INTRODUCTION

Medical imaging modalities such as CT, MRI, and positron emission tomography (PET) are routinely used in clinical practice for disease screening, diagnosis, staging, therapeutic monitoring, evaluation of residual disease, and assessment of disease recurrence. Traditionally, image contrast-based qualitative interpretations of medical images are the most commonly employed radiology practice. With advances in imaging technologies in recent years, imaging metrics that can quantify tissue biological and physiological properties, in addition to those that quantify tissue morphology such as disease size, are increasingly used in research and early phase clinical trials to characterize disease and response to treatment. A recent review by a group of principal investigators from the Quantitative Imaging Network (National Cancer Institute, National Institutes of Health) has called for wider incorporation of quantitative imaging methods into clinical trials, and ultimately, clinical practice. Because many QIBs are derived through multiple steps in image data acquisition and data analysis, QIB measurements can produce large variabilities, posing a significant challenge in translating QIBs into clinical trials, and ultimately, clinical practice. Both repeatability and reproducibility constitute the reliability of a QIB measurement. In this article, we review the statistical aspects of repeatability and reproducibility of QIB measurements by introducing methods and metrics for assessments of QIB repeatability and reproducibility and illustrating the impact of QIB measurement error on sample size and statistical power calculations, as well as predictive performance with a QIB as a predictive biomarker.

Both repeatability and reproducibility constitute the reliability of a QIB measurement. Repeatability refers to the precision of a QIB measured under identical conditions (e.g., using the same measurement procedure, same measurement system, same image analysis algorithm, and same location over a short period of time; also known as repeatability condition), which is mainly a measure of the within-subject variability. On the other hand, reproducibility refers to the precision of a QIB measured using different scanners over a long period of time, which is mainly a measure of the between-subject variability. Following this, we introduce statistical considerations for both repeatability and reproducibility of QIB measurements.
under different experimental conditions\textsuperscript{13,14} (also known as the reproducibility condition), which is mainly a measure of the variability associated with different measurement systems, imaging methods, study sites, and population. In recent years, many studies have been conducted to investigate the reliability of different QIBs. For example, Yokoo et al\textsuperscript{2} studied the precision of hepatic proton-density fat-fraction measurements by using MRI; Lodge\textsuperscript{15} examined the repeatability of standardized uptake value (SUV) in oncologic\textsuperscript{16} 18-fluorodeoxy glucose (FDG) PET studies; Shukla-Dave et al\textsuperscript{17,18} reviewed and emphasized the need for assessment of reproducibility and repeatability of MRI QIBs in oncologic studies; Park et al\textsuperscript{19} reviewed the challenges in reproducibility of quantitative radiomics metrics; Fedorov et al\textsuperscript{14} and Schwier et al\textsuperscript{15} assessed the repeatability of radiomics features from multiparametric MRI of small prostate tumors; Hernando et al\textsuperscript{20} quantified the reproducibility of MRI-based proton-density fat-fraction measurements in a phantom across vendor platforms at different field strengths; Kalpathy et al\textsuperscript{11} investigated the repeatability and reproducibility of tumor volume estimate from CT images of lung cancer; Lin et al\textsuperscript{16} evaluated the repeatability of \textsuperscript{18}F-NaF PET–derived SUV metrics; Baumgartner et al\textsuperscript{21} studied the repeatability of \textsuperscript{18}F-FDG PET brain imaging; Jafer et al.,\textsuperscript{22} Winfield et al.,\textsuperscript{20} Weller et al.,\textsuperscript{21} Lecler et al.,\textsuperscript{22} and Lu et al\textsuperscript{15} estimated the repeatability and reproducibility of the apparent diffusion coefficient (ADC) derived from diffusion-weighted MRI; Hagiwara et al\textsuperscript{23} studied the repeatability and reproducibility of quantitative relaxometry with a multidynamic multiecho MRI sequence using a phantom and normal healthy human subjects; Jafari-Khouzani et al\textsuperscript{24} appraised the repeatability of brain tumor perfusion measurement using dynamic susceptibility contrast MRI; Han et al\textsuperscript{25} studied the repeatability and reproducibility of 2D and 3D hepatic MR elastography markers in healthy volunteers; and Olin et al\textsuperscript{26} investigated the reproducibility of the MR-based attenuation correction factors in PET/MRI and its impact on \textsuperscript{18}F-FDG PET quantification in patients with non-small cell lung cancer.

In this article, we review the statistical aspects on repeatability and reproducibility of QIB. We introduce methods and metrics for assessment of QIB repeatability and reproducibility, and illustrate the impact of QIB measurement error on sample size and statistical power calculations, as well as performance as a predictive biomarker.

**MEASUREMENT ERROR MODEL**

Precision of a QIB is defined as the closeness of agreement between repeated measurements of the QIB,\textsuperscript{2} and repeatability and reproducibility comprise different sources of variability that may impact the precision of a given QIB. Measurement error is defined as the difference between a measured quantity and its true value.\textsuperscript{27} Any sources of variability can cause measurement errors in QIB measurements. Though it is critical to identify and obtain valid inference on the impact of every component of variation (see Section 3), we start by introducing a general model on measurement error. Table 1 lists commonly used symbols in this article.

![Image](https://via.placeholder.com/150)

**Table 1. List of symbols and notations**

| ~ | Distributed as |
|---|---|
| $N(a,b)$ | Normal distribution with mean $a$ and variance $b$ |
| $df$ | Degree of freedom |
| $\chi^2_{df}$ | Quantile function of chi-square distribution |
| $Var$ | Variance |
| $Y_{itl}$ | Measured QIB value from the $l$th measurements of a repeated QIB measurement made at time $t$ for subject $i$ |
| $X_i$ | True QIB value for subject |
| $n$ | Number of subjects included in the study |
| $m$ | Number of replicates for each subject |
| $\mu$ | Mean of $X$ |
| $\epsilon$ | Measurement error |
| $\delta$ | Within-subject error (under repeatability condition) |
| $\gamma$ | Between-condition error (under reproducibility condition) |
| $\gamma\delta$ | Interaction between subject and condition |
| $\sigma^2_X$, $\sigma^2_\epsilon$, $\sigma^2_\delta$, $\sigma^2_{\gamma\delta}$ | Variance of $X$, $\epsilon$, $\delta$, $\gamma$, and $\gamma\delta$, respectively |
| $\sigma^2_\gamma$ | Represents the estimator of a parameter |
| $H_0$ | Null hypothesis |
| $H_1$ | Alternative hypothesis |

QIBs, quantitative imaging biomarkers.

In a measurement error model, instead of the true QIB value, we can only observe QIB values with errors that are random across different QIB measurements. If the errors are constant for all measurements, then the error is called bias.\textsuperscript{2} Since both repeatability and reproducibility mainly concern random errors, we assume that there is no bias and that the random errors are independent and identically distributed with the mean equal to zero and variance of $\sigma^2_\epsilon$. Thus, the only unknown parameter $\sigma^2_\epsilon$ measures the level of variability, with larger values indicating larger variability or worse precision. Let $Y_{itl}$ be the measured QIB value from the $l$th measurements of a repeated QIB measurement made at time $t$ for subject $i$ with $X_0$ being the corresponding true value, the measurement error model can be expressed as

$$Y_{itl} = \beta_0 + \beta_1 X_{itl} + \epsilon_{itl}, \epsilon_{itl} \sim N(0, \sigma^2_\epsilon),$$

(1)

where $\beta_0$ represents the bias of the QIB measurement and $\beta_1$ represents the proportional bias of the QIB measurement. The random error $\epsilon_{itl}$ is assumed to be normally distributed (other commonly used distribution assumptions are discussed below). Because both bias and proportional bias cannot be identified solely through the QIB measurements, it is usually assumed that these values are constant and known in advance from
The terms \( X_i \) are made simplified setting where all measurements for subject \( i \) under experimental condition \( j \) (different experimental conditions may include different measurement systems, imaging methods, and countries/regions), in a similar fashion to model (1) we employ a linear relationship between \( Y_{ijk} \) and \( X_{ij} \), but further breaking down the measurement error \( \epsilon_{ijl} \) into different components of repeatability- and reproducibility-related errors. Similar to model (2), we assume there are no bias or proportional bias, and the model that accounts for both repeatability- and reproducibility-related errors can be written as

\[
Y_{ijl} = X_{ij} + \gamma_j + \epsilon_{ijl}.
\]

(2)

**REPEATABILITY AND REPRODUCIBILITY**

Repeatability and reproducibility represent different components of variability (causing measurement error \( \epsilon_{ijl} \)) in model (1). Repeatability refers to the precision of a QIB measured under identical conditions, while reproducibility refers to the precision of a QIB measured under different experimental conditions. Kessler et al.\(^3\) recommends identifying and evaluating each experimental component separately that contributes to reproducibility-related variability. However, in many real-life settings, it is very difficult to independently investigate each source of variability under the reproducibility condition. Thus, the statistical model introduced below only assumes a single parameter representing all sources of variability associated with the reproducibility condition.

Following the measurement error model (2), we consider a simplified setting where all measurements for subject \( i \) are made at a relatively short time interval so that the true value \( X_i \) remains unchanged. That is, let \( Y_{ijk} \) be the \( k \)th repeated QIB measurement made on subject \( i \) under experimental condition \( j \) (different experimental conditions may include different measurement systems, imaging methods, and countries/regions), in a similar fashion to model (1) we employ a linear relationship between \( Y_{ijk} \) and \( X_{ij} \),\(^4,5,32\) but further breaking down the measurement error \( \epsilon_{ijl} \) into different components of repeatability- and reproducibility-related errors. Similar to model (2), we assume there are no bias or proportional bias, and the model that accounts for both repeatability- and reproducibility-related errors can be written as

\[
Y_{ijk} = X_{ij} + \delta_{ik} + \gamma_j + \epsilon_{ijl}.
\]

(3)

The terms \( \delta_{ik} \), \( \gamma_j \), and \( \epsilon_{ijl} \) represent different components of measurement error caused by within-subject variability (under repeatability condition), between-condition variability (under reproducibility condition), and the interaction between subject and condition, respectively, and we assume they follow normal distributions. In general, the random error variances \( \sigma_\delta^2 \), \( \sigma_\gamma^2 \), \( \sigma_\epsilon^2 \), and \( \sigma_{\delta,\epsilon}^2 \) are the key performance characteristics used in repeatability and reproducibility studies (see details below).

**Repeatability**

Many studies have been conducted to investigate the repeatability of QIBs for different imaging modalities, including but not limited to CT, MRI, PET, and DCE MRI scans are usually required to be performed at least 24 h apart, “coffee-break” experiments where there is only a short break between repeated scans are not always possible. Thus, for repeatability studies with long interval between scans, possible changes in the true values should also be considered in the model.

Specifically, for a test–retest study with \( n \) subjects and \( m \) replicates for each subject, since all experiments are conducted under the same experimental condition, without the loss of generalizability, we set \( j = 1 \) and model (3) becomes

\[
Y_{ijk} = X_{ij} + \delta_{ik} + \gamma_1
\]

(4)

for \( i = 1, \ldots , n \) and \( k = 1, \ldots , m \). The random effect variance \( \sigma_\delta^2 \) is the key performance characteristic used in a repeatability study, with smaller value corresponding to better repeatability. Because we only consider QIB measurements in a single-site and single measurement system study for a test–retest repeatability study, the random error \( \gamma_j \) in (3), which measures the reproducibility-related variability such as between-site variability, becomes a condition-specific systematic error \( \gamma_j \) in (4), representing the condition-specific bias. As discussed in Section 2, the constant bias \( \gamma_1 \) cannot be identified through the test–retest study and should be estimated through a phantom study with ground-truth values. The term \( \{\gamma_1\}_{ij} \) in model (3) vanishes because there is only one study condition and the interaction effect cannot be observed.

Many metrics related to the estimate of within-subject variance \( \sigma_\gamma^2 \) have been proposed to quantify the magnitude of repeatability.\(^4\) Table 2 shows the metrics considered in this article for repeatability measurement. The within-subject standard deviation (wSD) is the most commonly used metrics for assessing repeatability. It is the standard deviation (\( \sigma_\delta \)) of repeat measurements for a single subject. If we assume all subjects have the same \( \sigma_\delta \) and the ground true values \( X_i \) are independent and normally distributed, wSD can be obtained by fitting a linear mixed-effects model with subject-specific random intercepts using maximum likelihood.\(^33\) Although other estimation procedures such as method of moment estimators can also provide consistent estimation of wSD, the maximum likelihood method is usually preferred for small sample size when model (3) is correctly
In model (7), with the extra constraint of $\delta_{ik} > 0$, it is more adequate to assume $\delta_{ik}$ follows log-normal or Weibull distributions with the mean of $\delta_{ik}$ equal to one. Raunig et al. suggest using log-normal distribution so that the log-transformed QIB measurement $\log(Y_{ik})$ is normally distributed (after adjusting for the site-specific bias $\gamma_i$)

$$\log(Y_{ik} - \gamma_i) = \log(X_i) + \log(\delta_{ik}), \log(\delta_{ik}) \sim N\left(0, \frac{\sigma^2}{\delta}ight).$$

(8)

Under model (8), $wCV$ only depends on the log-transformed within-subject variance $\sigma^2_{\delta}$ through the form

$$wCV = \sqrt{e^{\sigma^2_{\delta}} - 1}$$

(9)

Therefore, we can apply any of the estimators of $wSD$ on the log-transformed QIB measurements to obtain valid estimates of $\sigma^2_{\delta}$, and $wCV$ can be estimated by plugging the $\sigma^2_{\delta}$ into model (9). Without the log-normal distribution assumption, by mimicking estimator (5), when $m = 2$, we can still estimate $wCV$ by pooling and averaging the within-subject sample coefficient variation

$$\hat{mCV} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \frac{1}{m-1} \sum_{k=1}^{m} \left(\frac{Y_{ik} - \hat{\gamma}_i}{\hat{\gamma}_i}\right)^2}.$$  

(10)

Both ICC and $wCV$ have the benefit of being dimensionless, which make them useful for comparing quantities measured on different scales.

Instead of a single point estimates of the repeatability metric (either $wSD$, ICC, or $wCV$), it is desirable to make inference on these values through either constructing confidence intervals (CIs) or performing hypothesis testing. Both confidence interval and hypothesis testing involve estimating the distribution of the estimator and thus depend on the choice of the estimation method. For $wSD$ (denoted as $\sigma$ and the corresponding estimator denoted as $\hat{\sigma}$), if the ANOVA-type estimators such as equations (5) and (10) are used, $n(m-1)\hat{\sigma}^2/\sigma^2$ follows $\chi^2$ distribution with a degree of freedom $(df)$ of $n(m-1)$. Thus, the $(1 - \alpha) \times 100\%$ CI of $\sigma$ is

\[\frac{\hat{\sigma}}{1 - ((1 - \alpha)/2)^{(df/(df+2))}} \leq \sigma \leq \frac{\hat{\sigma}}{1 - ((1 - \alpha)/2)^{(df/(df+2))}}.\]
where $\chi^2_{df,\alpha/2}$ and $\chi^2_{df,1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles, respectively, of the $\chi^2$ distribution with degree of freedom $df$. To test whether the level of $wCV$ is greater than a threshold value $c$, we conduct hypothesis test with null hypothesis $H_0: \sigma^2 \leq c^2$ vs alternative hypothesis $H_1: \sigma^2 > c^2$. The corresponding test statistic is $T = \frac{\hat{\sigma}^2}{\frac{\chi^2_{df,\alpha/2}}{n(m-1)\alpha} - \frac{\hat{\sigma}^2}{\frac{\chi^2_{df,1-\alpha/2}}{n(m-1)[1-\alpha]}}}$.

Reproducibility

Reproducibility concerns the consistency or precision of the QIB measurement made on the same subject with the same experimental design but under different experimental conditions, such as different measurement device. Reproducibility of QIBs for different imaging modalities, such as CT, MRI, PET, and PET, have also been extensively studied. Although many experimental factors can be included for the reproducibility study condition, it is practically impossible to consider all conditions for a single reproducibility study. Lawrence et al. proposed a list of conditions that can be tested in reproducibility studies. Depending on which condition is being tested, reproducibility studies can be classified into two categories: (1) repeated measurement design and (2) cohort measurement design. For example, the former can be used to study the variability caused by different scanners, while the latter can be used to study the variability caused by different study sites. Because the within-subject variability is generally embedded in the variability under different experimental conditions, a reproducibility study can generate repeatability results for each experimental condition. Specifically, for a repeated measurement design, subject $i$ is repeatedly measured $m$ times for each of the $j$ experimental conditions; for a cohort measurement design, each subject will be repeatedly measured $m$ times under one of the experimental conditions. That is, a repeated measurement design requires each subject being measured $m \times j$ times, while a cohort measurement design requires each subject only being measured $m$ times.

Model (3) is valid for both experimental designs, and the key performance characteristics is the sum of the random effect variances ($\sigma^2 = \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2$, which represents the total variation under the reproducibility study condition. However, subjects in the cohort measurement design are only measured under a single experimental condition and thus, the subject-condition interaction effect $\gamma_{ij}$ does not exist. In this case, we can assume $\gamma_{ij} = 0$, and the total variation for the cohort measurement design becomes $\sigma^2 + \sigma^2$. Similar to reproducibility studies, either linear mixed effects models or two-way ANOVA can be used to fit the data and obtain valid estimates of $\sigma^2$, $\sigma^2$, and $\sigma^2$. The square root of the total variance $\sigma^2$, denoted as total SD (tSD) (Table 2) or reproducibility SD, can be used as a metric to quantify the magnitude of reproducibility. If estimators based on linear mixed effects models are considered, the sampling distributions for these estimators are unknown and numerical methods such as bootstrap or permutation should be used to make valid inferences on these parameters. On the other hand, if ANOVA-type estimators are considered, all the three terms $df\hat{\sigma}^2$, $df\hat{\sigma}^2$, $df\hat{\sigma}^2$, and $df\hat{\sigma}^2$ follow $\chi^2$ distributions and the corresponding CIs and test statistics can be easily obtained. However, because the sampling distribution of $\hat{\sigma}^2$, which equals to $\hat{\sigma}^2 + \hat{\sigma}^2 + \hat{\sigma}^2$, is unknown, numerical methods are recommended.

In a scenario where only two experimental conditions are being compared, e.g. comparing two scanner platforms, the variance of $\gamma_{ij}$ ($\sigma^2$) is no longer estimable (only $\gamma_{1}$ and $\gamma_{2}$ exist). We then use the agreement between these two experimental conditions as the measure of reproducibility. For such a situation where $m = 1$ for a repeated measurement design, Lawrence and Lin proposed the concordance correlation coefficient (CCC) (Table 2) as an agreement measure of reproducibility, defined as

$$CCC = \frac{\sigma_{12}^2 + \rho_{12}(\gamma_{1} + \gamma_{2})}{\sigma_{1}^2 + \sigma_{2}^2 + (\gamma_{1} + \gamma_{2})^2}$$

to evaluate the QIB agreement between the two experimental conditions, where $\sigma_1$ and $\sigma_2$ are the standard deviations of the measured QIB under experiment condition 1 and 2, respectively, and $\rho_{12}$ is the Pearson correlation between the measured QIB values $Y_1$ and $Y_2$. Similar to Pearson correlation, CCC is also ranged between $-1$ and 1, with values close to 1 (or $-1$) representing good concordance (or good discordance) and 0 representing no correlation.

The method introduced by Lawrence and Lin is commonly used to estimate CCC and its CI. Lin and Williamson provide a simple method to perform sample size calculation of CCC.

Plots for repeatability and reproducibility studies

Varye plots can be used to visually study the impact of measurement errors on repeatability and reproducibility studies. For reproducibility studies, although Box-Whisker plots can provide information on the variability for each subject, they are usually not feasible for test–retest studies since only few repeated measurements (usually two) are available for each subject. Bland–Altman plots are commonly included in repeatability
studies to visualize trends in variability within measurement intervals. Figure 1 demonstrates the Bland–Altman plots based on 100 simulated data with $X$ generated uniformly at random from interval $0$ and $5$. When gold-standard or reference values are available (see the setting described in Obuchowski et al.32), the difference between the QIB measurements and the corresponding reference values are plotted against the reference values in Bland–Altman plots. Figure 1a illustrates the case of additive error (model (4) with $\sigma = 0.8$), while Figure 1b shows the case of multiplicative error (model (7) with $\sigma = 0.3$). When reference values are not available, e.g. test–retest studies, standard deviations (or differences for the case of $m = 2$) of repeated measurements are plotted against the averages of repeated measurements in Bland–Altman plots. Figure 1c and d show the cases of additive and multiplicative errors based on two repeated measurements and same values of $\sigma$ as in Figure 1a and Figure 1b, respectively. When multiplicative errors are observed (e.g. in Figure 1(b) and (d)), log-transformed QIB measurements should be considered as suggested in Section 3.1.

When only two experimental conditions are being compared, Bland–Altman plots can also be used in reproducibility studies, especially when gold-standard or reference values are not available.4 In addition, scatter plot of experimental Condition 1 vs Condition 2 with a fitted regression line can provide useful visualization of the agreement between these two methods (Figure 2). When there exists more than two experimental conditions, we can follow the same procedure as above for each pair of the experimental conditions.

EXAMPLES OF THE IMPACT OF QIB MEASUREMENT ERRORS ON CLINICAL STUDIES

QIB as trial end point

QIBs can serve as a clinical trial’s endpoint to assess treatment efficacy, where subjects enrolled in the study are scanned before and after treatment, and the difference of the mean QIB measurements over the treatment course is used to determine the efficacy.
of the treatment. Since it is often nearly impossible to perform repeated measurements at a single time point in a longitudinal study, it is difficult to assess repeatability and/or reproducibility-related QIB measurement errors. From model (2), for a setting without repeated measurements, let $Y_{it}$ and $Y_{i2}$ be the QIB measurements (or log-transformed QIB measurements) before and after intervention for subject $i$, respectively, and we further assume that the corresponding true value $X_{it}$ follows normal distribution with mean $\mu_i$ and variance $\sigma_X^2$ for $t = 1, 2$ and $i = 1, \ldots, n$. The common approach to assess treatment efficacy is to test if the mean difference is greater than a threshold value $\epsilon$ so that the difference is practically meaningful (i.e. null hypothesis $H_0$: $\mu_2 - \mu_1 \leq \epsilon$ vs alternative hypothesis $H_1$: $\mu_2 - \mu_1 > \epsilon$). Under model (2), the corresponding test statistic $Z$ is

$$Z = \frac{n^{-1/2} \sum_{i=1}^{n}(Y_{i2} - Y_{i1} - \epsilon)}{\sqrt{2(1-\rho)(\sigma_X^2 + \sigma_\epsilon^2)}} \sim N(0, 1)$$

where $\rho$ is the Pearson correlation (ranged between 0 and 1) between $Y_{i1}$ and $Y_{i2}$. Thus, for a statistically significant level of $\alpha$, the formula of minimum sample size required to achieve $\beta$ power is

$$n = \left(\frac{2(1-\rho)(\sigma_X^2 + \sigma_\epsilon^2)/(\epsilon_0 + \epsilon_1)}{c(\mu_2 - \mu_1)}\right)^2$$

(12)

where $\epsilon_0$ and $\epsilon_1$ are the $\alpha$ and $\beta$ quantiles of standard normal distribution. For example, under the setting of $\epsilon = 0$ and $\mu_2 - \mu_1 = 0.5$, which represents 50% change against no change in hypothesis testing if log-transformed QIB is considered, Figure 3 illustrates the required sample sizes to achieve 80% power with a significance level of 5% for different values of $\sigma_\epsilon$, $\sigma_X$, and $\rho$. From both Figure 3 and equation (12), we can notice that the required sample size is an increasing function of $\sigma_\epsilon^2$ and $\sigma_X^2$, and is a decreasing function of $\rho$.

For a longitudinal study, the correlation parameter $\rho$ measures the level of dependence between the QIB measurements before and after treatment, with longer time interval between the measures generally resulting in smaller $\rho$. Obuchowski et al. showed that the range of $\rho$, which depends on the time interval of the two measurements, is between 0 and $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_\epsilon^2}$. Using equation (12), the range of sample size is from $\left(\frac{2\sigma_X^2(\epsilon_0 + \epsilon_1)}{c(\mu_2 - \mu_1)}\right)^2$ to $\left(\frac{2(\epsilon_0 + \epsilon_1)}{c(\mu_2 - \mu_1)}\right)^2$. Although the required sample size $n$ is a decreasing function of $\rho$ and $\rho$ is a decreasing function of the time interval between the measurements, we cannot jump into the conclusion that studies with smaller time interval require smaller sample sizes. This is because a smaller time interval usually also results in a smaller difference between $\mu_1$ and $\mu_2$.

QIB as predictive biomarker

In addition to serving as clinical trial end points, QIBs can also be used as predictive biomarkers for early prediction of treatment effect, or as intermediate endpoints in multiarm, multistage trials. Under this scenario, it is usually assumed that the true value of QIB is associated with the primary trial end point but we only measure QIB with error (see model (1)). Many methods have been proposed to adjust the measurement error on covariates in regression models. However, using such adjustment requires knowledge of the distribution of measurement errors that can only be obtained from additional studies such as repeatability or reproducibility studies. This requirement, on the one hand, emphasizes the importance of repeatability or reproducibility studies, but on the other hand, may not always be applicable and standard approach that ignores the measurement error is conducted. In this study, we use simulation to illustrate the impact of measurement error when standard approach is used.

Figure 3. Sample size required to achieve 80% power with a significance level of 5% against measurement error standard deviation $\sigma_\epsilon$. The true difference $\mu_2 - \mu_1$ is 0.5, commonly represents 50% change for log-transformed QIB; $\rho = 0.4$, 0.6, or 0.8; $\sigma_X = 1$ or 1.5; and $\epsilon = 0$. QIB, quantitative imaging biomarker.
Because the sample sizes in published studies where QIBs were used as predictive biomarkers are usually small, it is difficult to analytically evaluate the impact of measurement error on QIBs as predictive biomarkers. Based on the study design and assumptions, Monte Carlo simulations can be used to numerically approximate the impact of measurement error. For illustration purposes, we designed our simulations based on the study by Tudorica et al,\(^5\) where DCE-MRI QIBs were used for early prediction of breast cancer response [pathologic complete response (pCR) vs non-pCR] to neoadjuvant chemotherapy (NACT). We denoted \(Z_i\) as the indicator of pCR and \(X_i\) as the true value of a DEC-MRI QIB for subject \(i\). The true DCE-MRI QIB \(X_i\) was generated from normal distribution. Tudorica et al\(^5\) provided a list of DEC-MRI QIBs with their means and SDs for pCR and non-pCR patients. Here, we considered the percent change in QIB \(K_{trans}\) (transfer rate constant) after the first cycle of NACT relative to baseline (pCR: mean = −64%, SD = 9%; non-pCR: mean = −14%, SD = 41%), which showed the best predictive performance for pCR vs non-pCR in that study.\(^5\) Consistent with the sample size of the study,\(^5\) we included a total of 28 subjects in this simulation study. Here, without loss of generality, we can only observe the DEC-MRI QIB with error (model (2)). The best approach was to fit a univariate logistic regression model using the observed DEC-MRI QIB \(Y_i\) as covariate, i.e.

\[
\log \left( \frac{P(Z_i=1)}{1-P(Z_i=1)} \right) = \theta_0 + \theta_1 Y_i
\]  

(13)

Area under the receiver operating characteristic curve (AUC) was used to evaluate the QIB predictive performance. Sample size calculation of AUC can be conducted using the formula provided by Obuchowski et al.\(^5\) Because the effect of measurement error on AUC is still not clear, we perform a simulation study to evaluate this effect. The true \(K_{trans}\) percent change values were repeatedly generated 1000 times, and the average AUC across these 1000 simulated data sets and the corresponding 95% CIs were calculated. Figure 4 illustrates the average AUCs against different values of \(\sigma_e\) (\(\sigma_e = 5\%, 10\%, \ldots, 30\%\)), the measurement error standard deviation in \(K_{trans}\) percent change. Our simulation results show that the predictive performance as measured by the average AUC decreases, and the length of 95% CIs increases with increased measurement error (\(\sigma_e\)).

Discussion

In this review article, we provided a general introduction on the study design, statistical model, and statistical metrics that can be used to assess repeatability and reproducibility of QIB measurements. We also illustrated the impact of repeatability- and reproducibility-related QIB measurement errors on QIB applications, e.g. on sample size calculation when a QIB is used as a clinical trial end point.

The statistical models presented here assume that the measurement errors are normally distributed and independent from the true QIB values. If the measurement errors increase in proportion to the true QIB values, i.e. multiplicative errors (see model (7)), log-transformed QIB values can be used as the relationship between error and true value becomes additive after the transformation. In practice, when QIB measurements have values equal or close to zero, we can add a small amount on all QIB values before taking the log-transformation. For more complex error structures such as non-Gaussian or heterogeneous measurement errors, the statistical methods introduced in this article can provide a reasonable approximation of the repeatability and reproducibility metrics of interest, e.g. wSD and ICC, but statistical inferences on these estimates can be biased and may lead to false conclusions.

Test–retest studies are commonly used to study repeatability and reproducibility of a QIB, where each object, e.g. a phantom or a
human subject, is being repeatedly measured. This approach may sometimes be impractical for human subject studies for reasons such as costs, time, and the invasiveness of the imaging scan. As an alternative strategy, Obuchowski et al.\textsuperscript{13} proposed a method to estimate the measurement error under repeatability condition when a test–retest study is not feasible. The method requires a reference (golden standard) value is available for each subject. By assuming the reference value to be the true QIB value ($X_i$), it can serve as the second measured value for wSD or wCV estimation:

$$mSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i)^2}$$

$$mCV = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - X_i)^2}{X_i^2}}$$

Repeatability and reproducibility can be part of the same study. It is possible to study repeatability in a restricted subset of a reproducibility study to ensure repeatability is acceptable, e.g. in an initial subset of subjects going into the study to ensure it is worth pursuing.

There is increasing need to accelerate clinical translation of QIBs. However, significant challenges remain. Using solid tumor therapy response as an example, the 1D imaging tumor size measurement based on the RECIST (Response Evaluation Criteria In Solid Tumors) 1.1 guidelines\textsuperscript{26} is the only widely used QIB in today's standard of care and clinical trials. Many QIBs that interrogate tumor biology and physiology and thus are well suited for evaluation of response to increasingly used and effective molecular targeted therapies find it difficult to be translated into clinical trials and practice. This is mainly due to the variabilities in quantifying these QIB parameter values caused by differences in vendor imaging platforms, imaging data acquisition methods, and imaging data analysis algorithms and software tools. Because of the lack of sufficient repeatability and reproducibility studies to understand the variabilities of these functional QIBs, unlike RECIST tumor size measurement, there is currently no consensus on the magnitudes of changes in these QIBs for defining clinical response endpoints such as complete response, stable disease, etc. In order to establish a path to clinical translation for functional QIBs, there is a clear need for not only standardization of data acquisition and analysis to minimize variability\textsuperscript{1}, but also more efforts in assessment of QIB repeatability and reproducibility.\textsuperscript{1} It is our hope that the statistical tools presented in this article may contribute to this endeavor.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their insightful comments.

COMPETING INTERESTS

None.

FUNDING

R01 CA248192, National Institutes of Health, U.S.A.

REFERENCES

1. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mourtz JM, et al. Quantitative imaging in cancer clinical trials.\textit{ Clin Cancer Res} 2016; 22: 284–90.\ DOI: \texttt{10.1158/1078-0432.CCR-14-3336}

2. Pinker K, Chin J, Melsaether AN, Morris EA, Moy L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment.\textit{ Radiology} 2018; 287: 732–47.\ DOI: \texttt{https://doi.org/10.1148/radiol.2018172171}

3. Kessler LG, Barnhart HX, Buckler AJ, Choudhury KR, Kondratovich MV, Toledano EA, Moy L. Precision medicine and radiogenomics: guidelines and statistical methodology.\textit{ Stat Methods Med Res} 2015; 24: 68–106.\ DOI: \texttt{https://doi.org/10.1177/0962280214537390}

4. Obuchowski NA, Mozley PD, Matthews D, Buckler A, Bullen J, Jackson E. Statistical considerations for planning clinical trials with quantitative imaging biomarkers.\textit{ J Natl Cancer Inst} 2019; 111: 19–26.\ DOI: \texttt{https://doi.org/10.1093/jnci/djy194}

5. Obuchowski NA, Reeves AP, Huang EP, Wang X-F, Buckler AJ, Kim HG, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons.\textit{ Stat Methods Med Res} 2015; 24: 68–106.\ DOI: \texttt{https://doi.org/10.1177/0962280214537390}

6. Obuchowski NA, Reeves AP, Huang EP, Wang X-F, Buckler AJ, Kim HG, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons.\textit{ Stat Methods Med Res} 2015; 24: 68–106.\ DOI: \texttt{https://doi.org/10.1177/0962280214537390}

7. Sullivan DC, Obuchowski NA, Kessler LG, Raung DL, Gatsonis C, Huang EP, et al. Metrology standards for quantitative imaging biomarkers.\textit{ Radiology} 2015; 277: 825: 813–25.\ DOI: \texttt{https://doi.org/10.1148/radiol.2015142202}

8. Huang EP, Wang X-F, Choudhury KR, McShane LM, Grönem Y, et al. Meta-analysis of the technical performance of an imaging procedure: guidelines and statistical methodology.\textit{ Stat Methods Med Res} 2015; 24: 141–74.\ DOI: \texttt{https://doi.org/10.1177/0962280214537394}

9. Yokoo T, Serai SD, Pirasteh A, Bashir MR, Hamilton G, Hernandez D, et al. Linearity, bias, and precision of hepatic proton density fat fraction measurements by MR imaging: a meta-analysis.\textit{ Radiology} 2018; 286: 486–98.\ DOI: \texttt{https://doi.org/10.1148/radiol.201710550}

10. Lodge MA. Repeatability of SUV in oncologic 18F-FDG PET. Journal of Nuclear Medicine. 2017;58(4):523–532.

11. Baumgartner R, Joshi A, Feng D, Zanderigo F, Ogden RT. Statistical evaluation of test-retest studies in PET brain imaging. \textit{EJNMMI Res} 2018; 8: 13: 1–9.\ DOI: \texttt{https://doi.org/10.1186/s13550-018-0366-8}

12. Shukla-Dave A, Obuchowski NA, Cheneveruti TL, Jambawalikar S, Schwartz LH, Malyarenko D, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DW1 and DCE-MRI derived biomarkers in multicenter oncology trials. \textit{J Magn Reson Imaging} 2019; 49: e1021–21.\ DOI: \texttt{https://doi.org/10.1002/jmri.26538}

13. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in...
radiomics modeling: possible strategies in radiologic and statistical perspectives. Korean J Radiol 2019; 20: 1124. https://doi.org/10. 3348/kjr.2018.0070

14. Fedorov A, Vangel MG, Temppany CM, Fennessy FM. Multiparametric magnetic resonance imaging of the prostate: repeatability of volume and apparent diffusion coefficient quantification. Invest Radiol 2017; 52(9): 538.

15. Schwier M, Griebthuesen J, Vangel MG, Peiper S, Peled S, Temppany C, et al. Repeatability of multiparametric prostate MRI radiomics features. Sci Rep 2019; 9: 1–16. https://doi.org/10.1038/s41598-019- 45766-z

16. Hernando D, Sharma SD, Alisay Ghasabeh M, Alvis BD, Arora SS, Hamilton G, et al. Multisite, multivendor validation of the accuracy and reproducibility of proton-density fat-fraction quantification at 1.5T and 3T using a fat-water phantom. Magn Reson Med 2019; 77: 1516–24. https://doi.org/10. 1002/mrm.26228

17. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Nakazawa M, Suzuki Y, Kasahara A, et al. Linearity, bias, interscanner repeatability, and interscanner reproducibility of quantitative multidiagonal multicohe sequence for rapid simultaneous relaxometry at 3T: a validation study with a standardized phantom and healthy controls. Invest Radiol 2019; 54: 39–47. https://doi.org/10.1097/RLI. 0000000000000510

18. Jafar MM, Parsai A, Miquel ME. Diffusion-weighted magnetic resonance imaging in cancer: reported apparent diffusion coefficients, in-vitro and in-vivo reproducibility. World J Radiol 2016; 8: 21–49. https://doi.org/10.4329/wjr.v8i11.21

19. Lin C, Bradshaw T, Perk T, Harmon S, Eickhoff J, Jallow N, et al. Repeatability of quantitative 18F-naf PET: a multicenter study. J Nucl Med 2016; 57: 1872–79. https://doi.org/10.2967/jnumed.116.177295

20. Winfield JM, Tunario N, Rata M, Miyazaki K, Jerome NP, Germuska M, et al. Extracranial soft-tissue tumors: repeatability of apparent diffusion coefficient estimates from diffusion-weighted MR imaging. Radiology 2017; 284: 88–99. https://doi.org/10. 1148/radiol.2017161965

21. Weller A, Papoutsaki MV, Waterton JC, Chiti A, Stroobants S, Kuijser J, et al. Diffusion-weighted (DW) MRI in lung cancers: ADC test-retest repeatability. Eur Radiol 2017; 27: 4552–62. https://doi.org/10.1007/s00330- 017-4828-6

22. Weller A, Savatovsky J, Balvay D, Zmuda M, Sadik I-C, Galatoire O, et al. Repeatability of apparent diffusion coefficient and intravoxel incoherent motion parameters at 3.0 tesla in orbital lesions. Eur Radiol 2017; 27: 5094–5103. https://doi.org/10.1007/s00330- 017-4933-6

23. Lu Y, Hatzoglou V, Banerjee S, Stambuk HE, Gonen M, Shankaranarayanan A, et al. Repeatability investigation of reduced field-of-view diffusion-weighted magnetic resonance imaging on thyroid glands. J Comput Assist Tomogr 2015; 39: 334–39. https://doi.org/10.1097/RCT. 0000000000000227

24. Hagiwara A, Hori M, Cohen-Adad J, Nakazawa M, Suzuki Y, Kasahara A, et al. Linearity, bias, intrascanner repeatability, and interscanner reproducibility of quantitative multidiagonal multicohe sequence for rapid simultaneous relaxometry at 3T: a validation study with a standardized phantom and healthy controls. Invest Radiol 2019; 54: 39–47. https://doi.org/10.1097/RLI. 0000000000000510

25. Jafari-Khouzani K, EKEMEIE kalpathy-Cramer J, Bjernerud A, Vangel MG, Gerstner ER, et al. Repeatability of cerebral perfusion using dynamic susceptibility contrast MRI in glioblastoma patients. Transl Oncol 2015; 8: S1936-5233(15)00018-2: 137–46. . https://doi. org/10.1016/j.tranon.2015.03.002

26. Han A, Andre MP, Deiranieh L, Housman E, Erdman JW Jr, Loomba R, et al. Repeatability and reproducibility of the ultrasound attenuation coefficient and backscatter coefficient measured in the right lobe of the liver in adults with known or suspected nonalcoholic fatty liver disease. J Ultrasound Med 2018; 37: 1913–27. https://doi.org/10.1002/jum.14537

27. Hagiwara A, Fujita S, Ohno Y, Aoki S. Variability and standardization of quantitative imaging: monoparametric to multiparametric quantification, radiomics, and artificial intelligence. Invest Radiol 2020; 55: 601–16. https://doi.org/10.1097/RLI. 0000000000000666

28. Wang K, Manning P, Szeverenyi N, Wolfson T, Hamilton G, Middleton MS, et al. Repeatability and reproducibility of 2D and 3D hepatic MR elastography with rigid and flexible drivers at end-expiration and end-inspiration in healthy volunteers. Abdom Radiol (NY) 2017; 42: 2843–54. https://doi. org/10.1007/s00261-017-1206-4

29. A ‘Hern RP. Employing multiple synchronous and flexible drivers at end-expiration and end-inspiration. Abdom Radiol (NY) 2017; 42: 2843–54. https://doi.org/10.1007/s00261-017-1206-4

30. Olin A, Ladedof CG, Langer NH, Keller SH, Löfgrén J, Hansen AE, et al. Reproducibility of MB-based attenuation maps in PET/MRI and the impact on PET quantification in lung cancer. J Nucl Med 2018; 59: 999–1004. https://doi.org/10.2967/jnumed.117.198853

31. Fuller WA. Measurement error models. John Wiley & Sons; 2009.

32. Obuchowski NA, Bullen J. Quantitative imaging biomarkers: effect of sample size and bias on confidence interval coverage. Stat Methods Med Res 2018; 27: 3139–50. https://doi.org/10.1177/0962280217693662

33. Gulecchi A, Burzykowski T. Linear Mixed-Effects Models Using R. New York, NY: Springer; 2013., pp. 245–73. https://doi.org/10.1007/978-1-4614-3900-4

34. Girden ER. ANOVA. 2455 Teller Road, Thousand Oaks California 91320 United States of America : Sage; 1992. https://doi.org/10.4135/9781482983419

35. Gillip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? crucial concepts for primary care researchers. Ann Fam Med 2004; 2: 204–8. https://doi.org/10.1377/afm.141

36. A’Hern RP. Employing multiple synchronous outcome samples per subject to improve study efficiency. BMC Med Res Methodol 2021; 21: 1: 1–11. . https://doi.org/10.1186/s12874-021-01414-7

37. Rinne H. The Weibull Distribution. CRC press; 2008. https://doi.org/10.1201/ 9781420087444

38. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC press; 1994. https://doi.org/10.1201/9780429246593

39. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. Ann J Epidemiol 2014; 179: 252–60. https://doi.org/10.1093/aje/ kwt245

40. Cook JA, Bruckner T, MacLennan GS, Seiler CM. Clustering in surgical trials--database of intracluster correlations. Trials 2012; 13: 2: 1–8. . https://doi.org/10.1186/1745-6215-13- 2

41. Thompson DM, Fernald DH, Mold JW. Confidence intervals for intraclass correlation coefficients in variance and flexible drivers at end-expiration and end-inspiration in healthy volunteers. Abdom Radiol (NY) 2017; 42: 2843–54. https://doi.org/10.1007/s00261-017-1206-4

42. Demetsravili N, Wit E, Heavel ER. Confidence intervals for intraclass correlation coefficients in variance components models. Stat Methods Med Res 2016; 25: 2359–76. https://doi.org/10.1177/ 0962280214522787

43. Ionan AC, Polly MC, McMahan LS, Dobbin KK. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). BMC Med Res Methodol 2014; 14: 1–11. https://doi.org/10. 1186/1471-2288-14-121

44. Weerahandi S. Generalized confidence intervals. Journal of the American Statistical
45. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1993; 45: 255. https://doi.org/10.2307/2532051

46. Huang W, Li X, Chen Y, Li X, Chang M-C, Oborski MJ, et al. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. *Transl Oncol* 2014; 7: 166: 153–66: . https://doi.org/10.1593/tlo.13838

47. Chen CC, Barnhart HX. Comparison of ICC and CCC for assessing agreement for data without and with replications. *Computational Statistics & Data Analysis* 2008; 53: 554–64. https://doi.org/10.1016/j.csda.2008.09.026

48. Lin HM, Williamson JM. A simple approach for sample size calculation for comparing two concordance correlation coefficients estimated on the same subjects. *J Biopharm Stat* 2015; 25: 1145–60. https://doi.org/10.1080/10543406.2014.971163

49. Allman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 1983; 32: 307. https://doi.org/10.2307/2987937

50. Millen GC, Yap C. Adaptive trial designs: what are multiarm, multistage trials. In: *Archives of Disease in Childhood-Education and Practice*. ; 2020, pp. 376–78.

51. Yang M, Adomavicius G, Burtch G, Ren Y. Mind the gap: accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research* 2018; 29: 4–24. https://doi.org/10.1287/isre.2017.0727

52. Chesher A. The effect of measurement error. *Biometrika* 2018; 78: 451–62. https://doi.org/10.1093/biomet/78.3.451

53. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models.* Boston, MA: CRC press; March 2018. https://doi.org/10.1007/978-1-4899-4477-1

54. Tudorica A, Oh KY, Chui SY-C, Roy N, Troxell ML, Naik A, et al. Early prediction and evaluation of breast cancer response to neoadjuvant chemotherapy using quantitative DCE-MRI. *Transl Oncol* 2016; 9: S1936-5233(15)30017-6: 8–17:. https://doi.org/10.1016/j.tranon.2015.11.016

55. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004; 50: 1118–25. https://doi.org/10.1373/clinchem.2004.031823

56. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; 228–47.