Assessment Procedure Guidance

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

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

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

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

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

This guidance describes:

(1) The statistical assumptions underlying different types of Claims so that authors of the Profiles know which assumptions need to be assessed.

(2) Procedures appropriate for assessing the composite performance of a site. Testing of sites appears in the Profile in two places:
   a. The requirements (in Section 3 of the Profile) for the site to satisfy the assumptions.
   b. The procedure (in Section 4 of the Profile) for testing the metric that underlies the assumptions

(3) The procedures appropriate for testing individual actors

a. The requirements for each actor to satisfy the assumptions (in Section 3 of the Profile)
b. Procedures to assess the metric that underlies the requirement (in Section 4 of the Profile)

1. Statistical Assumptions Underlying Claims

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

Table 1: Statistical Assumptions for different Types of Claims

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

- Assessing a Technical Performance Claim (as stated in Section 2)
(e.g. Site is measuring … with a wCV …)
Note, we do not (yet?) have guidance for assessing longitudinal and cross-sectional claims (generally 95% confidence intervals) but we can assess the assumptions.
- Assessing individual Actor performance related to the assumptions underlying the Claim (Requirements in Section 3)
(e.g. Radiologist has a repeatability of X on test data)

2. Assessing Sites

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

This guidance will focus on assessing technical 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?>

2.1.1 Test Dataset Guidance

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

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

Desirable properties of a dataset for assessing composite precision:

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

Some examples from QIBA follow:

- In the amyloid profile, a DRO was developed specifically for testing.
- In the US SWS profile, ….
- In the CT volumetry profile, work is being done to create a phantom with real inserted lesions. In the meantime, images from a previously published clinical test-retest study are being made available to sites for testing.

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*> at time point 1 (denoted Y_{i1}) and at time point 2 (Y_{i2}) where *i* denotes the *i*-th case. For each case, calculate: $d_i = [(Y_{i1} - Y_{i2})/\{(Y_{i1} + Y_{i2})/2\}] \times 100$. Calculate: $wCV = \sqrt{\sum_{i=1}^{N} d_i^2 \,/(2 \times N)}$. Estimate the % Repeatability Coefficient as $\widehat{\%RC} = 2.77 \times wCV$.

2.1.3 Calculate the maximum allowable variability:

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

The maximum test-retest variability depends on several factors:

i. The number of subjects in the conformance dataset (described in section 2.1.1), and

ii. The estimate of precision used in the Profile claim.

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

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

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

2.2 Bias Assessment Procedure

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.

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 (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 allowed you to decide how to formulate your claims.

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 measurand<*name of QIB here*> (denoted Y_i), where *i* denotes the *i*-th case. Calculate the % bias:

$b_i = [(Y_i - X_i)/X_i] \times 100$, where X_i is the true value of the measurand. Over N cases estimate the population bias: $popbias = \sqrt{\sum_{i=1}^{N} b_i /N}$. The estimate of variance of the bias is $\widehat{Var}_b = \sum_{i=1}^{N}(\%b_i - \hat{b})^2 /(N-1)$. The 95% CI for the bias is $\hat{b} \pm t_{\alpha=0.025,(N-1)df} \times \sqrt{\widehat{Var}_b}$, where $t_{\alpha=0.025,(N-1)df}$ is from the Student's t-distribution with α=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:

i.) The bias is negligible, or

ii.) 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:

i. 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

ii. 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 chose 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
    - 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 Y_i), where *i* denotes the *i*-th case. Let X_i denote the true value for the i-th case. Fit an ordinary least squares (OLS) regression of the Y_i's on X_i's. A quadratic term is first included in the model to rule out non-linear relationships: $Y = \beta_o + \beta_1 X + \beta_2 X^2$. If $\beta_2 = 0$, then a linear model should be fit: $Y = \beta_o + \beta_1 X$, and R^2 estimated.

2.3.3 Specify the maximum allowable β_2 and minimum R-squared (R^2):
The estimate of β_2 should be <0.50 and R-squared (R^2) should be >0.90.

2.4 Regression Slope Assessment Procedure

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 DROs or 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).

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 Y_i), where *i* denotes the *i*-th case. Let X_i denote the true value for the i-th case. Fit an ordinary least squares (OLS) regression of the Y_i's on X_i's: $Y = \beta_o + \beta_1 X$. Let $\widehat{\beta_1}$ denote the estimated slope. Calculate its variance as $\widehat{Var}_{\beta_1} = \{\sum_{i=1}^{N}(Y_i - \widehat{Y_i})^2 / (N - 2)\} / \sum_{i=1}^{N}(X_i - \bar{X})^2$, where $\widehat{Y_i}$ is the fitted value of Y_i from the regression line and $\bar{X}$ is the mean of the true values. The 95% CI for the slope is $\widehat{\beta_1} \pm t_{\alpha=0.025,(N-2)df}\sqrt{\widehat{Var}_{\beta_1}}$.

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 Actors

3. Assessing Individual Actors

The first thing to consider is the relationship between assessing the performance of the site and assessing the performance of individual actors. For 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) being below a certain percentage 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.

<Insert description of bottom-up where you measure each of the actors first and then the site performance is the sum of that. In which case you can skip allocating and proceed to the experiments below>

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. acquisition device) 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 an example. In the CT lung mass profile, through a series of groundwork studies, the authors determined that the total imprecision in measuring lung mass volumes is apportioned as follows: 77% due to the combined effects of intra-software and intra-reader, and 23% due to intra-scanner variability. The component contributing the largest variability is the software-reader combination. In the CT volumetry profile, for lesions 10-34mm in diameter, the wCV used in the claim when holding scanners, readers, and software constant at the two time points is 10%. Since $(0.77) \times 0.10 = 0.077$, the maximum allowable wCV for the combined actors of software and reader would be 7.7%.

Suppose we want to test the conformance of a measurement software vendor. We might require a study where a single scan is performed on each of a sample of N patients. The vendor's measurement software is then used by a reader to make multiple repeat measurements from each scan. The vendor would estimate the wCV from the replicate measurements for each subject's scan, then take an average over the N subjects. We would probably require that the vendor perform this study for multiple readers. The readers' average wCV would need to be ≤7.7% in order for the measurement software vendor to be conformant. Alternatively, a DRO study can be conducted where a reader makes multiple repeat measurements on the DRO.

References:

[1] Obuchowski NA, Buckler A, Kinahan P, Chen-Mayer H, Petrick N, Barboriak DP, Bullen J, Barnhart H, Sullivan DC. Statistical Issues in Testing Conformance with the Quantitative Imaging Biomarker Alliance (QIBA) Profile Claims. *Academic Radiology* 2016; 23: 496-506.

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

[3] Raunig D, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *SMMR* 2015; 24: 27-67.

Appendix A:

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

$$H_o: \theta \geq \delta \quad \text{versus} \quad H_A: \theta < \delta.$$

The test statistic is: $T = \frac{N\times(\widehat{RC}^2)}{\delta^2}$. Conformance is shown if $T < \chi^2_{(\alpha),N}$, where $\chi^2_{(\alpha),N}$ 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, $\chi^2_{(0.05),N}$ in a table of chi square values. Then solve for $\widehat{RC}$ in the equation:

$$\chi^2_{0.05,N} = \frac{N\times(\widehat{RC}^2)}{\delta^2}.$$

For example, in the CT Volumetry Profile, N=31 and δ=21%. $\chi^2_{(0.05),31}$ = 19.3 from http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm. Then, solving for $\widehat{RC}$, 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%**
Var$_b$*=5%	22	8	≤5	≤5	≤5
Var$_b$=10%	42	13	7	≤5	≤5
Var$_b$=15%	61	17	9	7	≤5
Var$_b$=20%	80	22	12	8	6
Var$_b$=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.