October 1, 2014

Statistical Requirements for Testing Profile Claims

This document contains recommendations on the claim statement language to be used in the Profiles and the minimum requirements for meeting the statements in the claim. Future work should include descriptions of adequately powered study designs to test whether actors meet the claim statements, as well as descriptions of the statistical analysis methods for testing compliance.

I. Claim Statement Language

Claims involve one or more summary statements of the technical performance of the quantitative imaging biomarker (QIB) that is achievable by the imaging procedure. The summary statement should use simple language yet conform to standard statistical convention.

In describing the "imaging procedure" it is important to specify the imaging scenario applicable to the claim statement. For example, "This Profile permits the participating compliant actors (acquisition device, radiologist, image analysis tool, etc.) to be different at the two time points."

Table 1: Examples of Claim Statements

CROSS-SECTIONAL CLAIM: There is a 95% probability that the measured < *insert QIB name here* > ± < *insert precision value here* > encompasses the true value.

For example,
"There is a 95% probability that the measured SUVr ± 5% encompasses the true SUVr."

LONGITUDINAL CLAIM: There is a 95% probability that the measured change in < *insert QIB name here* > ± < *insert precision value here* > encompasses the true change.

For example,
"There is a 95% probability that the measured change in mass volume ±30% encompasses the true change in mass volume."

Following the summary statement, there should be a footnote or "Notes" section that briefly specifies the statistical assumptions behind the statement. Below are several options for these Notes.

Table 2: Examples of *Notes* Following Claim Statement

****

Cross-Sectional Claim:

i. For Unbiased QIB Measurements:

Notes:

- *The QIB measurements are unbiased.*
- *The precision value in the summary claim statement is 1.96 ×within-subject standard deviation* (for interval or ratio QIBs) or *1.96 ×within-subject CV* (for ratio QIBs).

ii. For QIB Measurements that are allowed to have some bias:

Notes:

- *The precision value in the summary claim statement is the Limits of Agreement (LOA) for 95% confidence* or *the total deviation index (TDI) for 95% coverage.*

Longitudinal Claim:

i. For Unbiased QIB Measurements where the same imaging procedures are used at the two time points:

Notes:

- *The QIB measurements demonstrate the property of linearity*
- *The change measurement is unbiased.*
- *The precision value in the summary claim statement is the repeatability coefficient (RC) at a single time point.*

ii. For Unbiased QIB Measurements where different imaging procedures are used at the two time points:

Notes:

- *The QIB measurements demonstrate the property of linearity*
- *The change measurement is unbiased.*
- *The precision value in the summary claim statement is the reproducibility coefficient (RDC) at a single time point.*

iii. For QIB Measurements that are allowed to have some bias where the same imaging procedures are used at both time points:

Notes:

- *The QIB measurements demonstrate the property of linearity*
- *The change measurement is unbiased.*
- *The precision value in the summary claim statement is the repeatability coefficient (RC) at a single time point.*

iv. For QIB Measurements that are allowed to have some bias where different imaging procedures are used at both time points:

Notes:

- *The QIB measurements demonstrate the property of linearity*
- *The precision value in the summary claim statement is the total deviation index (TDI) for the change measurement for 95% coverage.*

****

The precision value used in the summary claim statement should be based on i) a meta-analysis of the relevant published and unpublished literature, groundwork research, and/or field studies, and ii) clinically acceptable limits. The precision value should represent a balance between what can be achieved and what needs to be achieved for the QIB to be clinically useful. For example, from a meta-analysis the 95% CI for the RC might be [5%, 25%]. It will often not be appropriate to use 25% in the claim statement because the CI for the RC is impacted by the number of cases and number of studies in the meta-analysis. If the number of studies is small and/or the number of cases/study is small, then the CI will be wide. If the QIB is useful clinically only when it can be measured to within 15%, then a better value to use in the claim statement is 15%.

The longitudinal claim statement often depicts a worse-case, or best-case, scenario for the imaging procedure. Thus, following the summary longitudinal claim statement, the technical performance for different imaging scenarios can be presented, as illustrated in Table 3. This allows users to quickly identify the scenario applicable to the clinical situation and adjust the claim statement, as appropriate.

Table 3: Precision Table

****

Expected Precision for Alternative Scenarios, *Example*

	Same Scanner				Different Scanner			
	Same SWS		Different SWS		Same SWS		Different SWS	
	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader
Precision*	10%	12%	16%	18%	12%	14%	28%	30%

****

It is recommended that clinically useful interpretation statements follow the longitudinal claim. The longitudinal claim may be bi-directional (e.g. the volume of a mass may increase or decrease) or inherently uni-directional (e.g. emphysema can remain stable or worsen). Thus, there are two sets of interpretation statements, as follows:

Table 4: Clinical Interpretation

****

For longitudinal claims (bi-directional):

The following interpretations can be made with 95% confidence:

1. If the measured change in < *insert QIB name here* > is more than < *insert precision value here* >, then there is a real change.
2. The amount of change is: measured change ± < *insert precision value here* >.

For example,
"The following interpretations can be made with 95% confidence:
1. If the measured change in mass volume is more than 30%, then there is a real change.
2. The amount of change is between: measured change -30% and measured change +30%."

For longitudinal claims (uni-directional):

The following interpretations can be made with 95% confidence:
1. If the measured change in < *insert QIB name here* > is more than < *insert precision value here* >, then there is a real change.
2. The amount of increase is between: measured change < *insert -precision value here, or zero, whichever is larger* > and + < *insert precision value here* >.

For example,
"The following interpretations can be made with 95% confidence:
1. If the measured change in Perc15 is more than 17HU, then there is a real change.
2. The amount of increase in Perc15 is: measured change -17HU (or zero, whichever is larger) and measured change +17HU."

****

II. Minimum Requirements for Claim Statements

The minimum requirements depend on several factors:

1. Cross-sectional or longitudinal claim
2. For longitudinal claim, whether imaging procedures (i.e. scanner, reader, software) are held constant at the two time points or allowed to vary
3. QIB is expected to be biased or unbiased

Table 5 summarizes the minimum requirements for each scenario.

Table 5: Minimum Requirements

	Unbiased cross-sectional QIB Measurements	**QIB cross-sectional Measurements allowed to have some bias**
Cross-Sectional	1. Bias profile with 95% CIs; 2. Precision profile with test of hypothesis that wSD (or wCV) is ≤ claim value.	1. Coverage profile with test of hypothesis that $TDI_{95\%}$ (or LOA) is ≤ claim value.
Longitudinal		
Same imaging procedure at two time points	1. Test of Linearity; 2. Precision profile with test of hypothesis that RC ≤ claim value.	1. Test of Linearity; 2. Coverage profile with test of hypothesis that $TDI_{95\%}$ (or LOA) ≤ claim value.
Different imaging procedure at two time points	1. Test of Linearity; 2. Bias profile with 95% CIs; 3. Precision profile with test of hypothesis that RDC ≤ claim value.	1. Test of Linearity; 2. Coverage profile with test of hypothesis that $TDI_{95\%}$ ≤ claim value; 3. Test of interchangeability with other compliant actors.

Details:

Unbiased: the measurements do not tend to over- or under-estimate the true value

Bias profile with 95% CIs. The sample size will need to be sufficient such that an actor can show that for the relevant range of true values and lesion characteristics, the bias is negligible. The meaning of “negligible bias” will vary between QIBs, but will take on some small pre-specified value. This pre-specified value should be the largest value where the clinical implications of the bias are considered negligible. For example, Table 6 illustrates how increasing bias affects volume measurements. If 5% bias is considered the clinical maximum allowable bias, then the sample size must be sufficient to rule out bias exceeding 5%.

Table 6: Effect of Bias on Measured Volume

True Volume	Measured volume if bias=5%	Measured volume if bias=10%	Measured volume if bias=15%
50	52.5	55	57.5
100	105	110	115
150	157.5	165	172.5
200	210	220	230

Precision profile with test of hypothesis. The claim will contain a specified maximum precision value. Hypothesis testing, in particular a test of non-inferiority, will need to be performed relative to this specified maximum value. The precision metric to be evaluated will depend on the profile: within-subject standard deviation (wSD), within-subject coefficient of variation (wCV), repeatability coefficient (RC), or reproducibility coefficient (RDC). The sample size will need to be sufficient such that an actor can show that for the relevant range of true values and lesion characteristics, the precision is equal to or less than the value in the claim.

Coverage profile with test of hypothesis. The claim will contain a specified maximum deviation value. The deviation value is the value such that 95% of the differences between the QIB measurement and the true value are less than it. Hypothesis testing will need to be performed relative to the specified maximum value. The sample size will need to be sufficient such that an actor can show that for the relevant range of true values and lesion characteristics, the difference between the QIB measurement and the true value is equal to or less than the specified maximum value in the Claim statement 95% of the time.

Test of Linearity. A test, usually employing phantoms, must be performed to demonstrate linearity of the QIB measurements. QIB measurements should be taken at approximately 10 nearly equally-spaced values over the relevant range of the true value, as well as for different characteristics of the measurand. The sample size will need to be sufficient such that an actor can show that for the relevant range of true values and lesion characteristics, the coefficient for the quadratic term is near zero, and the coefficient for the linear term is near one. The meaning of "near zero" and "near one" will vary between QIBs, but will take on some small pre-specified value. The effect of non-zero quadratic terms and/or non-one linear terms can be calculated and clinical judgment can be used to guide identification of the pre-specified value. For example, if the 95% CI for the slope was [0.98, 1.02], what does this mean in terms of bias in measuring tumor volume change? From Table 7 it may be judged that 2% deviation from a slope of one is clinically negligible.

Table 7: Effect of Slope ≠ 1 on Bias

Volume over time:	Absolute True change	Measured change if slope=0.98*	Measured change if slope=1.02
50 to 55	+5	+4.9	+5.10
50 to 75	+25	+24.5	+25.5
50 to 100	+50	+49	+51
50 to 200	+150	+147	+153
50 to 300	+250	+245	+255
50 to 400	+350	+343	+357
50 to 500	+450	+441	+459

*(true volume at t_0×0.98 – true volume at t_1×0.98)

Test for Interchangeability. When QIB measurements are expected to have some bias and imaging procedures are allowed to change during the QIB measurement (e.g. measuring change over time where the imaging procedure can vary at the two time points), then the presence of linearity and adequate coverage profiles will not be sufficient for ensuring that the claim statement is met. In addition to the test for linearity and coverage profile, a test for interchangeability of the actor with other actors, at the individual case level, will be needed.

III. Study Designs for Testing Compliance *(DRAFT! DRAFT!)*

Two compliance studies are needed:

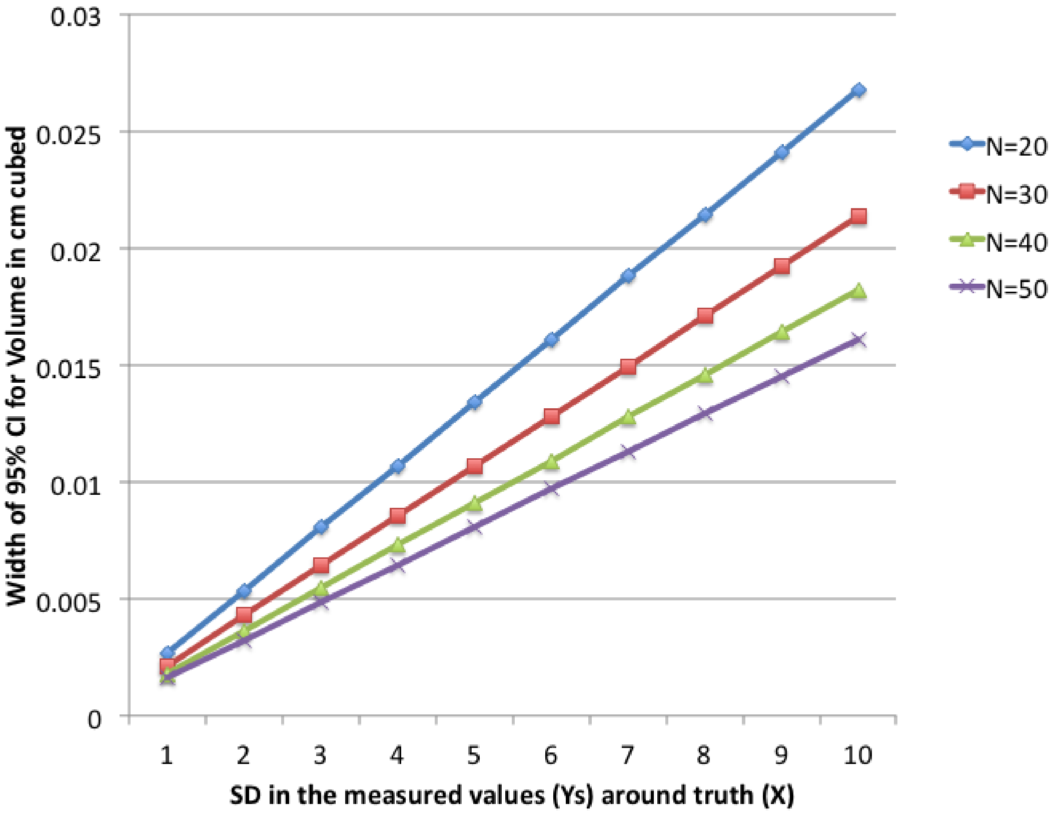
1. Phantom study to assess linearity
2. Clinical study to measure bias and precision

1. Phantom Study Used to Assess Compliance:

Purpose: To evaluate the assumption of linearity

Study Design: Measure lesion volume at 10 nearly equally spaced values over the relevant range: X = 0.5, 1, 5, 10, 50, 100, 200, 300, 400, and 500cm^3.

Sample Size for Slope: How many observations are needed around each of these volumes to estimate slope to within the desired precision? *Probably 3-4, for a total sample size of 30-40.*



You could measure the slope to within ±0.02 with 3-4 observations at each value of X.

If the 95% CI for the slope was [0.98, 1.02], what does this mean in terms of bias in measuring change? See Table 1.

Table 1: Effect of Slope ≠ 1 on Bias

Volume over time:	Absolute True change	Measured change if slope=0.98*	Measured change if slope=1.02
50 to 55	+5	+4.9	+5.10
50 to 75	+25	+24.5	+25.5
50 to 100	+50	+49	+51
50 to 200	+150	+147	+153
50 to 300	+250	+245	+255
50 to 400	+350	+343	+357
50 to 500	+450	+441	+459

*(true volume at t_0×0.98 – true volume at t_1×0.98)

2. Clinical Study

Purpose: Measure precision to complete the precision table.

Precision Table

	Same Scanner				Different Scanner			
	Same SWS		Different SWS		Same SWS		Different SWS	
	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader
Precision								

Study Design: Three-factorial crossed test-retest design. Each patient is scanned twice in a short period of time. Each image is processed with each algorithm by each reader. Suppose there are 3 scanners representing three major manufacturers, 3 algorithms, and 3 readers.

Table 3: Study Design (Option A) to Estimate Precision

	Scanner	Processing
Patient Group 1	A	3 algorithms x 3 readers = 9 measurements
	B	3 algorithms x 3 readers = 9
Patient Group 2	A	3 algorithms x 3 readers = 9
	C	3 algorithms x 3 readers = 9
Patient Group 3	B	3 algorithms x 3 readers = 9
	C	3 algorithms x 3 readers = 9
Patient Group 4	A	3 algorithms x 3 readers = 9
	A	3 algorithms x 3 readers = 9
Patient Group 5	B	3 algorithms x 3 readers = 9
	B	3 algorithms x 3 readers = 9
Patient Group 6	C	3 algorithms x 3 readers = 9
	C	3 algorithms x 3 readers = 9

Within each group there should be representation of lesion size (roughly a range of 0.5 to 500cm^3).

Notes:

- You could sequester part of the data for compliance testing. So you would use some of the data for informing your profile, but you would save out some of the scans from each patient group for later compliance testing.

APPENDIX:

Precision Table

	Same Scanner				Different Scanner			
	Same SWS		Different SWS		Same SWS		Different SWS	
	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader	Same Reader	Diff. Reader
Precision	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2	σ_7^2	σ_8^2

σ_1^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{within-scanner}} + \sigma^2_{\text{within-SWS}} + \sigma^2_{\text{within-reader}}$
σ_2^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{within-scanner}} + \sigma^2_{\text{within-SWS}} + \sigma^2_{\text{betw-reader}}$
σ_3^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{within-scanner}} + \sigma^2_{\text{betw-SWS}} + \sigma^2_{\text{within-reader}}$
σ_4^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{within-scanner}} + \sigma^2_{\text{betw-SWS}} + \sigma^2_{\text{betw-reader}}$
σ_5^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{betw-scanner}} + \sigma^2_{\text{within-SWS}} + \sigma^2_{\text{within-reader}}$
σ_6^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{betw-scanner}} + \sigma^2_{\text{within-SWS}} + \sigma^2_{\text{betw-reader}}$
σ_7^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{betw-scanner}} + \sigma^2_{\text{betw-SWS}} + \sigma^2_{\text{within-reader}}$
σ_8^2	$\sigma^{\mathbf{2}}_{\text{within-subject}} + \sigma^2_{\text{betw-scanner}} + \sigma^2_{\text{betw-SWS}} + \sigma^2_{\text{betw-reader}}$

An estimate of the precision is

$$1.96\sqrt{2\sigma_\varepsilon^2} = 2.77\sigma_\varepsilon.$$

where:

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n}\sum_{k=1}^{k}(Y_{ik} - \bar{Y}_i)^2 / n(K-1)$$

and

$$\bar{Y}_i = \sum_{k=1}^{K} Y_{ik} \,/\, K$$

is the average over K replications for case i (i=1, 2, … n).

4. Statistical Analyses for Testing Compliance

Note: In the longitudinal claims when no only the size of the lesion will change, but also other characteristics (e.g. shape of the mass), then must take additional bias of the new characteristic into account in the precision value estimate.