscopy (not a gold standard but is an acceptable reference measure of lipid fraction in liver – especially for bias assessment) – TODO Briefly discuss considerations/rationale of when it's worth going away from phantom – e.g. pull a paragraph worth of material from the related paper. It's a judgement call for the committee. But state the basis for the judgement. Strict a cautionary tone but show it as an option. (Between Ed and Scott can draft some text)).*

### 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, of test-retest subjects scanned at the site with measurements made by readers at the site using algorithms available at the site.

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)
* Replicate measurements can be ethically obtained (i.e. ?radiation/contrast considerations of test-retest??? )

Details about the conformance precision 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>* for the first replicate measurement (denoted Yi1) and for the second replicate measurement (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 variance:

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, suppose a site's conformance dataset has N=30 cases with test-retest data; the cases represent a mixture of sizes. In the Profile, a wCV of 8.5%, 10.3%, and 14.1% is 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%. Given the sample size, a site’s estimated wCV must be <7.1% for the conformance dataset in order to be 95% confident that the precision requirements are 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; 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 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, software, etc) in the form of Profile requirements, and assessed based on the guidelines in Section 3 of this document.

TODO – Nancy will sketch figures for 2.2/3/4 and Kevin can tidy up.



**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.

Assessing a clinical site’s composite performance needs a different kind of dataset. Ideally, a single sequestered dataset should be used, often designed from phantoms.

Desirable properties of a dataset for assessing composite bias follow:

* Ground truth is known.
* 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)
* Have permission to distribute on QIDW (although if not possible, it should at least be publically available elsewhere)

Some examples from QIBA follow:

* In the US SWS profile, …. <TODO Brian – nice example of deliberately targeting ground truth phantom>
* 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)..
* PET DRO?

### 2.2.2 Procedure Guidance

Based on groundwork studies or the literature, you should have a good understanding of the characteristics of the bias of your biomarker and how it varies over different measurement targets (i.e. the biomarker’s bias profile). For example, you may know 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 your 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 you will want to instruct the sites how to generate a bias profile so that you can evaluate the site’s bias relative to the assumptions you have made about the bias 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 bias profile from various QIBA profiles:

* In the CT Volumetry Profile, sites must stratify the lesions by shape. For each stratum actors estimate the population bias.
* For the US SWS profile, sites must estimate the bias for …

In Section 4 of the Profile you also need to 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. Standard language for estimating the % bias is given here:

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

For each case, calculate the value of the *<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.

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

### 2.2.3 Calculate the sample size for testing the bias:

In section 3 of the Profile you must specify the maximum allowable bias, in other words, 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 you expect sites to have little bias, then you can choose a sample size that will give wider CIs because you feel certain that sites will still have CIs below the maximum allowable bias. If you expect sites to have bias near the maximum allowable bias, then you should choose a sample size that will give tighter CIs.)

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; you can also consult a statistician for calculating this value.

In addition, in Section 3 of the Profile you should also specify the maximum allowable bias for each of the strata specified in the bias profile (e.g. nodules grouped by shape). 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 estimated *popbias* (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 sequestered dataset should be used to assess linearity. The dataset may be generated from DROs or phantoms which has the advantage of knowing ground truth. Since these bypass the influence of variability in the patient handling and, for the DRO, image acquisition, it would be good to first determine that those earlier activities are not expected to be a source of non-linearity.

Desirable properties of a dataset for assessing linearity follow:

* 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)
* Ground truth is known (the actual correct values of several measurements is 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 measurand values are available (See Figure 2)
  + During the testing procedure the system will make 5-10 observations per measurand value (a total of 50 measurements is recommended).

Some examples from QIBA follow:

* In the Amyloid profile, a DRO was designed with 5 true values to test for linearity.
* 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 conformance 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 you will need to 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. Standard language is given here:

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

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.

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

2.3.3 Specify the maximum allowable and minimum R-squared (R2):

The absolute value of the estimate of should be <0.50 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.

Desirable properties of a dataset for estimating the slope follow:

* 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)
* Ground truth is known or at least multiples of ground truth can be formulated (i.e. X, 2X, 3X, etc.)
* 5-10 nearly equally-spaced measurand values are available with 5-10 observations per measurand value (a total of 50 measurements is recommended) (See Figure 2).

Some examples from QIBA follow:

* In the Amyloid profile, a DRO was designed to estimate the slope.
* 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 conformance 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. Standard language is given here:

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

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 .

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

2.4.3 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 should be specified in Section 3.

# 3. Assessing Individual Actors

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 section 3 we describe how to assess conformance of individual actors in the biomarker production chain.

<TODO Kevin - turn this into a section on Choosing Assessment Metrics for each Actor? Should somehow capture their key contribution to the final Claim performance>

Note that some actors in the biomarker production chain generate intermediate data. For example, the acquisition device 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 acquisition device. 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) will be below a certain threshold for the biomarker measurement.

Since each of the actors included in the profile contribute to the generation of that measurement, the site performance is a composite of the performance of the individual actors. In contrast, the performance metric for a given actor will depend on the nature of its task in the biomarker production chain. Setting a performance target for each actor can be approached several ways. In the *Bottom-Up Approach*, we design studies that isolate an actor so that we can estimate this actor’s technical performance (start from "what's reasonably achievable" for each actor. Trick to then get a computational estimate of the Site performance since there are likely "interesting" interactions between the actors and we might have missed sources. It would be best to measure site Experimentally in a study.)(Doing this computational approach is handy when you don't have test-retest data and can't easily get it. If you are "off" that will be caught in the Claim Confirmation stage and you will have a new experimental value. So that lets you defer the test-retest until Claim Confirmed stage. But if you don't do Claim Confirmed, you are on thinner ice with your Claim).

In the *Top-Down Approach*. , we use data from a full-scale study, e.g. test-retest, and then determine the allocation of the composite measurement error to the individual actors.

For some QIBs there are few or no good published test-retest studies where subjects were scanned two or more times over a short period of time following the imaging methods described in the Profile. Thus, we do not have an estimate of the technical performance of the entire biomarker production chain, e.g. “composite wCV” referred to in Section 2 of this document. Instead, the literature or our groundwork studies provide information about the technical performance of segments of the biomarker production chain.

Each of these approaches is described below.

<TODO – diagram of the biomarker production chain to re-inforce this idea – at least include the modality, measurement algorithm and reader. Maybe put a box around the algorithm and reader since they are an operational pair – not always separable (or worth the effort of separating) so often one box in the diagram.>

TODO Diagram here showing, say, one image each from several patients, and how the "study design" would flow. In principle it is easier to assess the actors compared to the site.

## 3.1 Bottom-Up Threshold Selection Approach:

<TODO intro sentence> In the bottom-up approach, studies are performed that isolate an actor so that this actor’s technical performance can be estimated. ofTable x describes several study designs to assess the precision of individual actors. Many of these designs do not require a full test-retest study, and thus are more feasible to conduct.

Table x: Estimating Actor Precision

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

Table y: Estimating Actor Bias

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

Consider the following example. In the CT volumetry Profile, a software/reader repeatability study was performed with multiple software 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 the Profile, to assess conformance of algorithms, images from a previously published clinical test-retest study are being made available for software vendors to test their algorithms. (It is implicitly assumed that the acquisition device used for this dataset met the Profile requirements for conformance.) The profile provides a link to the QIDW to obtain the images for actors to download. The Profile also includes the statistical methods for estimating wCV (these methods are given in Section 2.1.2), and it provides a maximum allowable wCV in order for algorithms to meet the conformance requirement (the rationale for determining the maximum allowable wCV is given in Section 2.1.3).

Continuing with this example, the CT Volumetry Profile authors also conducted a reader reproducibility study and a software reproducibility study. From these studies they generated a table of reproducibility coefficients (RDC). The table gives the RDCs for various situations such as when different readers make measurements at two time-points or when different algorithms are used to make measurements at two time-points. With this table they were able to expand their longitudinal claim to include not only the situation where the same imaging methods are used at both time points but also when different readers and/or different algorithms are used at the two time points.

## 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. If an actor (e.g. technician) is known to contribute negligibly to the bias and/or imprecision, 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.

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 software 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-software 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 Amyloid profile. 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 acquisition device and that attributable to the workstation (including the reader using the workstation). The authors then built a DRO to test the precision of workstation vendors. Workstation vendors’ measurement software is then used by a reader to make multiple repeat measurements of each target on the DRO. In the profile the authors provided steps describing how many measurements to make, how to estimate wCV (see section 2.1.2), and what the maximum allowable wCV is for conformance (see section 2.2.2).

# 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **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.