ORIGINAL RESEARCH

Evaluating agreement between solid tumor measurements used to assess response

Chaya S. Moskowitz *, Mithat Gönen

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, United States

Received: November 5, 2014 **Accepted:** January 8, 2015 **Online Published:** May 8, 2015
DOI: 10.5430/jbgc.v5n2p1 **URL:** http://dx.doi.org/10.5430/jbgc.v5n2p1

Abstract

Objective: To propose and demonstrate using Bland-Altman plots and Limits of Agreement based on the relative difference (RD) in solid tumor measurements to assess agreement.
Methods: A modification to the Bland-Altman plot which involves replacing the difference plotted on the vertical axis with the relative percent difference between two measurements of tumor burden is discussed. Quantifying tumor burden requires summing skewed individual tumor measurements. This quantity is the same one used in assessing tumor response to therapeutic agents and is familiar to radiologists and clinicians working with cancer patients. The distribution of the relative difference is explored and revised equations for the limits of agreement are presented. The methods are then applied to positron emission tomography data studying two radiotracers.
Results: The distribution of the relative difference is highly skewed and can be approximated by a negative shifted lognormal distribution. The limits of agreement for the RD need to appropriately reflect this distribution. The standard equations for the 95% limits of agreement assume the differences are approximately normally distributed and may not be appropriate for the relative difference.
Conclusions: The modified Bland-Altman plot is based on a clinically meaningful quantity and provides a method for handling data with multiple lesions per patient.

Key Words: Inter-reader variability, Limits of agreement, Method comparison, Reproducibility

1 Introduction

Serial measurements of solid tumors are used to gauge whether a tumor is responding to an anti-cancer therapeutic agent when managing a patient's care and testing new treatments in clinical trials. This assessment is frequently done based on change in tumor burden as seen on functional or metabolic imaging. Under guidelines established for response evaluation in solid tumors using functional imaging (RECIST), the relative percent difference (RD) between tumor size measured at a baseline, pre-treatment time and a follow-up time after treatment has commenced is calculated.[1,2] Other recommendations for evaluating response in solid tumors using metabolic imaging with 2-[18 F]Fluoro-2-deoxyglucose positron emission tomography (FDG-PET) have suggested using the RD to quantify tumor response.[3] Because patients with aggressive forms of cancer may have multiple tumors, a patient-level assessment of tumor burden requires evaluating a composite form of individual tumors. The RECIST criterion, for example, stipulates that up to five tumors be measured and the sums of their dimensions at baseline and follow-up, respectively, should be used in calculating RD. Because small changes in

*Correspondence: Chaya S. Moskowitz; Email: moskowc1@mskcc.org; Address: Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 E 63rd Street, 3rd Floor, New York, United States.

the RD have the potential to affect early-phase clinical trial outcomes and patient treatment, ensuring tumor measurements are reproducible is crucial.[4–6]

Bland-Altman plots and 95% limits of agreement (LoA) are frequently used to evaluate the reproducibility of solid tumor measurements on functional and metabolic imaging across readers or "techniques" such as tracers, reconstruction algorithms, or machinery.[6–13] The Bland-Altman plot is typically constructed by plotting the simple difference between two measurements, d, by the mean of the measurements. The 95% LoA are constructed under the assumption that the differences are approximately normally distributed. They are defined as $\bar{d} \pm 1.96s$, where $\bar{d}$ is the average difference and s is the standard deviation of d[14,15] and are usually included on the Bland-Altman plot. One concludes that the inter- or intra-observer agreement is acceptable if the differences within the LoA are not clinically important.

Although d is the most commonly used measure of distance, across applications, including those in the radiology literature, there is variation in how the Bland-Altman plots and LoA are constructed in terms of how the difference is quantified. Here we suggest basing Bland-Altman plots and the LoA on RD when evaluating agreement between tumor measurements. Because this quantity is regularly used to make clinical decisions, an analysis based on RD is easily interpreted by radiologists and clinicians involved with imaging cancer patients. The difficulty, however, lies in appropriately constructing LoA when using the percent difference between sums of tumor measurements to quantify change instead of the simple difference. The RD is a ratio of a numerator and denominator that are correlated and are themselves sums of correlated measurements that in practice frequently do not have a symmetric distribution consistent with the normal distribution. Below we discuss the methodology and demonstrate its application to an analysis of data collected on prostate cancer patients imaged with positron emission tomography (PET) using two different radiotracers.

2 Illustrative data example

Although FDG is the most widely used radiotracer for assessing therapy response,[16] there are other radiotracers available that could be used for this purpose. Fox and colleagues studied patients with progressive prostate cancer who had multiple metastatic bone and soft-tissue lesions.[17] They highlight that there is a need to standardize imaging analysis of lesions to allow lesions to be appropriately tracked. Part of this process entails looking at agreement in measurements made using different tracers. High reproducibility between radiotracers not only ensures comparability across studies, but would theoretically also allow radiotracers to be switched in the course of following a single patient.

In their study, each patient was imaged with PET/CT using two different radiotracers, FDG and ^{18}F-16β-fluorodihydrotestosterone (FDHT) within a 24-hour window in order to study the reproducibility of the measurements between the two radiotracers. This study was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board (MSKCC IRB). Written informed consent was obtained for all subjects. The consent procedure was approved by the MSKCC IRB. Details of the data collection, image acquisition, co-registration of the scans from the two tracers, and PET images showing the uptake of the two radiotracers have been published previously.[17]

We will use this data to illustrate the methods described below. We have data available on the maximal standardized uptake value $\mathrm{SUV}_{\max}$ measurements for FDG-PET and FDHT-PET on 167 lesions in 42 patients. The number of lesions per patient ranges from one to sixteen.

3 Methods

3.1 Definition of the relative difference

In general, a patient may have multiple lesions, and we denote these by l with l = 1,...,n_i where n_i is the number of the lesions for the i^{th} patient and N is the total number of patients. For each lesion there are two measurements, X_1 and X_2, (for instance representing two radiotracers, radiologists, or other conditions being evaluated). Thus, a patient has n_i pairs, (X_{1il}, X_{2il}), where the pairs may be correlated within patient. Using the same approach that is taken in clinical practice and is suggested by RECIST, the relative difference in total measured tumor burden between the replicated measurements is calculated by summing up tumor measurements within patients separately for X_1 and X_2 and then taking the relative difference between these summed quantities. That is, we sum across the lesions within a patient so that each patient has only a single number quantifying (total) tumor burden for a given method. The relative difference in tumor burden for the i^{th} patient is then defined as:

$$RD_i = 100 \times \frac{\sum_{l=1}^{n_i} X_{1l} - \sum_{l=1}^{n_i} X_{2l}}{\sum_{l=1}^{n_i} X_{1l}} \quad (1)$$

In the case where there is only a single lesion for a patient, this definition still holds; the sums are simply replaced by the single tumor measurement.

3.2 Distribution of the relative difference in tumor measurements

In order to work with the RD and produce a Bland-Altman plot that includes LoA we need to know the distribution of the RD. In practice, we have found that tumor measurements tend to have a skewed distribution that is not consistent with a normal distribution but is more characteristic of a lognor-

mal distribution. In this case, the RD is a function of the ratio of summands of correlated lognormal measurements and is not normally distributed. That is, the measurements made by the same method on the different lesions within a patient are correlated, the measurements made by the two different methods on the same lesion are also correlated, and these measurements are components of sums that appear both in the numerator and the denominator of a ratio inducing another level of correlation that needs to be taken into account. Identifying the appropriate distribution upon which to base the LoA poses a challenge that has not been directly addressed in the literature previously.

Assuming tumor measurements follow a lognormal distribution, calculating the RD as we described results in sums of correlated lognormal random variables in both the numerator and denominator of RD. In this case, there is no closed form analytic expression for the distribution of the RD. It is a well-known fact that the distribution of the sum of lognormal random variables does not have a closed form expression. Multiple papers have discussed this point and proposed ways to work with this sum.[18–20] One widely used result is that the distribution of this sum can be approximated by a lognormal random variable. The Fenton-Wilkonson approximation, developed for the case when the summands are independent, is frequently used to approximate the mean and variance of this distribution by matching the first two moments of the distribution of the sum with the first two moments of the approximating single lognormal distribution.[18] This work was extended by Abu-Dayya and Beaulieu[21] to accommodate correlated summands and later by Ligeti[22] to show that the distribution of the ratio of correlated sums of lognormals is well-approximated by a lognormal distribution. Hence, in accordance with this previous work, the distribution of the RD can be approximated by a negative shifted lognormal distribution.

3.3 Bland-Altman plots and limits of agreement

A Bland-Altman plot based on the RD consists of a single point for each patient with RD_i plotted against the patient's average tumor burden,

$$Avg_i = \frac{1}{2}\left(\sum_{l=1}^{n_i} X_{1il} + \sum_{l=1}^{n_i} X_{2il}\right) \tag{2}$$

To obtain LoA, we:

(1) Define Y to be the normalizing transformation of RD. Calculate it for each patient as:

$$Y_i = \ln\left(1 - \frac{RD_i}{100}\right) \tag{3}$$

where ln is the natural logarithm.

(2) Calculate the sample mean and standard deviation of Y,

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4}$$

and

$$sd = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5}$$

(3) The 95% LoA of RD are then:

$$100 \times (1 - e^{\bar{y} \pm 1.96 sd}) \tag{6}$$

3.4 Relationship between the relative difference and tumor size

In some applications, the level of agreement may be a function of lesion size causing the RD to vary in a systematic way with the average tumor burden. For instance, if small lesions are more difficult to measure, we may see better agreement for larger lesions. This would be noticeable on the Bland-Altman plot by a larger spread of the RD as tumor burden increased. The formula for the LoA given above may be too wide for patients with small tumor burden and too wide for patients with a larger tumor burden.

In this situation, we can use linear regression to obtain LoA that vary with tumor burden. We fit the regression model:

$$Y = \alpha_0 + \alpha_1 Avg + \varepsilon \tag{7}$$

obtain estimates of the regression coefficients which we write as $\hat{\alpha_0}$ and $\hat{\alpha_1}$ and an estimate of the square root of variance of the random error, denoted as $\hat{\sigma}$. The LoA adjusted for average tumor burden are then:

$$100 \times (1 - e^{\hat{\alpha_0} + \hat{\alpha_1} Avg \pm 1.96\hat{\sigma}}) \tag{8}$$

To determine whether it is necessary to use the LoA based on a regression model as opposed to using the constant LoA detailed above, beyond examining the Bland-Altman plot for a noticeable trend, we can test whether α_1 is significantly different from zero. If it is, then this would aid in suggesting that the agreement between measurements depends on the size of the measurements and the regression-based LoA should be used.[23] We have also found, however, that if this approach is used when agreement does not substantially depend on the size of the measurements, the resulting regression-based LoA may be a poor fit for the data.

4 Simulation study

We conducted a simulation study in order to explore how well the approach we outline above estimates the LoA. We generated N = 25, 50, and 100 observations or "patients" from a multivariate lognormal distribution $(X_{1i}, X_{2i})^T$

~MLVN ($(\mu_1, \mu_2)^T$, $\sum \sigma_x$), where X_{1i} and X_{2i} were allowed to be vectors of length n_i = 3, 5, or 10 representing the number of lesions measured. μ_1 and μ_2, the distribution means, are vectors of the same length as the corresponding X_{1i} and X_{2i}. We based the parameter distribution values on previous studies done in patients with non-small-cell lung cancer where the average tumor measurement was between 34 mm and 37 mm and the standard deviation of the tumor measurements ranged from 16 mm to 20 mm.[11,13] We show results for when the component values of μ_1 and σ_1 are fixed at 35 and 16, respectively. We allowed the values of μ_2 to vary from between 35 to 25, and the values of σ_2 to vary from between 16 to 19. We chose the covariance between X_{1i} and X_{2i} in order to give relatively high values of correlation, ρ between X_{1i} and X_{2i}, as would be expected in comparing two different measurements of the same tumor, and show results for when ρ is 0.5 and 0.9. Shown are results for when the correlation between X_{1ij} and X_{1ik}, and similarly X_{2ij} and X_{2ik}, $j \neq k$, is equal 0.1. Results for when this correlation is larger do not vary substantively from what is shown.

All simulations were performed 1,000 times. Because we needed an automated algorithm for the simulations, for each generated dataset, a linear regression was fit. If the slope of the fitted line was significantly different from 0, the regression approach to the LoA was used. Otherwise, the constant LoA were used. We present the average proportion of data that falls outside the upper and lower limits across the 1,000 simulations.

For comparison we also evaluate what happens if the log-normal distribution of the RD is ignored and the LoA are estimated assuming it is normally distributed (*i.e.* $\overline{RD} \pm 1.96 s_{RD}$) where $\overline{RD}$ is the mean relative difference and s_{RD} is the standard deviation of the relative differences.

Table 1 shows results for when there is a moderate amount of agreement between the two measurements and Table 2 shows results for when there is a high amount of agreement between the two measurements. Ideally we would expect to see 0.025 of the data falling above the upper LoA and 0.025 of the data falling below the lower LoA. For larger sample sizes, the approximated distribution of the RD appears to serve as a good basis for constructing LoA. For smaller sample sizes it does not work quite as well. At least in part, this is simply a function of the difficulty with constructing the LoA using small samples. For instance, with N = 25 we want 95% of the data points to lie within the 95% LoA which leaves 1 data point (25×0.05 = 1.25) to lie outside the LoA, either above the upper limit or below the lower limit. In comparison, the LoA naively constructed using the normal LoA, while on average including 95% of the data within the LoA, uniformly include too much data within the upper limit and leave too little data out of the lower limit. This performance is a result of the fact that the limits based on the normal distribution are necessarily symmetric around the average RD. For a fixed sample size, N, changing the number of lesions that are measured on each subject does not appear to affect the performance of the LoA.

5 Illustrative data example results

Using the illustrative PET data example, Figures 1a and 1b show plots of the $\mathrm{SUV}_{\max}$ measurements made on the individual lesions by FDG-PET and FDHT-PET. They both depict measurements that are skewed to the right. In contrast in Figure 1c, when the measurements are aggregated and RD is calculated at the patient-level, the distribution is skewed to left and clearly does not follow a normal distribution. In Figure 1d, the transformed variable, Y, is plotted and depicts a distribution that more closely follows a normal distribution.

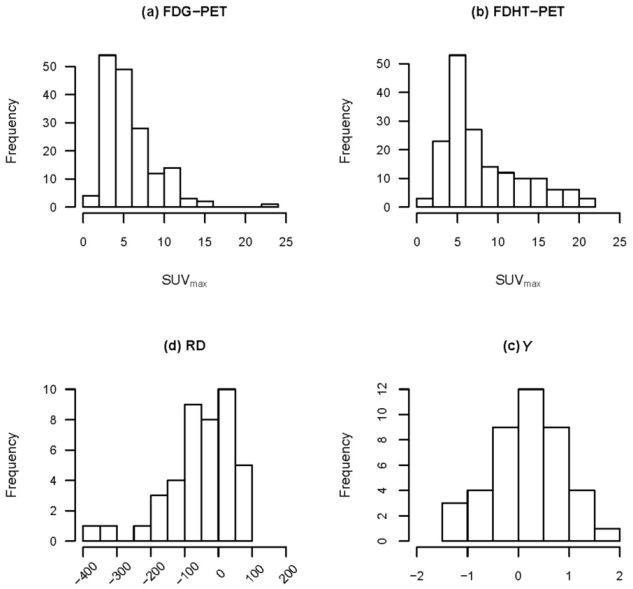

Figure 1: FDG-PET and FDHT-PET measurements
Histograms showing the distribution of the lesion: (a) $\mathrm{SUV}_{\max}$ measurements for FDG-PET, (b) $\mathrm{SUV}_{\max}$ measurements for FDHT-PET and the patient-level, (c) relative difference (%), and (d) normalized relative difference, Y, between the two tracers.

Figure 2 contains a Bland-Altman plot of this data which includes the 95% LoA. The mean RD, -52.8%, is shown by the thick, solid, black line. The 95% LoA calculated using the formula above, (-382.1%, 68.7%), are shown by the thick, dashed, black lines. They suggest that the relative difference in patient tumor burden between most pairs of FDG-PET and FDHT-PET $\mathrm{SUV}_{\max}$ measurements taken at essentially the same time will fall within this range.

Although there are no standard thresholds defining response categories for metabolic imaging in the same way that RECIST defines response thresholds for functional imaging, it has been suggested that a decrease of 30% in serial SUV measurements from FDG-PET indicates that a patient is

having a partial response to treatment.[3,9] Figure 2 demonstrates a clear lack of agreement between FDG-PET and FDHT-PET. It shows, for instance, that a patient with no real change in their tumor burden had a seemingly substantial decrease in metabolic activity with an SUV_{max} decreasing by over 380% due to changing the radiotracer from FDG to FDHT.

Table 1: Limits of agreement for moderate correlation between measurements

N	Number of lesions	μ_1	σ_1	Average RD	Proportion of data outside 95% LoA			
					Lognormal		Normal	
					Lower	Upper	Lower	Upper
25	3	35	18.7	-5.3	0.021	0.022	0.044	0.003
		34	18.7	-2.3	0.020	0.021	0.042	0.003
		34	16.6	-2.7	0.021	0.022	0.043	0.002
		25	16.6	25.6	0.021	0.022	0.043	0.002
25	5	35	18.7	-3.8	0.020	0.020	0.039	0.005
		34	18.7	-0.4	0.021	0.021	0.040	0.004
		34	16.6	-1.2	0.022	0.022	0.040	0.005
		25	16.6	26.5	0.022	0.021	0.043	0.003
25	10	35	18.7	-2.5	0.020	0.022	0.036	0.007
		34	18.7	0.6	0.021	0.022	0.038	0.007
		34	16.6	0.5	0.020	0.021	0.036	0.007
		25	16.6	27.3	0.020	0.022	0.039	0.005
50	3	35	18.7	-5.3	0.023	0.022	0.044	0.002
		34	18.7	-2.1	0.023	0.023	0.044	0.002
		34	16.6	-2.6	0.022	0.023	0.042	0.002
		25	16.6	25.8	0.023	0.023	0.045	0.001
50	5	35	18.7	-3.7	0.023	0.023	0.041	0.004
		34	18.7	-0.7	0.023	0.023	0.043	0.004
		34	16.6	-1.0	0.023	0.023	0.040	0.005
		25	16.6	26.6	0.023	0.024	0.043	0.003
50	10	35	18.7	-2.5	0.024	0.022	0.040	0.007
		34	18.7	0.4	0.023	0.024	0.038	0.006
		34	16.6	0.2	0.023	0.023	0.039	0.007
		25	16.6	27.3	0.023	0.024	0.041	0.004
100	3	35	18.7	-5.4	0.024	0.025	0.043	0.002
		34	18.7	-1.9	0.024	0.024	0.045	0.002
		34	16.6	-2.6	0.024	0.023	0.042	0.002
		25	16.6	25.9	0.023	0.025	0.045	0.001
100	5	35	18.7	-3.6	0.024	0.024	0.041	0.004
		34	18.7	-0.7	0.024	0.024	0.043	0.004
		34	16.6	-0.8	0.024	0.024	0.041	0.004
		25	16.6	26.6	0.023	0.025	0.043	0.002
100	10	35	18.7	-2.5	0.024	0.024	0.039	0.007
		34	18.7	0.5	0.024	0.024	0.040	0.006
		34	16.6	0.3	0.024	0.024	0.038	0.007
		25	16.6	27.3	0.024	0.025	0.041	0.004

Note. The average proportion of data outside of the 95% LoA across 1,000 simulations is shown for measurements generated from several different lognormal distributions all with $\rho = \text{cor}(Y_{1i}, Y_{2i}) = 0.5$, $\mu_1^T = 35^T$ and $\sigma_1 = 18.7$. The lognormal LoA are calculated assuming individual tumor measurements follow a lognormal distribution and approximate the distribution of RD with a lognormal distribution. The normal LoA are calculated assuming RD follows a normal distribution.

For comparison, the grey, solid lines demonstrate what would happen to the estimates of the LoA had we used the original normal-based equations for the LoA. The resulting interval, (-251.0%, 145.4%), is markedly different particularly at the upper bound of the interval which overestimates the increase we would expect to see even if there were no difference in the lesions. This reflects the results from the simulation study; a symmetric interval is not appropriate for data with a skewed distribution.

In this application, the level of agreement does not vary systematically with tumor burden. The regression-based LoA shown in Figure 2 appear too wide for the data.

Table 2: Limits of agreement for high correlation between measurements

N	Number of lesions	μ_1	σ_1	Average RD	Proportion of data outside 95% LoA			
					Lognormal		Normal	
					Lower	Upper	Lower	Upper
25	3	35	18.7	-1.5	0.022	0.022	0.036	0.008
		34	18.7	1.7	0.021	0.021	0.033	0.009
		34	16.6	0.6	0.023	0.021	0.035	0.008
		25	16.6	28.8	0.022	0.021	0.036	0.007
25	5	35	18.7	-1.5	0.021	0.022	0.035	0.010
		34	18.7	1.6	0.021	0.021	0.033	0.009
		34	16.6	1.0	0.020	0.021	0.033	0.010
		25	16.6	28.5	0.022	0.021	0.038	0.007
25	10	35	18.7	-1.4	0.020	0.021	0.034	0.010
		34	18.7	1.5	0.021	0.020	0.033	0.009
		34	16.6	1.3	0.022	0.021	0.035	0.009
		25	16.6	28.1	0.022	0.022	0.036	0.008
50	3	35	18.7	-1.7	0.024	0.024	0.037	0.009
		34	18.7	1.4	0.023	0.023	0.036	0.009
		34	16.6	0.6	0.024	0.023	0.037	0.010
		25	16.6	28.8	0.023	0.024	0.039	0.007
50	5	35	18.7	-1.6	0.023	0.023	0.035	0.010
		34	18.7	1.6	0.023	0.023	0.036	0.010
		34	16.6	1.0	0.024	0.023	0.036	0.010
		25	16.6	28.6	0.023	0.024	0.038	0.007
50	10	35	18.7	-1.3	0.023	0.022	0.035	0.010
		34	18.7	1.8	0.024	0.023	0.036	0.009
		34	16.6	1.5	0.023	0.024	0.036	0.010
		25	16.6	28.2	0.022	0.024	0.037	0.008
100	3	35	18.7	-1.6	0.024	0.023	0.037	0.009
		34	18.7	1.5	0.024	0.024	0.037	0.010
		34	16.6	0.7	0.025	0.024	0.038	0.010
		25	16.6	28.8	0.023	0.025	0.039	0.006
100	5	35	18.7	-1.4	0.024	0.024	0.037	0.011
		34	18.7	1.6	0.024	0.024	0.037	0.010
		34	16.6	1.0	0.024	0.024	0.036	0.010
		25	16.6	28.5	0.024	0.024	0.039	0.007
100	10	35	18.7	-1.4	0.024	0.024	0.036	0.011
		34	18.7	1.7	0.024	0.024	0.036	0.010
		34	16.6	1.4	0.024	0.024	0.036	0.010
		25	16.6	28.3	0.023	0.025	0.037	0.008

Note. The average proportion of data outside of the 95% LoA across 1,000 simulations is shown for measurements generated from several different lognormal distributions all with $\rho = \text{cor}(Y_{1i}, Y_{2i}) = 0.9$, $\mu_1^T = 35^T$ and $\sigma_1 = 18.7$. The lognormal LoA are calculated assuming individual tumor measurements follow a lognormal distribution and approximate the distribution of RD with a lognormal distribution. The normal LoA are calculated assuming RD follows a normal distribution.

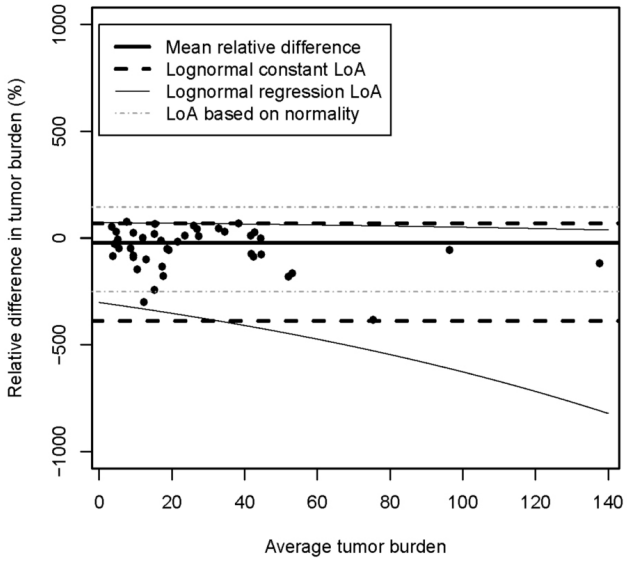

Figure 2: Bland-Altman plot for patient tumor burden with PET data

A Bland-Altman plot evaluating the agreement between $SUV_{\max}$ measurements from FDG- PET and FDHT-PET for total patient tumor burden. The solid, thick line shows the mean RD of -52.8% between the two measurements. The thick, dashed lines show the constant lognormal 95% LoA, (-382.1%, 68.7%). The thin, solid line reflects the regression-based LoA based on the fitted regression line Y =.04 + 0.006Avg. The grey dashed-dotted lines demonstrate the estimated 95% LoA incorrectly calculated assuming RD is normally distributed.

6 Discussion

A critical point when interpreting Bland-Altman plots and the LoA is gauging how far apart measurements can be before it is decided that there is not sufficient agreement between the measurements. Statisticians can easily produce Bland-Altman plots of absolute change and simple differences, but may have difficulty providing guidance on what constitutes acceptable agreement to their radiology collaborators particularly when talking about differences in tumor measurements on the absolute scale. As Bland and Altman point out, this is not a statistical question but a clinical one.[23] A key feature of our approach is the use of a relative difference metric that has long become standard of clinical care. Using a Bland-Altman plot and LoA based on the RD to evaluate agreement between solid tumor measurements allows inter- and intra-observer variability to be assessed with a clinically meaningful quantity and may facilitate interpretation. Furthermore, it provides a method for handling multiple lesions measurements per patient and analyzing data on a patient-level basis. It has not previously been clear how to use this metric for evaluating agreement in solid tumor measurements.

Using the RD for evaluating reproducibility and reliability is not an entirely novel suggestion. Others have published papers looking at the agreement in tumor measurements by plotting the RD in place of the simple difference in Bland-Altman plots.[8, 11, 13, 24] This approach, however, is far from consistently used and computation of the LoA has not reflected the skewed distribution of the RD.

To our knowledge, there is no literature discussing the distribution of the RD in tumor burden, and certainly no prior work describing how to construct appropriate LoA for it. This is true both for the case when all patients contribute a single lesion to the analysis and when patients may contribute multiple lesions, however the latter is of more clinical interest and presents a bigger methodological challenge. Instead of keeping the multiple within-patient measurements separate as has been done previously in the statistical literature,[23] the RD collapses the data into one measurement per patient posing a methodological challenge that has not previously been addressed.

It is important to note that this metric has some undesirable properties, such as lack of symmetry with respect to X_1 and X_2. While in some cases there may be a particular reason to chose one method of measurement as X_1 and another as X_2, in other cases there are not, and the values of the RD and the LoA may differ based on how X_1 and X_2 are assigned. For instance, in the PET data we described we have a specific reason for choosing FDG-PET to be X_1. FDG, while not truly a "reference standard" in the sense that its use with PET is not considered a definitive test, is what is commonly used in practice and is the standard tracer used with PET. FDHT is newer and not commonly used. Hence, interest lies in evaluating tumor measurements made with the new tracer relative to the existing one. In contrast, if we were evaluating inter-reader agreement between two radiologists, there may be no particular reason why one radiologist should be chosen over the other to be assigned as X_1. This could potentially give rise to situations where there is a substantive difference depending on how X_1 is assigned. We recommend examining the RD and the LoA when one measurement method is assigned to be X_1 and a second time when the second measurement method is assigned to be X_1 to check for marked differences.

We noticed in our applications that level of agreement is sometimes a function of lesion size. This is sensible from a clinical standpoint; it has always been difficult to accurately image and evaluate small lesions. We incorporated regression methods to adjust our analysis of agreement for size. Despite the fact that such regression adjustments have long been a part of statistical literature, it appears that they have not been embraced by the radiology community since we have not seen them in our review of clinical publications reporting on agreement.

Acknowledgements

This work was supported by National Cancer Institute grants P50 CA086438 to the MSKCC Center for Molecular Imaging in Cancer and by the MSKCC core grant P30 CA008748.

References

[1] Eisenhauer EA, Therasse P, Bogaerts J, *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009; 45: 228-247. PMid: 19097774. http://dx.doi.org/10.1016/j.ejca.2008.10.026

[2] Miller AB, Hoogstraten B, Staquet M, *et al.* Reporting results of cancer treatment. Cancer. 1981; 47: 207-214. http://dx.doi.org/10.1002/1097-0142(19810101)47:1<207::AID-CNCR2820470134>3.0.CO;2-6

[3] Wahl RL, Jacene H, Kasamon Y, *et al.* From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. J Nucl Med. 2009; 50 Suppl 1: 122S-150S. PMid: 19403881. http://dx.doi.org/10.2967/jnumed.108.057307

[4] Erasmus JJ, Gladish GW, Broemeling L, *et al.* Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol. 2003; 21: 2574-2582. PMid: 12829678. http://dx.doi.org/10.1200/JCO.2003.01.144

[5] Sargent DJ, Rubinstein L, Schwartz L, *et al.* Validation of novel imaging methodologies for use as cancer clinical trial end-points. Eur J Cancer. 2009; 45: 290-299. PMid: 19091547. http://dx.doi.org/10.1016/j.ejca.2008.10.030

[6] Provenzale JM, Ison C, Delong D. Bidimensional measurements in brain tumors: assessment of interobserver variability. AJR Am J Roentgenol. 2009; 193: W515-522. PMid: 19933626. http://dx.doi.org/10.2214/AJR.09.2615

[7] de Langen AJ, Lubberink M, Boellaard R, *et al.* Reproducibility of tumor perfusion measurements using 15O-labeled water and PET. J Nucl Med. 2008; 49: 1763-1768. PMid: 18927324. http://dx.doi.org/10.2967/jnumed.108.053454

[8] Bauknecht HC, Romano VC, Rogalla P, *et al.* Intra- and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. Invest Radiol. 2010; 45: 49-56. PMid: 19996757. http://dx.doi.org/10.1097/RLI.0b013e3181c02ed5

[9] Gietema HA, Schaefer-Prokop CM, Mali WP, *et al.* Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT– influence of inspiration level, nodule size, and segmentation performance. Radiology. 2007; 245: 888-894. PMid: 17923508. http://dx.doi.org/10.1148/radiol.2452061054

[10] Goodman LR, Gulsun M, Washington L, *et al.* Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. AJR Am J Roentgenol. 2006; 186: 989-994. PMid: 16554568. http://dx.doi.org/10.2214/AJR.04.1821

[11] Oxnard GR, Zhao B, Sima CS, *et al.* Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol. 2011; 29: 3114-3119. PMid: 21730273. http://dx.doi.org/10.1200/JCO.2010.33.7071

[12] Wormanns D, Kohl G, Klotz E, *et al.* Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. Eur Radiol. 2004; 14: 86-92. PMid: 14615902. http://dx.doi.org/10.1007/s00330-003-2132-0

[13] Zhao B, James LP, Moskowitz CS, *et al.* Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009; 252: 263-272. PMid: 19561260. http://dx.doi.org/10.1148/radiol.2522081593

[14] Altman DG, Bland JM. Measurement in Medicine - the Analysis of Method Comparison Studies. Statistician. 1983; 32: 307-317. http://dx.doi.org/10.2307/2987937

[15] Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet. 1986; 1: 307-310. http://dx.doi.org/10.1016/S0140-6736(86)90837-8

[16] Kelloff GJ, Hoffman JM, Johnson B, *et al.* Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. Clin Cancer Res. 2005; 11: 2785-2808. PMid: 15837727. http://dx.doi.org/10.1158/1078-0432.CCR-04-2626

[17] Fox JJ, Autran-Blanc E, Morris MJ, *et al.* Practical approach for comparative analysis of multilesion molecular imaging using a semiautomated program for PET/CT. J Nucl Med. 2011; 52: 1727-1732. PMid: 21984797. http://dx.doi.org/10.2967/jnumed.111.089326

[18] Fenton LF. The sum of log-normal probability distibutions in scattered transmission systems. IRE Transactions on Communication Systems. 1960; 8: 57-67. http://dx.doi.org/10.1109/TCOM.1960.1097606

[19] Naus JI. Distribution of Logarithm of Sum of 2 Log-Normal Variates. Journal of the American Statistical Association. 1969; 64: 655. http://dx.doi.org/10.1080/01621459.1969.10501004

[20] Schwartz SC, Yeh YS. On the Distribution Function and Moments of Power Sums with Log-Normal Components. Bell System Technical Journal. 1982; 61: 1441-1462. http://dx.doi.org/10.1002/j.1538-7305.1982.tb04353.x

[21] Abu-Dayya AA, Beaulieu NC. Outage probabilities in the presence of correlated log- normal interferers. IEEE Transactions on Vehicular Technology. 1994; 1: 164-173. http://dx.doi.org/10.1109/25.282277

[22] Ligeti A. Outage probabilities in the presence of correlated lognormal useful and interfering components. IEEE Communication Letters. 2000; 4: 15-17. http://dx.doi.org/10.1109/4234.823535

[23] Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999; 8: 135-160.24.

[24] Velasquez LM, Boellaard R, Kollia G, *et al.* Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. J Nucl Med. 2009; 50: 1646-1654. PMid: 19759105. http://dx.doi.org/10.2967/jnumed.109.063347