Published in final edited form as:

Acad Radiol. 2018 March ; 25(3): 372–379. doi:10.1016/j.acra.2017.09.023.

Interpreting Change in Quantitative Imaging Biomarkers

Nancy A. Obuchowski, PhD

Quantitative Health Sciences /JJN3, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195-5196

Abstract

Introduction—Quantitative imaging biomarkers (QIBs) are becoming increasingly adopted into clinical practice to monitor changes in patients' conditions. The repeatability coefficient (RC) is the clinical cut-point used to discern between changes in a biomarker's measurements due to measurement error versus changes that exceed measurement error, thus indicating real change in the patient. Imaging biomarkers have characteristics that make them difficult for estimating the repeatability coefficient, including non-constant error, non-Gaussian distributions, and measurement error that must be estimated from small studies.

Methods—We conducted a Monte Carlo simulation study to investigate how well three statistical methods for estimating the repeatability coefficient perform under five settings common for QIBs.

Results—When the measurement error is constant and replicates are normally distributed, all of the statistical methods perform well. When the measurement error is proportional to the true value, approaches that use the log transformation or coefficient of variation perform similarly. For other common settings, none of the methods for estimating the repeatability coefficient perform adequately.

Conclusion—Many of the common approaches to estimating the repeatability coefficient perform well for only limited scenarios. The optimal approach depends strongly on the pattern of the within-subject variability; thus, a precision profile is critical in evaluating the technical performance of QIBs. Asymmetric bounds for detecting regression vs. progression can be implemented and should be used when clinically appropriate.

Keywords

Quantitative Imaging Biomarker; repeatability; repeatability coefficient; measurement error

1. Introduction

As quantitative imaging biomarkers (QIBs) become increasingly adopted into clinical practice for diagnosis, prognosis, and disease monitoring, it's critical that clinicians be able to interpret them properly. Clinicians must be able to discern between changes in a

corresponding author: obuchon@ccf.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

biomarker's measurements that are expected due to measurement error by the imaging system versus changes that exceed measurement error and thus indicate a real change in the patient.

Studies of the technical performance of QIBs, particularly studies of their repeatability, are used to help radiologists interpret change. Organizations such as the Quantitative Imaging Biomarker Alliance (QIBA) conduct groundwork studies to estimate biomarkers' performance. They also perform meta-analyses to summarize biomarkers' performance over multiple published studies [1, 2]. From these studies of repeatability, investigators calculate a cutpoint, or threshold, for discerning when a measured change is attributable to measurement error vs. when the measured change should be interpreted as a real change in the patient, with some stated degree of confidence [3-5]. For illustration, when measuring the change in volume of a pulmonary lesion from baseline, a measured change <25% might be considered purely measurement error, whereas a measured change exceeding 25% might be considered a real change, with 95% confidence [6].

In this paper we present the results of a Monte Carlo simulation study investigating how well these clinical cutpoints perform with respect to their declared confidence levels. We consider scenarios common for imaging biomarkers, including non-constant measurement error, non-Gaussian distributions, and small technical performance studies from which the cutpoints are estimated. We consider three statistical methods for estimating the cutpoint. We investigate the performance of the cutpoints when no real change has occurred and also when a real change has occurred (both regression and progression). Based on these findings we provide recommendations for how best to estimate the cutpoint for interpreting change in QIBs.

2. Methods

2.1 Defining change

There are three common ways in which change in a quantitative biomarker can be defined. Let Y_{ib} denote the biomarker measurement taken on the i^{th} subject at baseline and let Y_{it} denote the biomarker measurement taken on the same subject at some follow-up timepoint. Table 1 summarizes three ways to define change: (1) difference from baseline expressed in the original units of the biomarker, $\hat{d}$, (2) percentage change from baseline, $\% \hat{d}_b$, and (3) percentage difference in the two measurements, $\% \hat{d}_{mean}$. We will examine cutpoints for each of these in our study.

2.2 Methods for estimating cutpoints from technical performance studies

The biomarker measurements are assumed to be a function of the true biomarker value, with some possible bias and measurement error. Let X_{ik} denote the true biomarker value at timepoint k for subject i . The biomarker measurement is written as a linear function of the true value [3]:

$$Y_{ik} = \beta_0 + \beta_1 X_{ik} + \varepsilon_{ik} \quad (4)$$

where β_o is the fixed bias, β_I is the proportional bias, and ϵ_{ik} is the measurement error. It is often assumed that the measurement error is constant and follows a normal distribution, $\epsilon_{ik} \sim N(0, \sigma^2)$. We vary these assumptions in our Monte Carlo simulation study in order to consider scenarios typical in imaging. The measurement error, σ^2 , is often referred to as the within-subject variance. Its square root is referred to as the within-subject standard deviation, or wSD [3,4].

To estimate the cutpoint at which changes in the biomarker measurements can be considered real changes, technical performance studies are conducted to estimate σ^2 , the within-subject variance. These studies, often called test-retest studies, are designed so that a subject is imaged in the same fashion on multiple occasions over a short period of time to ensure that no biological change has occurred [4]. Since it is often difficult to justify imaging a subject multiple times over a short period of time (often the same day), these studies are usually limited to two replicates per subject [3, 4]. Obuchowski and Bullen [7] have recommended that at least 35 test-retest subjects are needed for providing 95% coverage of the true biomarker value of future subjects. For a test-retest study with N subjects and two replicates per subject, an estimate of the within-subject variance is:

$$\hat{\sigma}^2 = \sum_{i=1}^N \{Y_{i1} - Y_{i2}\}^2 / 2N, \quad (5)$$

where Y_{ik} denotes the biomarker measurement for the k^{th} replicate on the i^{th} subject. Assuming constant within-subject variance and normality, the cutpoint associated with 95% confidence is then

$$\widehat{RC} = 1.96 \times \sqrt{2\hat{\sigma}^2} \quad (6)$$

where RC is the repeatability coefficient, or the least significant difference between two repeat measurements taken under identical conditions [3-5, 8]. Changes in the biomarker that exceed $+\widehat{RC}$ or are less than $-\widehat{RC}$ are considered real changes (not due to just measurement error), with 95% confidence. Specifically, measured differences $< -\widehat{RC}$ are considered true regression, while measured differences $> +\widehat{RC}$ are considered true progression. We refer to this simple approach as Approach 1 (see first row of Table 2).

The cut-points in Approach 1 are symmetric, such that a similar magnitude of progression or regression is considered true change (see Figure 1). Consider an example for lung tumor volume measurements. From a prior technical performance study, the RC was estimated as 150mm^3 (i.e. $\hat{\sigma} = 54$). Now suppose that in a future subject the measured volume of a lung nodule is 500mm^3 at baseline, and after treatment the measured volume is 300mm^3 . The measured change is $(300-500)$, or a 200mm^3 reduction in volume. Is this difference due to measurement error or should we conclude that a real change has occurred? Since $-200 < -\widehat{RC}$, we would conclude that a real change has occurred with 95% confidence.

It is common for biomarkers' wSD to be correlated with the magnitude of the biomarker [8]. When there is a relationship between the wSD and the magnitude of the measurements, there

are two common approaches to minimizing this relationship [2, 3, 8]: (i) within-subject coefficient of variation (wCV) approach, and (ii) log transformation. These approaches lead to symmetric and asymmetric intervals, respectively (see Figure 1). We discuss these approaches next.

From the technical performance study, the wCV is estimated by first calculating the mean of the replicate measurements for each of the N subjects. Denote the mean for the i^{th} subject as $\bar{Y}_i = (Y_{i1} + Y_{i2})/2$. The estimate of the wCV² is

$$\widehat{wCV}^2 = \sum_{i=1}^N \{(Y_{i1} - Y_{i2})^2 / (2\bar{Y}_i^2)\} / N \quad (7)$$

The %RC is estimated as

$$\%RC = 1.96 \times \sqrt{2\widehat{wCV}^2}$$

and the symmetric cutpoints are defined as Approach 2 in Table 2. Note that %RC could be used to define a cutpoint for the percentage change from baseline (% $\hat{d}_b$) or for the percentage difference (% $\hat{d}_{mean}$). These are presented as Approaches 2A and 2B in Table 2, respectively.

In the log transformation approach the variance of the log transformed data is calculated as [8, 9]:

$$\hat{\sigma}_{ln}^2 = \sum_{i=1}^N \{\ln(Y_{i1}) - \ln(Y_{i2})\}^2 / 2N. \quad (8)$$

The cutpoint associated with 95% confidence is

$$\widehat{RC}_{ln} = 1.96 \times \sqrt{2\hat{\sigma}_{ln}^2}. \quad (9)$$

Back-transforming, we get the asymmetric cutpoints (see Figure 1):

$$RC_{ln,lower} = (\exp(-\widehat{RC}_{ln}) - 1) \text{ and } RC_{ln,upper} = (\exp(+\widehat{RC}_{ln}) - 1). \quad (10)$$

Approaches 3A and 3B in Table 2 describe how these cutpoints are used to identify real change in the percentage change from baseline (% $\hat{d}_b$) and a real change in the percentage difference (% $\hat{d}_{mean}$).

2.2 QIB Repeatability Data

For evaluating the above methods of estimating RC, we conducted a Monte Carlo simulation study (described in detail in Section 2.3). In order to generate realistic data for the simulation study, we first examined some QIB data.

In the literature we found evidence that QIBs commonly have a within-subject variance that varies proportionately with the magnitude of the biomarker [8, 9, 10, 11, 12]. There is also some evidence that for some biomarkers, there is a range where the within-subject variance is modest, but outside the range the variance is much higher [8]. There is very little literature on the distribution of replicate measurements from a subject. Ideally, QIB repeatability data would come from clinical studies where real subjects are each scanned multiple times using the same imaging methods within a short period of time; the QIB is then measured from each scan. For several reasons, including costs and ethical concerns about patient radiation and side-effects of contrast agents, these studies rarely involve more than two repeat scans per subject [4]. Although we can estimate wSD and wCV from these studies, we cannot reliably determine the distribution of the replicate measurements from a subject. Bland and Altman [8] suggest that measurement error is likely to follow a normal distribution. To confirm this, we relied on two other types of studies for guidance: phantom studies and reproducibility studies.

With phantoms, it is easy to obtain multiple scans on the same simulated target and to assess the distribution of the measurements. The disadvantage is that phantoms do not provide the biological variation inherent in a human subject. Nevertheless, we examined the distribution of replicate measurements from two phantom studies, assuming that the distribution of measurements from repeated imaging of simulated objects is similar to the distribution of replicate scans of human subjects. We recognize, though, that the magnitude of the variability is likely underestimated in phantoms [3,4].

In reproducibility studies, multiple image processing algorithms and/or multiple radiologists perform the QIB measurements on a single scan of a human subject. The distribution of the QIB measurements from multiple algorithms and/or radiologists on the same subject may also be similar to the distribution of replicate scans subjects. Thus, data from two reproducibility studies were also examined.

Table 3 summarizes results from the phantom and clinical reproducibility studies available to us. The results suggest that the distribution of intra-subject replicate measurements commonly follows a normal distribution except sometimes there can be clustering at extreme values, for example when the imaging system cannot measure the biomarker at very low (or very high) values.

2.3 Monte Carlo Simulation Study

We simulated data from five scenarios (see Table 4). The first two scenarios represent best-case scenarios, where the replicates within a subject are normally distributed (scenario 1) or log-normally distributed (scenario 2). The third scenario represents a positive-valued QIB with many values at zero (to simulate the clustering of values seen at extreme values when the imaging system is unable to measure the true biomarker value). Scenario 4 represents the

situation where the wSD is proportional to the magnitude of the biomarker. For this scenario we simulated the data such that the wCV was constant over the range of biomarker values. Finally, scenario 5 represents the situation where measurements in mid-range have a constant within-subject SD but outside of the mid-range, the SD is much higher. We arbitrarily set the mid-range to ± 1 SD around the mean true value.

We evaluated the statistical approaches summarized in Table 2 under three scenarios: i) there is no true change in the biomarker, ii) there is true progression in the biomarker value (defined as an increase in the true value of the biomarker of two times the within-subject SD), and iii) there is true regression in the biomarker value (defined as a decrease in the true value of the biomarker of two times the wSD). The following steps were performed:

1. A technical performance study was conducted to estimate RC, %RC, and RC_{In} . We simulated data from Equation 4. We started with N=35 subjects and 2 replicates per subject, but we also considered a larger technical performance study with N=100 test-retest subjects. From the technical study the cutpoints for Approaches 1-3 (from Table 2) were estimated.
2. 100 new future patients were simulated using Equation 4. Each patient had two measurements (baseline and follow-up). Change was estimated in different ways using Equations 1, 2 and 3 (defined in Table 1).
3. The proportion of new future patients whose change exceeds the cutpoint from step 1 was recorded for each of the methods.
4. Steps 1-3 were repeated 1000 times (thus 100,000 total replications).

3. Results

Figure 2 describes the situation when there is no true change in the biomarker and technical performance studies of N=35 were performed. The figure illustrates the proportion of subjects inside the “measurement error” range for the five scenarios. Similarly, Figure 3 illustrates the no true change situation but when the cutpoints are estimated from a much larger technical performance study of N=100. Figure 4 and 5 represent the power to detect progression and regression in the biomarker’s true value, respectively. We discuss each scenario separately.

3.1 Scenario 1 (Constant wSD and normally distributed replicates)

All five approaches performed near the nominal level of 95% when the technical performance study was large, dropping off only slightly to 94% when the technical performance study was based on N=35.

The power to detect real progression and regression varied because of how change was defined (i.e. % change from baseline vs. % difference between the two measurements) and because of the asymmetry of the log-based approach. Summing the power over the progression and regression scenarios, the five methods performed very similarly, with Approach 1 performing the best (60.4% overall), and Approach 2A demonstrating the lowest overall power (58.1%).

3.2 Scenario 2 (Constant wSD and log-normally distributed replicates)

Even with a large technical performance study, the coverage of Approaches 2A and 2B was only 92.6% and 91.1%, respectively. The coverage of Approaches 3A and 3B was closer to the nominal level at 93.6% and 93.9%, respectively, while Approach 1 was conservative at 97.2%. For a more practically-sized ($N=35$) technical performance study, Approaches 2A and 2B dropped to 91.8% and 90.7%, and Approaches 3A and 3B dropped to 92.6% and 92.8%, respectively. Approach 1 was at the nominal level (95.1%).

The power to detect true change was low for all five methods, but among the three methods closest to nominal coverage for the no change scenario, Approach 3A had the highest power (14.9% overall), followed by Approach 3B (14.6%), and then Approach 1 at 9.9%.

3.3 Scenario 3 (Clustering of values at one extreme)

Approaches 2A, 3A, and 3B had very poor coverage of only about 80%. Approach 2B was very conservative (>99%). Approach 1 performed closest to the nominal level (93.9% with a large technical performance study and 93.2% with a typically sized study) and demonstrated reasonable power (51.7% overall).

3.4 Scenario 4 (Constant wCV and normally distributed replicates)

All five methods performed near the nominal coverage, even when the technical performance study is small. Approaches 2B and 3A each had coverage of 94.1%, while the other approaches were slightly less than 94%.

All five methods also had good power. Approaches 2B and 3A had overall power of 60.6% and 60.4%, while the other methods had power slightly less.

3.5 Scenario 5 ("Sweet-spot" with high wSD at each end and normally distributed replicates)

For the no-change scenario with a practically-sized technical performance study, Approach 3 performed the best with coverage of 94.6% (for 3A) and 93.3% (for 3B). Approach 1 had coverage of 92.8% and Approach 2 had the lowest coverage (91.7% for 2A and 92.5% for 2B).

Among the three best performing approaches, the overall power was highest for Approach 3B (46.3%), followed by Approach 1 (44.6%), then Approach 3A (33.0%).

4. Discussion

As quantitative imaging biomarkers (QIBs) become increasingly adopted into clinical practice, it's critical that investigators and clinicians are provided simple and intuitive tools for measuring and interpreting them. In this study we focused on tools for discriminating between real change in a patient from change due to measurement noise. We found that none of the commonly used statistical methods perform well in all scenarios. The best method depends on the pattern of the measurement error, as well as the statistical distribution of replicate measurements. For QIBs, the pattern of error can be quite complicated and often

the distribution of replicate measurements is difficult to determine because of small sample size. Thus, our recommendations for detecting change must be based on the type of information typically available to investigators.

Our study results emphasize the importance of thoroughly investigating the pattern of measurement error of QIBs. This is the first recommendation from our work. The pattern of the measurement error can be assessed by constructing precision profiles [3, 8], i.e. tables or figures that illustrate estimates of the wSD (or wCV) over a range of variables, e.g. the magnitude of the measurement. For example, for lung CT, the wCV of volume measurements of nodules is not constant for all nodule sizes; the wCV tends to decrease as the nodule size increases [11]. Regarding the distribution of replicate measurements, many investigators make the mistake of taking the distribution of the QIB measurements over the subjects as the assumed distribution of replicate measurements within a subject. Bland and Altman [8] also point out this misunderstanding. In our study for all 5 scenarios considered we used a normal distribution to describe the QIB values in the population, i.e. the distribution of values between subjects. We varied, however, the distribution of the errors within a subject to emphasize differences in method performance based on this within-subject variability.

Another issue to consider in deciding on an optimal statistical method for identifying real change is the desirability of an asymmetric range, i.e. different magnitudes for the cutpoints for detecting progression and regression. With the log-based approach (Approach 3), an asymmetric range is generated due to the log transformation and exponentiation back to the original units of the biomarker. Weber et al [9] explain that the asymmetric range is appropriate when changes are measured relative to a baseline value. They present the scenario where a decrease from 5 to 4 represents a 20% decrease, whereas a change from 4 to 5 represents a 25% increase. Since the difference is the same, they argue that an asymmetric range is needed. They support an asymmetric range for standardized uptake values (SUVs) for monitoring lung cancer tumors with the lower bound of -28% and the upper bound of +39%. There could also be clinical reasons for an asymmetric range. When there are different clinical consequences for detecting progression vs. regression (e.g. detection of progression changes treatment, whereas detection of regression is associated with no change in patient management), these clinical needs should be taken into account. Approaches 1 and 2 can be modified to produce an asymmetric range based on clinical demands. For example, we might choose to allocate our type I error rate as $a:b$ for regression:progression, where $a+b=0.05$, and $a<b$ if detection of progression is more important. The modified cutpoints for Approach 1 become (and similarly for Approach 2):

$$\widehat{RC}_{lower}=Z_a \times \sqrt{2\hat{\sigma}^2} \quad \widehat{RC}_{upper}=Z_b \times \sqrt{2\hat{\sigma}^2}.$$

Recall that for a symmetric range, $a=b=0.025$ and the Z_a and Z_b values are -1.96 and +1.96. Consider an example. Suppose that detection of progression is more important than detection of regression. We might choose to allocate our type I error rate as 0.01:0.04. The new values for Z_a and Z_b would be -2.328 and +1.751. Using our previous lung CT volume example, where $\hat{\sigma} = 54$, recall that the symmetric range was [-150, +150]. The asymmetric

range would be [-178, +134]. Note that we evaluated the coverage of these asymmetrics. We found similar coverage using the asymmetric range as we reported for symmetric ranges in Section 3.

A final issue that we considered is the performance of the statistical methods for subgroups of patients. In Section 3 we presented the coverage probability and power of the statistical methods pooled over all future subjects. Based on these results, Approach 1 appeared to perform well even in situations where the assumption of constant wSD was not met. For example, in scenario 4 where the wSD changes with the magnitude of the measurements, Approach 1 had a coverage probability of 93.9%, which is very similar to the coverage probability of other methods that take into account the changing magnitude of the wSD (i.e. 93.6%-94.1% for Approaches 2 and 3). When we investigated the performance of Approach 1 for subjects with low and high measurements, however, we found that the performance varied greatly. Specifically, for subjects with low measurements, the coverage was 100%, while for subjects with high measurements, the coverage was only 80%, with an overall pooled average of 93.9%. Note that the coverage probability for Approaches 2 and 3 varied little for these subgroups of subjects.

Taking into account all of these various factors, we make the following recommendations for statistical methods for detecting change in QIBs. Based on the precision profile of the QIB, a determination needs to be made of whether or not the wSD is reasonably constant over the whole range of the QIB or even over smaller ranges of the QIB (e.g. Scenarios 3 and 5 in our simulation study). When the wSD is constant for identifiable ranges and there is no reason (biological or mathematical) to expect that replicate measurements will deviate seriously from a normal distribution, Approach 1 can be used, as it is simple and has good performance. In contrast, when the wSD varies with the magnitude of the biomarker measurements, Approach 2 or 3 should be considered. Real scenarios are likely to be more complex than in our simulations. For example, in lung CT where the wCV of nodule volume measurements decreases as volume increases, nodules can be divided into ranges of small, medium, and large, where the wCV is fairly constant within a range [6]. The %RC would be estimated for each range. Nodules of different sizes at baseline would then have different cut-points for discriminating between real change in the nodule from measurement error.

In our Monte Carlo simulation study we evaluated three common approaches to detecting change in a QIB. Bland and Altman discussed a fourth modeling-based approach [8]. Their modeling approach involves constructing a model for wSD^2 as a function of the magnitude of the biomarker value. Bland and Altman propose modeling the difference between the measurements in order to estimate the limits of agreement. Our application is slightly different so we considered an approach that directly modeled the wSD^2 , but the general idea is the same. The general model is

$$\hat{\sigma}_i^2 = b_0 + b_1(\bar{Y}_i)$$

where the estimated within-subject variance of subject i is a function of the mean of the measurements from subject i . A quadratic term could also be added and tested using

ordinary least squares regression. The estimated regression coefficients, $\hat{b}_0$ and $\hat{b}_1$, are then used to define dynamic cutpoints, i.e. cutpoints that vary with a future subject's baseline measurement. For example, for the i -th future subject, the cutpoint associated with 95% confidence is then

$$\widehat{RC}_{dynamic} = 1.96 \times \sqrt{2(\hat{b}_0 + \hat{b}_1 Y_{ib})}$$

where $\hat{b}_0$ and $\hat{b}_1$ are estimated from a technical performance study of a biomarker's repeatability, Y_{ib} is the baseline measurement for future subject i . When we evaluated this approach in our Monte Carlo simulation study with $N=35$, this model approach did not have nominal coverage, with coverage of only 89-92%. With a larger sample size ($N=100$), the coverage improved somewhat but it did not reach the nominal level. Thus, this sort of approach might be extremely useful for QIBs but much larger technical performance studies than are typically performed now would be needed.

Our study has several limitations. We explored only five scenarios in our simulation study. While we believe that they cover many situations in quantitative imaging, other plausible scenarios were not explored. Also, while we were able to identify statistical methods appropriate for some scenarios, there was no single statistical method that worked well in all scenarios. More research is needed in developing statistical methods to estimate robust cutpoints for discriminating real change from measurement error.

Acknowledgments

This work was funded in part with federal funds from the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Department of Health and Human Services, under contracts HHSN268201300071C and HHSN268201500021C.

References

1. Radiological Society of North America (RSNA). Quantitative Imaging Biomarkers Alliance. www.rsna.org/QIBA/
2. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. *Radiology*. 2015; 277:813–825. [PubMed: 26267831]
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.
4. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative Imaging Biomarkers: A Review of Statistical Methods for Computer Algorithm Comparisons. *SMMR*. 2015; 24:68–106.
5. Barnhart HX, Barboriak DP. Applications of the Repeatability of Quantitative Imaging Biomarkers: A review of statistical analysis of repeat data sets. *Transl Oncol*. 2009; 2:231–235. [PubMed: 19956383]
6. <http://qibawiki.rsna.org/index.php/Profiles>
7. Obuchowski NA, Bullen J. Quantitative Imaging Biomarkers: Effect of Sample Size and Bias on Confidence Interval Coverage. *Statistical Methods in Medical Research*. 2017 In press.
8. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999; 8:135–160. [PubMed: 10501650]
9. Weber WA, Gatsonis CA, Mozley PD, Hanna LG, Shields AF, Aberle DR, Govindan R, Torigan DA, Karp JS, Yu JQ, Subramaniam RM, Halvorsen RA, Siegel BA. Repeatability of ^{18}F -FDG

- PET/CT in advanced non-small cell lung cancer: prospective assessment in 2 multicenter trials. *J Nucl Med.* 2015; 56:1137–1143. [PubMed: 25908829]
10. Winfield JM, Tunariu N, Rata M, et al. Extracranial soft-tissue tumors: repeatability of apparent diffusion coefficient estimates from diffusion-weighted MR imaging. *Radiology.* 2017;14. [PubMed: 28318443]
 11. Obuchowski NA, Barnhart HX, Buckler AJ, Pennello G, Wang XF, Kalpathy-Cramer J, Kim HJ, Reeves AP. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. *Stat Methods Med Res.* 2015; 24:107–140. [PubMed: 24919828]
 12. Lodge M. Repeatability of SUV in Oncologic ^{18}F -FDG PET. *J Nucl Med.* 2017; 58:523–532. [PubMed: 28232605]

Figure 1.

Illustration of lower and upper bounds of RC based on three statistical approaches. The green range is based on Approach 1 where the difference between two measurements (i.e. $\hat{d}$) is quantified in the original units of the biomarker (mm³ used here for illustration). Measurement error is distinguished from real change based on estimates of the wSD. The red and blue ranges are based on statistical Approaches 2 and 3, respectively, where the percent change from baseline (i.e. $\hat{d}_b$) or the percent difference in the two measurements (i.e. $\hat{d}_{mean}$) is quantified. Measurement error is distinguished from real change based on estimates of either the wCV (leading to a symmetric range) or the wSD after log transformation (leading to an asymmetric range).

Figure 2.

Coverage of the RC range when no true change in the subject has occurred. The different statistical methods for calculating RC described in Section 2.1 are illustrated in green (Approach 1), shades of red (Approach 2A and 2B), and shades of blue (Approach 3A and 3B). The nominal coverage is 95%. The repeatability metrics were estimated from technical performance studies of N=35. The five scenarios are: 1=constant wSD with normally distributed replicates; 2=constant wSD with log-normally distributed replicates; 3=clustering of values at one extreme where the imaging system is unable to measure; 4=constant wCV with normally distributed replicates; and 5=wSD “sweet spot” in the mid-range with large wSD at both extremes.

Figure 3.

Coverage of the RC range when no true change in the subject has occurred. RC was estimated from technical performance studies with $N=100$. The different statistical methods for estimating RC described in Section 2.1 are illustrated in green, shades of red, and shades of blue, respectively. The nominal coverage is 95%.

Figure 4.

Proportion of subjects falling outside the RC range when a true increase in the biomarker has occurred. RC was estimated from technical performance studies with $N=35$. The different statistical methods for estimating RC are illustrated in green, shades of red, and shades of blue, respectively.

Figure 5.

Proportion of subjects falling outside the RC range when a true decrease in the biomarker has occurred. RC was estimated from technical performance studies with $N=35$. The different statistical methods for estimating RC are illustrated in green, shades of red, and shades of blue, respectively. Note that for scenario 3, method 3a demonstrated power slightly over 50%.

Table 1

Three Definitions of Change in QIBs

Difference from baseline	$\hat{d} = (Y_{it} - Y_{ib})$ (1)
Percentage change from baseline	$\% \hat{d}_b = (Y_{it} - Y_{ib}) / Y_{ib} \times 100$ (2)
Percentage difference	$\% \hat{d}_{mean} = \frac{(Y_{it} - Y_{ib})}{(Y_{ib} + Y_{it}) / 2} \times 100$ (3)

* where positive values are indicative of an increase in the biomarker value at follow-up (referred to here as “progression” without loss of generality) and negative values are indicative of a decrease in the biomarker value at follow-up (“regression”).

Table 2

Three Approaches to Discerning Measurement Error from Real Change

	“The observed change in QIB is attributable to measurement error.”	“The observed change in the QIB represents true change in the patient, with 95% confidence.”
Approach 1	$-\widehat{RC} < \hat{d} < +\widehat{RC}$	$\hat{d} < -\widehat{RC} \text{ or } \hat{d} > +\widehat{RC}$
Approach 2A	$-\% \widehat{RC} < \% \hat{d}_b < +\% \widehat{RC}$	$\% \hat{d}_b < -\% \widehat{RC} \text{ or } \% \hat{d}_b > +\% \widehat{RC}$
Approach 2B	$-\% \widehat{RC} < \% \hat{d}_{mean} < +\% \widehat{RC}$	$\% \hat{d}_{mean} < -\% \widehat{RC} \text{ or } \% \hat{d}_{mean} > +\% \widehat{RC}$
Approach 3A	$RC_{lower} < \% \hat{d}_b < RC_{upper}$	$\% \hat{d}_b < RC_{lower} \text{ or } \% \hat{d}_b > RC_{upper}$
Approach 3B	$\% RC_{lower} < \% \hat{d}_{mean} < \% RC_{upper}$	$\% \hat{d}_{mean} < \% RC_{lower} \text{ or } \% \hat{d}_{mean} > \% RC_{upper}$

Table 3

Assessment of the Distribution of Intra-Subject Repeated Measurements

QIB	Setting	# observations	Distribution for intra-subject replicates
Magnetic Resonance Imaging T1 values	Phantom study with replicate scans of each target	15 replicates per target	Close to normally distributed except for targets at low T1 values where there's clustering at the minimum values
Ultrasound Shear wave speed	Phantom study with replicate scans of each target	10 replicates per target	Close to normally distributed
Magnetic Resonance Elastography values	Clinical study with 1 scan per subject, evaluated by multiple algorithms	10 replicates per subject	Close to normally distributed except for subjects with low values where there's clustering at these low values.
Computed Tomography acetabular angle	Clinical study with 1 scan per patient, interpreted by multiple readers	4 replicates per subject	Close to normally distributed

Table 4

Five Scenarios for Monte Carlo Simulation

Scenario	Description	Distribution of True Value of Biomarker ¹	Distribution of Replicate Measurements within Subject ²
1	QIB with constant wSD and normally distributed replicates	$X \sim N(100, 100)$	$\epsilon \sim N(0, 100)$
2	QIB with constant wSD and log-normally distributed replicates	$X \sim N(100, 100)$	$\epsilon = \exp(Z)$ where $Z \sim N(0, 10)$
3	QIB with clustering at one extreme ³	$X \sim N(10, 100)$	$\epsilon \sim N(0, 100)$, but values less than zero were set to zero
4	QIB with constant wCV and normally distributed replicates	$X \sim N(20, 10)$	$\epsilon \sim N(0, \sigma^2)$, with σ increasing with true biomarker value: $\sigma = (0.1) \times (X)$
5	QIB where wSD is larger for both small and large values of QIB ⁴	$X \sim N(20, 10)$	$\epsilon \sim N(0, \sigma^2)$, with $\sigma^2 = 100$ for $X < 17$ or > 23 ; otherwise $\sigma^2 = 20$ and values less than zero were set to zero

¹The notation $X \sim N(\mu, \lambda)$ denotes that the true value of the biomarker, X is a realized value from a normal distribution with mean μ and between-subject variance of λ .

²The notation $\epsilon \sim N(\nu, \xi)$ denotes that the measurement error of the biomarker is a realized value from a normal distribution with mean ν and within-subject variance of ξ .

³About 1/3 of subjects have one or both measurements equal to zero under this scenario. When a measurement was zero, the value was set to 0.001 for division by the measurement or for taking the log of the measurement.

⁴About 1% of subjects have one or both measurements equal to zero under this scenario.