

Applications of the Repeatability of Quantitative Imaging Biomarkers: A Review of Statistical Analysis of Repeat Data Sets

Huiman X. Barnhart* and Daniel P. Barboriak†

*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA;

†Department of Radiology, Duke University Medical Center, Durham, NC 27710, USA

Abstract

Repeat imaging data sets performed on patients with cancer are becoming publicly available. The potential utility of these data sets for addressing important questions in imaging biomarker development is vast. In particular, these data sets may be useful to help characterize the variability of quantitative parameters derived from imaging. This article reviews statistical analysis that may be performed to use results of repeat imaging to 1) calculate the level of change in parameter value that may be seen in individual patients to confidently characterize that patient as showing true parameter change, 2) calculate the level of change in parameters value that may be seen in individual patients to confidently categorize that patient as showing true lack of parameter change, 3) determine if different imaging devices are interchangeable from the standpoint of repeatability, and 4) estimate the numbers of patients needed to precisely calculate repeatability. In addition, we recommend a set of statistical parameters that should be reported when the repeatability of continuous parameters is studied.

Translational Oncology (2009) 2, 231–235

Introduction

Validation of quantitative imaging biomarkers for specific clinical applications has been a focus of many recent research efforts. The success of these endeavors may potentially be of significant benefit to a large group of patients with a variety of diseases [1]. In the field of oncology, imaging biomarkers may be used to increase our understanding of tumor and patient heterogeneity, to document drug delivery to target tissues, to elucidate drug mechanisms of actions, to monitor treatment side effects, and, perhaps most importantly, to evaluate treatment response [2,3].

The potential utility of an imaging biomarker to measure a biologic process can be severely hampered by a lack of repeatability or reproducibility. (For the purpose of this article, we refer to *repeatability* as consistency of quantitative results obtained when the same imaging test is performed at short intervals on the same subjects or test objects using the same equipment in the same center, as distinguished from *reproducibility*, the consistency of results obtained when the same imaging test is performed at short intervals on the same subjects or test objects using different equipment in different centers.) The derivation of reproducible quantitative parameters from imaging can pose enormous challenges [4–6]. The level of repeatability or reproducibility needed to successfully use a biomarker is critically dependent on the intended application; for example, the

requirements needed for an imaging test to detect a biologic effect of a treatment in a large cohort of patients are considerably less stringent than the requirements for the same test to reliably predict a successful treatment response in a single patient.

Unfortunately, although an understanding of a technique's repeatability is necessary to interpret the significance of a test result for individual patients, repeat image sets to evaluate the repeatability of quantitative imaging techniques are relatively rarely obtained [7]. Recent efforts to promote the performance of repeat imaging studies and to make the image data publicly available for use in research applications, including the Reference Image Database to Evaluate Response to Therapy (RIDER) project of the National Cancer Institute's Cancer Imaging Program [8,9] and related efforts discussed in this publication, are in the early stages of addressing this need.

Address all correspondence to: Daniel P. Barboriak, MD, Department of Radiology, Box 3808, Duke University Medical Center, Durham, NC 27710. E-mail: barbo013@mc.duke.edu

Received 8 September 2009; Revised 22 September 2009; Accepted 23 September 2009

Copyright © 2009 Neoplasia Press, Inc. All rights reserved 1944-7124/09/\$25.00
DOI 10.1593/tdo.09268

Appropriate Statistical Methods

The increasing availability of public databases of repeat examinations raises the question of what are appropriate statistical methods for analyzing quantitative parameters obtained from repeat imaging data sets. Unfortunately, inappropriate statistical methods, such as the use of correlation coefficients to measure the agreement between two repeated measurements, have been common in the published literature [10]. Further complicating matters, the types of statistical analyses that are most appropriate depend to a great extent on the nature of the research question being addressed.

Bland and Altman [11] provided an intuitive methodology using the concept of limits of agreement for assessing agreement between two methods of clinical measurement. Barnhart et al. [12] provided an overview on the evolution of various statistical methodologies on assessing agreement over the past several decades. In this article, we focus on the issue of repeatability only and address appropriate statistical analysis that can use information from repeat imaging data sets in four different scenarios. In addition, we will suggest statistical parameters that should be reported in studies of repeatability and discuss other potential applications for statistical analysis of repeat imaging data sets.

For the scenarios described below, we assume that the presence of *change or lack of change* in a quantitative imaging biomarker is of interest to the researcher. In this situation, the presence of variability in the extracted parameter, whether due to patient-related or imaging system-related factors, can make a measured change in the parameter difficult to interpret. Statistical tests derived from the analysis of repeat data sets can be used, for example, to help determine whether this variability can account for an observed level of change in parameter. In addition, we assume that the parameter extracted from imaging is continuous and normally distributed, that the repeatability is similar across subjects and does not depend on the subject's true parameter value or other characteristics of the subject, and that there is no systematic change in the subject or test object imaged during the interval between repeated studies.

In the first two scenarios, we try to determine if there is true change in an individual patient or not. Owing to random error, individual patients may fall into one of three categories: 1) definite change is present with 95% confidence, 2) definite no change is present with 95% confidence, and 3) unable to determine if there is a change or not. Scenario 1 deals with the first category and scenario 2 deals with the second. There may be a large group of patients in the third category; it may be possible to reduce the number of patients in this category by estimating the repeatability with higher precision. To accomplish this, we can design a larger study of repeat measurements with a prespecified level of precision on the estimate of repeatability, as described in scenario 4. Scenario 3 deals with interchangeability of devices when we only know the repeatability of particular parameters in the repeat data sets, where different imaging devices were used in different sets of patients.

The following notations are used for the repeated data sets. Let Y_{ik} be the observed value for the i th subject at the k th replication, $i = 1, \dots, n$; $k = 1, \dots, K$. The one-way analysis of variance (ANOVA) model $Y_{ik} = \mu_i + \varepsilon_{ik}$ can be used to express the observed value as the true value plus the within-subject error with between-subject variance $\sigma_B^2 = \text{Var}(\mu_i)$ and within-subject variance $\sigma_W^2 = \text{Var}(\varepsilon_{ik})$. Hence, the total variance is $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. Let BMS and WMS be the between-subject means of squares and within-subject means of squares obtained from the one-way ANOVA model or calculated by

$BMS = K \sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2 / n$ and $WMS = \sum_{i=1}^n \sum_{k=1}^K (Y_{ik} - \bar{Y}_i)^2 / n(K-1)$, where \bar{Y}_i is the average over replication for subject i and \bar{Y} is the grand mean over all observations. We can also calculate $tSD = \sqrt{(BMS + (K-1)WMS) / K}$, $bSD = \sqrt{(BMS - WMS) / K}$, and $wSD = \sqrt{WMS}$, which are the unbiased estimates of σ_T , σ_B , and σ_W , respectively. The tSD is useful for the design of future studies if Y is chosen as the primary outcome because tSD can be used as an estimate of the standard error of the outcome. The repeatability coefficient (RC) is defined as $RC = 1.96\sqrt{2\sigma_W^2} = 2.77\sigma_W$ and can be estimated by $\hat{RC} = 2.77wSD$. Note that the interpretation of RC is that the difference between any two readings on the same subject is expected to be from $-RC$ to RC for 95% of subjects. If there are only two repetitions ($K = 2$) per subject, wSD is the SD of the differences divided by $\sqrt{2}$. Because the WMS is distributed as $\chi_{n(K-1)}^2 \sigma_W^2 / n(K-1)$, the 95% confidence interval (CI) for σ_W is $(\hat{\sigma}_L, \hat{\sigma}_U)$ where $\hat{\sigma}_L = \sqrt{n(K-1)WMS / \chi_{n(K-1)}^2(0.975)}$, $\hat{\sigma}_U = \sqrt{n(K-1)WMS / \chi_{n(K-1)}^2(0.025)}$, and $\chi_{n(K-1)}^2(a)$ is the a th percentile of the χ^2 distribution with $n(K-1)$ degrees of freedom. The corresponding 95% CI for RC is $(RC_L, RC_U) = (2.77\hat{\sigma}_L, 2.77\hat{\sigma}_U)$. A relative measure of repeatability is the within-subject coefficient of variation (wCV) defined as $wCV = \sigma_W / \mu$ where $\mu = E(\mu_i)$ is the mean of true value. This measure is sometimes called error rate.

Scenario 1: What Level of Change in Parameter Should Be Observed to Be Confident That There Has Been a True Change in the Parameter in an Individual Patient?

The presence of change in a biomarker after therapy may be useful to identify a group of patients in whom therapy has resulted in a specific biologic effect. Straightforward examples of this in cancer imaging would include studies of change in volume of a lung nodule by volumetric computed tomography or change in standardized uptake value in ^{18}F -fluorodeoxyglucose positron emission tomography after therapy relative to pretreatment baseline values. If such changes can be validated as surrogate markers of treatment response, the potential impact on patient care is obvious. Alternatively, association of change in biomarker with particular patient or disease characteristics can be helpful to elucidate treatment mechanisms of action; for example, it may be of interest to identify whether only particular subgroups of tumors or patients demonstrate a particular biologic response.

To assess whether there is an overall change in the group after treatment, paired analysis (e.g., paired t -test if differences are normally distributed or nonparametric Wilcoxon test if not) is appropriate. But to determine whether treatment has resulted in a change in parameter for an individual patient, use of 95% CI for estimated RC [11] can provide a reasonable approach for determining whether the observed change is a true change. Specifically, let $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$ be the observed measurements before ($j = \text{pre}$) and after ($j = \text{post}$) the therapy for the i th subject, where μ_{ij} and ε_{ij} are the true value and measurement error for subject i at time j , respectively. A true individual change corresponds to $\mu_{i\text{post}} - \mu_{i\text{pre}} \neq 0$.

Because there may be small levels of change that are considered inconsequential in the context of an individual trial, it is helpful to consider a more general situation in which $(-\delta_c, \delta_c)$ is the interval within which the magnitude of parameter change is considered essentially unchanged; for example, a level of change so small that no therapeutic benefit is expected. In this case, a true individual change corresponds to $|\mu_{i\text{post}} - \mu_{i\text{pre}}| \geq \delta_c$. Assume that conditional on subject i , the measurement errors $\varepsilon_{i\text{pre}}, \varepsilon_{i\text{post}}$ are independent and have the

same error variance of σ_W^2 for all subjects. This assumption implies that the random repetition error does not depend on the subject's true parameter value or on subject characteristics. Then $Y_{ipost} - Y_{ipre}$ has mean $\mu_{ipost} - \mu_{ipre}$ and variance of $2\sigma_W^2$. Thus, the 95% CI for $\mu_{ipost} - \mu_{ipre}$ is $(Y_{ipost} - Y_{ipre} - 1.96 \times \sqrt{2\sigma_W^2}, Y_{ipost} - Y_{ipre} + 1.96 \times \sqrt{2\sigma_W^2})$ or $(Y_{ipost} - Y_{ipre} - RC, Y_{ipost} - Y_{ipre} + RC)$. If this CI is outside $(-\delta_c, \delta_c)$, then we are 95% confident that a change has occurred for this individual, assuming we know the true value of RC. Thus, the RC provides a reasonable approach for determining whether the observed change is a true change or not.

One problem in implementing this approach in practice is that the RC is not known with certainty. Moreover, an RC estimated from a large number of repeat data sets is generally more reliable than a coefficient estimated from a smaller number. By calculating 95% CI for the RC, (RC_L, RC_U) , we can take into account the effect of the number of repeat data sets on the degree of uncertainty around the estimate of the RC and generate a conservative estimate of the level of parameter change that would allow individual patients with true change to be identified. Using this approach, we can only be 95% confident that a change has occurred for an individual patient if the interval $(Y_{ipost} - Y_{ipre} - RC_L, Y_{ipost} - Y_{ipre} + RC_U)$ lies outside $(-\delta_c, \delta_c)$.

If only a single direction of change is of interest, then a one-sided confidence limit may be applied. For example, to determine if the measurement has increased or not, we would like to determine if $\mu_{ipost} - \mu_{ipre} > \delta_c$. The lower limit for the one-sided 95% CI is $Y_{ipost} - Y_{ipre} - 1.645\sqrt{2\sigma_L}$. If this limit is greater than δ_c , then we are 95% confident that a positive change has occurred.

For the specific instance in which we are studying whether a change of any magnitude occurred in a patient, no matter how small (i.e., when $\delta_c = 0$), when both directions of change are of interest, we can only be 95% confident that a change has occurred for an individual patient if the interval $(Y_{ipost} - Y_{ipre} - RC_L, Y_{ipost} - Y_{ipre} + RC_U)$ does not contain zero. When only a single direction of change is of interest, a one-sided CI is used; for example, if only an increase in parameter is of interest, we can only be 95% confident that a change has occurred when $Y_{ipost} - Y_{ipre} - 1.645\sqrt{2\sigma_L}$ is greater than zero.

Scenario 2: What Level of Change in Parameter Should Be Observed to Be Confident There Has Been No Change in the Parameter in an Individual Patient?

Although the utility of identifying patients who demonstrate an actual change in a parameter may be obvious, the ability to reliably identify individual patients who do not exhibit change after treatment may also be of value. Given the expense and side effects of a particular therapy, one application of quantitative imaging may be to definitively identify patients who did not experience, for example, a beneficial pharmacodynamic change from therapy. This result of the imaging might be used to justify discontinuation of therapy, particularly if further studies show that lack of pharmacodynamic change corresponds reliably to poor clinical outcomes.

In considering this scenario, two situations should be considered. In the first situation, we would like to identify patients who show no change in a parameter and distinguish this patient group from those who either had an increase or decrease in the parameter. In this situation, it is assumed that it remains unclear whether a change in parameter in either direction may represent a therapeutic effect. In the second situation, we would like to identify patients who show either

no change in a parameter or change in one direction in the parameter and distinguish this patient group from those who have the opposite direction of change. In this situation, a direction of change of parameter corresponding to a therapeutic effect is presumed to be known.

Again, consider $(-\delta_c, \delta_c)$ to be the interval within which magnitude of parameter change is so small that no therapeutic benefit is expected. Then the patient is considered to have no change if the interval $(Y_{ipost} - Y_{ipre} - RC_L, Y_{ipost} - Y_{ipre} + RC_U)$ is contained inside of $(-\delta_c, \delta_c)$. However, if the direction of parameter change corresponding to therapeutic benefit is known and if (for example) increase in parameter values indicates benefit, we can similarly use the RC to identify individuals who had either no change or decrease in parameter value and thus warrant discontinuation of therapy. In this case, we are 95% confident that subjects with $Y_{ipost} - Y_{ipre} + 1.645\sqrt{2\sigma_U} < \delta_c$ had either no change or decrease in parameter value.

Scenario 3: Assuming Two Imaging Devices Are Measuring the Same Parameter, Are Two Devices Interchangeable as Far as Repeatability Is Concerned?

The question of whether different imaging devices are interchangeable is of considerable practical importance in cancer imaging, particularly when devising clinical trials. To reduce variability in clinical trials in which repeated quantitative imaging measurements are made, imaging studies are often required to be performed on the same piece of equipment. This requirement may result in difficulty with patient scheduling; moreover, imaging equipment hardware or software upgrades occurring while the clinical trial is taking place may affect the accuracy or reproducibility of the quantitative imaging parameter extraction.

The question of whether the quantitative results of a specific image acquisition and analysis protocol are interchangeable with another is also of interest outside the clinical trials context. The possibility that these results can be influenced by the characteristics of imaging devices (e.g., field strength of MR magnets used to study proton spectroscopy [13]), acquisition protocols (e.g., scan timing and duration used in oncologic ^{18}F -fluorodeoxyglucose positron emission tomography studies [14]), and image analysis software (e.g., software used for measuring tumor perfusion from multidetector computed tomography data [15]) is well known. Different modalities may be used to extract similar physiological parameters such as tumor blood volume or flow [16]; if parameter accuracy is validated for a modality that is not widely available, expensive, or exposes patients to risks, it may be possible to show that parameters extracted by a more widely available, inexpensive, and/or safer modality give equivalent clinical information.

To compare the *accuracy* of two imaging devices, an independent ground truth is necessary, which repeat imaging sets do not provide. Repeat image sets do, however, allow the repeatability of two techniques to be compared, assuming that the imaged patient groups are similar. If the same group of patients were measured by two different imaging devices in the repeat data sets, both intradevice repeatability and interdevice agreement (variability) may be assessed. In this case, we can determine whether the two devices can be used interchangeably or not by assessing interdevice agreement with statistical approaches for assessing agreement [11,12].

Often, however, different image patient groups were used by different devices in the repeated data sets. In this case, if there are no systematic differences between the quantitative parameters extracted related to the device used to perform the imaging, the technique with

smaller repeatability may be preferred or the devices may be used interchangeably if the repeatability is similar.

Considering this situation, let $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, $i = 1, \dots, n_j$ be the observed measurements for the i th subject by the j th device at the k th replication, where μ_{ij} , ε_{ijk} are the “true” value and measurement errors by device j for subject i . If there is no systematic bias between the devices, then $\mu_{ij} = \mu_i$ for all j , the “true” parameter value produced by the devices. Let $\sigma_{w_j}^2$ be the variance of measurement error by device j . To determine if the repeatability in one device is different from the repeatability in the other device, one can test hypothesis $H_0: \sigma_{w_1}^2 = \sigma_{w_2}^2$ versus $H_1: \sigma_{w_1}^2 \neq \sigma_{w_2}^2$. If different sets of patients were used in the measurements by different devices, test of this hypothesis test can be accomplished by an F -test or Levene test for equality of variance. This can be accomplished by using the option “HOVTEST” in the mean statement in SAS procedure GLM with one-way ANOVA model for outcome variable of $Z_{ijk} = Y_{ijk} - \bar{Y}_{ij}$, where \bar{Y}_{ij} is the average reading for the i th subject by the j th device over the replications. Another simple way to test this hypothesis is to use the two group t -test for $|Z_{ijk}|$ because the mean of this absolute value is proportional to σ_{w_j} when the data is normally distributed. The paired t -test may be used if the same sets of patients were used for measurements by different devices. A note of caution is that the test for equality of variance may have low power and the failure to reject the null hypothesis only means that there is no sufficient evidence to conclude that the repeatability is unequal.

To establish whether the repeatability is equivalent or not, we would need to carry out an equivalence test. In this case, we need to choose an equivalence margin δ_v (>1), which is the acceptable ratio in RCs for the two devices to be considered interchangeable. The hypothesis we want to test then is in the form of $H_0: \sigma_{w_1}/\sigma_{w_2} \geq \delta_v$ or $\sigma_{w_1}/\sigma_{w_2} \leq 1/\delta_v$ versus $H_1: 1/\delta_v < \sigma_{w_1}/\sigma_{w_2} < \delta_v$. This hypothesis can be decomposed into two one-sided hypotheses that can be tested by two F -tests [17].

Scenario 4: How Can Analysis of Small Repeat Data Sets Be Used to Estimate How Many Patients Would Be Needed to Power a Larger Trial of Repeat Data Sets to Determine the Repeatability of a Quantitative Image Parameter to a Particular Degree of Accuracy?

In scenarios 1 and 2, we provided a statistical rationale for classifying individual patients as “changed” or “unchanged” based on the observed change in quantitative imaging parameter. Because the RC is never known with absolute certainty, there is always a range of observed changes from which the change classification is undetermined. Narrowing the CIs for the RC as well as the magnitude of the RC is needed to have a smaller group of patients fall into this undetermined range; this can be partially achieved if the sample size in repeat data sets sufficiently large that the RC can be estimated precisely. Small repeat data sets can be used to help design these larger studies to have the desired precision.

Ideally, we would like to have $RC_U \leq \delta_c$, where δ_c is the magnitude of parameter’s positive change that is of interest for measurement in the trial; for example, δ_c could be the minimum amount of change in parameter value considered to indicate a therapeutic benefit to a patient. When designing a study of repeat data sets to estimate RC, a sufficient number of subjects and replications should be performed so that if the study finds that $RC_U > \delta_c$, this result is unlikely to be due to chance alone. The equation $\chi_{\alpha/(K-1)}^2 / (n(K-1)) \geq (RC/\delta_c)^2$ can be used to determine the number of trial subjects (n) and replications per subject (K) required to calculate CIs at the level of $1 - \alpha$. For

example, if $K = 2$ and the RC is 80% of δ_c , we will need a sample size of 48 to be 95% confident ($\alpha/2 = 0.025$) that $RC_U \leq \delta_c$. However, if RC is 90% of δ_c , we will need a sample size of 192 subjects.

Small repeat data sets can be useful for planning larger trials of repeat imaging by providing initial estimates of RC for the equation above. Often, the parameter δ_c will also not be known with certainty, in which case an estimate obtained from other studies (e.g., from a prior treatment imaging trial or by extrapolation from animal studies) may be used.

Discussion

In the early stages of biomarker development, the technical performance of an imaging test is often evaluated by measurements of sensitivity and specificity or receiver operating characteristic analysis [18]. Single-center clinical trials to demonstrate proof of concept for clinical use often follow. Measurements of imaging biomarker reproducibility, although often mentioned as an important step in imaging biomarker development [5], are relatively less commonly obtained [7], perhaps for reasons of expense, patient tolerance, and/or increased risk to patients. As the analyses above indicate, repeatability of quantitative image parameter extraction is a major factor in determining whether a change in parameter value can be detected in a practical way for individual patients in a clinical care setting.

The analyses above can also be extended to consider the planning of a multicenter clinical trial, assuming that there are several candidate imaging protocol/parameter sets that could be implemented and that the degree of parameter change corresponding to a positive therapeutic result is reasonably well understood. If the RC for a particular parameter is large relative to the size of change (Δ) in parameter believed to correspond with a therapeutic effect, it is unlikely that measurements of change in this parameter can be measured with sufficient reliability to be used in clinical decision making; parameters that are highly repeatable and that show large changes after successful treatment (i.e., $RC/\Delta < 1$) have much more promise in this regard.

Given the high cost of implementation of multicenter imaging trials, comparison of the size of treatment effect to the RC could serve as a means to “triage” candidate imaging protocol/parameter sets; those with high treatment-related parameter change size relative to RC would be considered promising for inclusion in multicenter trials, particularly when identification of imaging techniques that can be translated into standard clinical practice is a study goal. In contrast, those with high RC relative to treatment effect size would require more development work aimed at increasing reproducibility before being considered in a multicenter trial.

Because of the relatively rarity of repeatability measurements for quantitative parameters in the radiology literature and the possible utility of these measurements for several purposes, we suggest that, at minimum, studies of repeatability report the following values for each parameter:

- Number of subjects (n) with nonmissing value
- Number of replications (K)
- Mean, total SD ($tSD = \sqrt{(\text{BMS} + (K-1)\text{WMS})/K}$), and range of observed parameter values
- Mean, within-subject SD (wSD , or SD of differences divided by $\sqrt{2}$ if $K = 2$), and range of differences in parameter value
- Repeatability coefficient estimated by $2.77wSD$ or $2.77\sqrt{\text{WMS}}$ with corresponding 95% CI (RC_L , RC_U)

One may also report the within-subject coefficient of variation (wCV), variance ratio ($VR = \text{between-subject variance}/\text{within-subject}$

variance), and the intraclass correlation coefficient (ICC), where the wCV can be estimated by $wCV = wSD/Y$, VR can be estimated by $VR = (tSD^2 - wSD^2) / wSD^2$, and the ICC can be estimated by $ICC = (tSD^2 - wSD^2) / tSD^2$ based on the above minimum statistics.

Although we illustrated that RC is useful in determining whether there is a change for an individual patient, we caution that this approach depends on a key assumption that RC is the same for all patients. If the RC depends on the magnitude of the parameter for that patient or if the RC varies for different subgroups of patients with particular characteristics, then RC as a function of the parameter or multiple RCs for different subgroups should be estimated to better evaluate parameter change for individual patients while taking into account the patient's specific characteristics. The dependency of RC on patient's characteristics can only be assessed with a large number of replications ($K > 2$) and/or sufficient sample sizes within the subgroups to get precise estimates of multiple RCs.

One limitation of this review is that we considered only a limited number of scenarios and assumed the quantitative imaging parameter was a single metric derived from each patient. Because images show the distribution of parameters in space, they allow hypotheses about locations and distributions of parameter value distribution to be tested (e.g., lesional analysis in which patients may have variable numbers of lesions), which are beyond the scope of this article.

In summary, the public availability of repeat imaging data sets provides new opportunities for study of variability of quantitative parameters derived from imaging. When use of these parameters in a clinical context is contemplated, issues of parameter reproducibility and repeatability become particularly important. In this article, we discuss four questions that can be addressed using data from repeat imaging data sets and appropriate statistical approaches that can be used to help answer these questions.

Because of their value as a test bed for activities such as software development and validation, the inclusion of repeat imaging data sets in publicly accessible databases should be encouraged. Recognizing that this is not always practical, we also recommend a minimal set of statistics that should be reported when particular results from these data sets are described in the literature. Inclusion of a common set of statistics may be helpful in several instances; for example, when comparing results obtained across several different laboratories; when determining whether two imaging devices are likely to be interchangeable as far as repeatability is concerned; and when planning larger trials of test-retest imaging.

References

- [1] Smith JJ, Sorensen AG, and Thrall JH (2003). Biomarkers in imaging: realizing radiology's future. *Radiology* **227**, 633–638.
- [2] O'Connor JP, Jackson A, Parker GJ, and Jayson GC (2007). DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents. *Br J Cancer* **96**, 189–195.
- [3] Zhao M, Pipe JG, Bonnett J, and Evelhoch JL (1996). Early detection of treatment response by diffusion-weighted $^1\text{H-NMR}$ spectroscopy in a murine tumour *in vivo*. *Br J Cancer* **73**, 61–64.
- [4] Schuster DP (2007). The opportunities and challenges of developing imaging biomarkers to study lung function and disease. *Am J Respir Crit Care Med* **176**, 224–230.
- [5] Sorensen AG (2006). Magnetic resonance as a cancer imaging biomarker. *J Clin Oncol* **24**, 3274–3281.
- [6] Willmann JK, van Bruggen N, Dinkelborg LM, and Gambhir SS (2008). Molecular imaging in drug development. *Nat Rev Drug Discov* **7**, 591–607.
- [7] Murphy PS, McCarthy TJ, and Dzik-Jurasz AS (2008). The role of clinical imaging in oncological drug development. *Br J Radiol* **81**, 685–692.
- [8] Armato SG III, Meyer CR, McNitt-Gray MF, McLennan G, Reeves AP, Croft BY, and Clarke LP (2008). The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. *Clin Pharmacol Ther* **84**, 448–456.
- [9] McLennan G, Clarke L, and Hohl RJ (2008). Imaging as a biomarker for therapy response: cancer as a prototype for the creation of research resources. *Clin Pharmacol Ther* **84**, 433–436.
- [10] Halligan S (2002). Reproducibility, repeatability, correlation and measurement error. *Br J Radiol* **75**, 193–194.
- [11] Bland JM and Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310.
- [12] Barnhart HX, Haber MJ, and Lin LI (2007). An overview on assessing agreement with continuous measurements. *J Biopharm Stat* **17**, 529–569.
- [13] Kim JH, Chang KH, Na DG, Song IC, Kim SJ, Kwon BJ, and Han MH (2006). Comparison of 1.5T and 3T ^1H MR spectroscopy for human brain tumors. *Korean J Radiol* **7**, 156–161.
- [14] Castell F and Cook GJ (2008). Quantitative techniques in ^{18}F FDG PET scanning in oncology. *Br J Cancer* **98**, 1597–1601.
- [15] Goh V, Halligan S, and Bartram C (2007). Quantitative tumor perfusion assessment with multidetector CT: are measurements from two commercial software packages interchangeable? *Radiology* **242**, 777–782.
- [16] Ng CS, Kodama Y, Mullani NA, Barron BJ, Wei W, Herbst RS, Abbruzzese JL, and Charnsangavej C (2009). Tumor blood flow measured by perfusion computed tomography and ^{15}O -labeled water positron emission tomography: a comparison study. *J Comput Assist Tomogr* **33**, 460–465.
- [17] Chow S-C, Shao J, and Wang H (2008). *Sample Size Calculations in Clinical Research*. Boca Raton, FL: Chapman & Hall/CRC.
- [18] Metz CE (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol* **3**, 413–422.