Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2011 *Paper* 21

Bland-Altman Plots for Evaluating Agreement Between Solid Tumor Measurements

Chaya S. Moskowitz[*] Mithat Gonen[†]

[*] Memorial Sloan-Kettering Cancer Center, moskowc1@mskcc.org

[†] Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

Bland-Altman Plots for Evaluating Agreement Between Solid Tumor Measurements

Chaya S. Moskowitz and Mithat Gonen

Abstract

Rationale and Objectives. Solid tumor measurements are regularly used in clinical trials of anticancer therapeutic agents and in clinical practice managing patients' care. Consequently studies evaluating the reproducibility of solid tumor measurements are important as lack of reproducibility may directly affect patient management. The authors propose utilizing a modified Bland-Altman plot with a difference metric that lends itself naturally to this situation and facilitates interpretation. Materials and Methods. The modification to the Bland-Altman plot involves replacing the difference plotted on the vertical axis with the relative percent change (RC) between the two measurements. This quantity is the same one used in assessing tumor response to therapeutic agents and is very familiar to radiologists and clinicians working with cancer patients.The distribution of the RC is explored and revised equations for the limits of agreement (LoA) are presented. These methods are applied to positron emission tomography (PET) data studying two radiotracers. Results. The RC can be calculated separately for each lesion measured or at the patient level by summing over lesions within patient. In both cases, the distribution of the RC is highly skewed and is approximated by a negative shifted lognormal distribution. The standard equations for the 95% LoA assume the differences are approximately normally distributed and are not appropriate for the RC. Conclusions. The modified Bland-Altman plot with correctly calculated LoA can aid in evaluating agreement between solid tumor measurements.

Bland-Altman Plots for Evaluating Agreement Between Solid Tumor Measurements

Chaya S. Moskowitz, Ph.D.*, Mithat Gönen, Ph.D.

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
307 E 63rd Street, New York, NY 10065
Phone: (646) 735-8100 Fax: (646) 735-0010

Abstract

Rationale and Objectives. Solid tumor measurements are regularly used in clinical trials of anti-cancer therapeutic agents and in clinical practice managing patients' care. Consequently studies evaluating the reproducibility of solid tumor measurements are important as lack of reproducibility may directly affect patient management. The authors propose utilizing a modified Bland-Altman plot with a difference metric that lends itself naturally to this situation and facilitates interpretation.

Materials and Methods. The modification to the Bland-Altman plot involves replacing the difference plotted on the vertical axis with the relative percent change (RC) between the two measurements. This quantity is the same one used in assessing tumor response to therapeutic agents and is very familiar to radiologists and clinicians working with cancer patients.The distribution of the RC is explored and revised equations for the limits of agreement (LoA) are presented. These methods are applied to positron emission tomography (PET) data studying two radiotracers.

Results. The RC can be calculated separately for each lesion measured or at the patient level by summing over lesions within patient. In both cases, the distribution of the RC is highly skewed and is approximated by a negative shifted lognormal distribution. The standard equations for the 95% LoA assume the differences are approximately normally distributed and are not appropriate for the RC.

Conclusions. The modified Bland-Altman plot with correctly calculated LoA can aid in evaluating agreement between solid tumor measurements.

Keywords: reproducibility, inter-observer variability, limits of agreement, multiple lesions

*Corresponding author
Email address: moskowc1@mskcc.org (Chaya S. Moskowitz, Ph.D.)

Preprint submitted to Elsevier *September 23, 2011*

1. Introduction

The Bland-Altman plot is a graphical tool that is frequently used to evaluate reproducibility and repeatability. In its original and most typically used form, it is constructed by plotting the differences between two measurements, $d = X_1 - X_2$, by the mean of the measurements, $\frac{1}{2}(X_1 + X_2)$. To accompany the plots, analysts usually present the 95% limits of agreement (LoA) which are defined $\bar{d} \pm 1.96s$ where $\bar{d}$ is the average difference and s is the standard deviation of d [1, 2].

Here we are concerned with the application of Bland-Altman plots to evaluating agreement between solid tumor measurements. Serial measurements of solid tumors are used to gauge whether a tumor is responding to an anti-cancer therapeutic agent both when managing a patient's care and when testing new treatments in clinical trials. This assessment is most frequently done using the change in tumor size as seen on anatomic imgaing. Under guidelines established for evaluating response in solid tumors using anatomic imaging [3, 4], the relative percent change in tumor size measured at a baseline, pre-treatment time, say X_B, and a follow-up time after treatment has commenced, X_F, is calculated as RC$= 100 \times \frac{X_B - X_F}{X_B}$. In patients who have multiple tumors measured, X_B and X_F are taken to be the sums of the tumor measurements at each time and signify patients' tumor burden. This continuous RC is usually divided into four response categories representing patients with a complete response, with a partial response, with stable disease, and with progressive disease. Non-anatomic serial measurements of solid tumors from 2-[^{18}F]Fluoro-2-deoxyglucose positron emission tomography (FDG-PET) are also used to evaluate tumor response in a more limited but growing number of oncologic settings [5]. While currently there is variability across studies in how tumor response is quantified using FDG-PET (e.g. measurements could be quantified with different metrics such as the maximal standardized uptake value (SUV$_{\text{max}}$), total lesion glycolysis, or SUV normalized to lean body mass to name a few), it has been suggested that the relative change from the baseline value should be used to quantify tumor response [6].

Many papers have looked at the reproducibility or repeatability of solid tumor measurements. While some of these papers focus on inter- and intra-observer agreement for the four response categories (using a kappa statistic, for example), multiple papers report on the agreement for the continuous measurements made on individual tumors. Given the previous studies and the recent calls by some that tumor response assessment in clinical trials should move away from the response categories and consider the change in tumor measurements on a continuous scale [6–10], it is important to ensure that agreement between continuous tumor measurements be analyzed using appropriate statistical methods which allow for meaningful interpretations.

In reviewing the literature both in clinical and radiology journals, we noted multiple cases where Bland-Altman plots and LoA were presented to look at agreement between solid tumor measurements. (For some recent examples, see [11–19].) Across the literature, there was variation in how the plots and LoA were constructed. While the method originally suggested by Bland and Altman uses the differences, d, there is no reason why the Bland-Altman plots and LoA cannot be based on other quantities that may be more useful in other situations. Some of these papers took advantage of this fact and present plots using other quantities. It appears, however, that there is a lack of consistency in how researchers looking at agreement in tumor measurements analyze such data. Moreover, in many papers there were errors in the calculation of the LoA.

The validity of LoA intervals depends on certain assumptions being met. In papers following their original publications suggesting the use of these plots, Bland and Altman point out that the calculation of the LoA as they initially presented it assumes that the pairs of measurements used in the calculations are independent across the observations [20, 21]. That is, suppose that X_1 is the first measurement (e.g. the measurement made by the first reader or at the first read) and X_2 is the second measurement (by the second reader or at the second read). Then the pair of measurements for the i^{th} observation is denoted as (X_{1i}, X_{2i}). Bland and Altman's point is that the standard LoA calculation assumes that (X_{1i}, X_{2i}) is independent of (X_{1j}, X_{2j}) for $i \neq j$. In our situation, if the observations consist of measurements made on multiple tumors within the same individual, the observations are correlated and the assumption of independence is violated. Although there have been a couple of papers discussing how to compute the LoA appropriately accounting for this correlation [20–22], we have not seen any examples of this methodology being used with measurements of tumor size. A second assumption made in the definition of the LoA is that the differences, d, are normally distributed. With studies basing their analysis on metrics other than d, there is a need to ensure that the distribution assumed in calculating the intervals aligns with the distribution of the chosen metric. If the wrong distribution is used, the interval does not yield 95% coverage. In other words, while the idea behind the 95% LoA is that we expect 95% of the differences (however "difference" is measured) to be within the 95% LoA, if the interval is not properly constructed then it is likely that a substantially lower (or, in some cases, higher) proportion of the differences may fall into the interval.

In light of these observations, we suggest a clinically relevant metric upon which to base Bland-Altman plots and the corresponding LoA when evaluating agreement between replicated tumor measurements. We also explore its distribution and suggest methods for constructing appropriate LoA. These methods are applied to data collected on prostate cancer patients imaged with PET using two different radiotracers.

2. Methods

While the differences $X_1 - X_2$ may be the most relevant quantity to study in some situations, within the context of tumor response assessment we suggest that the more useful quantity to work with is RC. This quantity is the one that is used to make clinical decisions on a daily basis. Recall that we conclude that the inter- or intra-observer agreement is acceptable if the differences within the LoA are not clinically important. While some clinicians and radiologists may be able to look at $X_1 - X_2$ and evaluate whether the differences are sufficiently small or unacceptably large, using RC as the metric translates the results of the analysis onto a scale that facilitates interpretation and permits a wider audience to have a better sense of the impact on clinical decision making. In other words, it allows a direct interpretation in terms of the most clinically relevant quantity.

We note that this is not an entirely novel suggestion. Others have published papers looking at the agreement in tumor measurements by plotting RC in place of $X_1 - X_2$ in Bland-Altman plots [15, 18, 19]. This approach, however, is not consistently used. Furthermore, difficulty lies in how to appropriately construct the LoA when using a fraction to quantify change instead of the simple difference. In our experience, tumor measurements are not normally distributed and the distribution of RC is not symmetric. Hence, estimating the LoA by taking the average RC and adding and subtracting a factor of 1.96 times the standard deviation of RC likely results in an interval that does not give the stated coverage. Below we suggest more appropriate methods for obtaining the LoA based on RC. We distinguish the cases depending on whether single or multiple tumors are being considered for each patient.

2.1. One tumor per patient

When only one tumor per patient is being used to evaluate reproducibility, there is a single pair of measurements, (X_{1i}, X_{2i}), for the i^{th} patient and the pairs are independent across patients. Similar to what others have observed [7, 8], in the applications we have studied we find that a lognormal distribution is a reasonable approximation for the distribution of tumor measurements. If X_1 and X_2 both follow lognormal distributions, then their ratio, $\frac{X_2}{X_1}$, also has a lognormal distribution. The distribution of RC, which can be rewritten as $\text{RC} = 100 - 100 \times \frac{X_2}{X_1}$ is called a negative shifted lognormal distribution [23] or simply lognormal [24].

To obtain the 95% LoA for RC, let $Y = \ln\,(1 - \frac{\text{RC}}{100})$. Calculate the sample mean of Y, $\bar{y}$, and the sample standard deviation of Y, s. The 95% LoA are then $100 \times (1 - e^{\bar{y} \pm 1.96 s})$. To obtain the LoA for other probability levels, the value of 1.96 is changed by taking different quantiles from the standard normal distribution.

2.2. Multiple tumors per patient

We discuss two options for analyzing RC when there are multiple lesions per patient. The first option is to keep each lesion as a separate observation. RC is calculated as above and each tumor is represented as a single point on the Bland-Altman plot. In contrast, the estimate of the variance of RC used in the LoA is different from above accounting for the clustering of lesions within patient. The second option is to calculate the relative change in total measured tumor burden between the replicated measurements by summing up tumor measurements at each time. This metric is in fact what is suggested by RECIST [3] and what is typically used in clinical practice. In this case, each point on the Bland-Altman plot represents a patient and both the definition and the distribution of the relative change are different from above. We describe this approach in further detail below.

2.2.1. Tumors as the unit of analysis

We use l to denote lesions with $l = 1, ..., n_i$ where n_i is the number of lesions measured for the i^{th} patient. Thus for each patient, there are n_i pairs, (X_{1il}, X_{2il}), where the pairs may be correlated within patient. For each lesion we calculate $RC_{il} = \frac{X_{1il} - X_{2il}}{X_{1il}}$. While several different correlation structures are theoretically possible, we focus on the compound symmetric correlation structure which assumes that any two lesions within an individual have the same degree of correlation and this correlation is the same across all patients. More formally, correlation$(\text{RC}_{il}, \text{RC}_{il'}) = \rho$ for $l \neq l'$. We have found that this structure works reasonably well in practice. Along the lines of what is suggested in [22], we can fit a random effects model to obtain the estimated variance.

Let $Y_{il} = \ln\,(1 - \frac{\text{RC}_{il}}{100})$. The model we fit is $y_{il} = \mu + b_i + \epsilon_{il}$ where μ is the overall mean, $b_i \sim N(0, \sigma_b^2)$ is the subject random effect, and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is the residual error. To obtain the 95% LoA, calculate the sample mean of Y, $\bar{y}$. Fit the random effects model using any standard statistical software package. Add together the variance estimates for the subject random effects and the error terms and take the square root, $s = \sqrt{\hat{\sigma}_b^2 + \hat{\sigma}_\epsilon^2}$, to use as an estimate of the standard deviation of Y. Plug these estimates into the same equation used above, $100 \times (1 - e^{\bar{y} \pm 1.96 s})$, to obtain the 95% LoA.

2.2.2. Patients as the unit of analysis

Using the same notation as above, here we sum across the lesions within a patient so that each patient has only a single measure of the change in (total) tumor burden. The relative change in tumor burden for the i^{th} patient is defined as

$$\text{RC}_{\text{total}} = 100 \times \frac{\sum_{j=1}^{n_i} X_{1j} - \sum_{j=1}^{n_i} X_{2j}}{\sum_{j=1}^{n_i} X_{1j}} = 100 - 100 \times \frac{\sum_{j=1}^{n_i} X_{2j}}{\sum_{j=1}^{n_i} X_{1j}}.$$

With tumor measurements following a lognormal distribution, we are now faced with a ratio whose numerator and denominator are sums of correlated lognormal random variables. It is a well-known fact that the distribution of the sum of lognormal random variables does not have a closed form expression. However, multiple people have shown that this distribution can be approximated by a lognormal distribution. One frequently used approximation is the Fenton-Wilkonson approximation [25] which was developed for the case when the summands are independent. This work was extended by Abu-Dayya and Beaulieu [26] to accomodate correlated summands and later by Ligeti [27] to show that the distribution of the ratio of correlated sums of lognormals is well-approximated by a lognormal distribution. Using these results and reasoning similar to when there is one tumor per patient, the distribution of RC_{total} is approximately negative shifted lognormal.

To form the 95% LoA, let $Y = \ln\,(1 - \frac{\text{RC}_{\text{total}}}{100})$. Calculate the sample mean of Y, $\bar{y}$, and the sample standard deviation of Y, s. The 95% LoA are then $100 \times (1 - e^{\bar{y} \pm 1.96 s})$.

3. Results

Fox and colleagues [28] studied patients with progressive prostate cancer who had multiple metastatic bone and soft-tissue lesions. Briefly, as part of an IRB-approved protocol each patient was imaged with PET/CT using two different radiotracers, FDG and ^{18}F-16β-fluoro-dihydrotestosterone (FDHT). Both scans were done within a 24-hour window in order to study the reproducibility of the measurements. More details of the data collection, image acquisition, and co-registration of the scans from the two tracers can be found in [28]. For this analysis we have data available on the SUV$_{\text{max}}$ measurements for FDG-PET and FDHT-PET on 167 lesions in 42 patients. The number of lesions per patient ranges from one to sixteen. Here we look at the agreement between FDG-PET SUV$_{\text{max}}$ and FDHT-PET SUV$_{\text{max}}$. Because there are multiple lesions per patient, we first consider the agreement between measurements of each lesion individually and then the agreement between the measurements of total tumor burden for each patient.

In Figure 1, we plot the SUV$_{\text{max}}$ lesion measurement for both radiotracers. Note that the skewed shapes of the histograms are consistent with lognormal distributions. Taking the SUV$_{\text{max}}$ for FDG-PET to be X_1 and the SUV$_{\text{max}}$ for FDHT-PET to be X_2 in the equations above, we calculate RC_{il} at the lesion level and plot the results in Figure 2a. With values of RC$_{il}$ ranging from -809% to 81%, this histogram demonstrates clearly that RC does not follow a normal distribution. The Shapiro-Wilks test for normality suggests that the distribution of $Y = \ln(1 - \frac{RC}{100})$ is consistent with a normal distribution (p=0.52), suggesting that the negative lognormal distribution is a reasonable fit for RC estimated with this data.

To evaluate the agreement between the FDG-PET and FDHT-PET lesion measurements, we look at the Bland-Altman plot displayed in Figure 3. The mean RC, -77.0%, is shown by the black dotted line. The 95% LoA, (-485.3%, 69.6%), are shown by the black dashed lines. They suggest that the relative difference between most pairs of FDG-PET and FDHT-PET SUV$_{\text{max}}$ measurements taken from the same lesion at essentially the same time will fall within this range. In other words, by changing the radiotracer from FDG to FDHT, we might conclude that a lesion with no real change had a seemingly substantial decrease in metabolic activity with an SUV$_{\text{max}}$ decreasing by over 400%. Although there are no standard thresholds defining response categories for nonanatomic imaging in the same way that RECIST defines response thresholds for anatomic imaging, it has been suggested that a decrease of 30% in serial SUV measurements from FDG-PET is suggestive that a patient is having a partial response to treatment [6]. Placed within this context, Figure 3 very clearly indicates a lack of agreement between FDG-PET and FDHT-PET. For comparison, the grey dashed-dotted lines demonstrate what would happen to the estimates of the LoA had we used the original normal-based equations for the LoA that assume that the multiple lesions within a patient are independent of one another. The resulting interval, (-378.8%, 224.8%), 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 lesion.

If we instead quantify the change in the SUV measurements at the patient level in terms of $\mathrm{RC}_{\mathrm{total}}$, the resulting values, shown in Figure 2(b), are skewed as expected. The Shapiro-Wilks test for Y suggests data consistent with a normal distribution (p=0.83). The Bland-Altman plot looking at the agreement in the measures of total tumor burden is shown in Figure 4. Even more important than the mean $\mathrm{RC}_{\mathrm{total}}$ of -52.8%, the 95% LoA, (-382.1%, 68.7%), are again very wide and likely to suggest that a patient is responding to treatment simply due to a switch in radiotracers. The LoA estimated by incorrectly assuming the $\mathrm{RC}_{\mathrm{total}}$ follows a normal distribution, (-251.0%, 145.4%), while still indicating a large degree of variability are quite different from the correctly estimated LoA.

4. Discussion

In looking at Figures 3 and 4, one may be struck by the observation that the LoA are not symmetric around the mean. In the usual calculation of the LoA assuming normality, by definition the intervals are symmetric reflecting the symmetric bell-shaped curve of the normal distribution. In contrast, it is precisely because the RC and $\mathrm{RC}_{\mathrm{total}}$ have highly skewed distributions that a symmetric interval is not appropriate.

In the context of estimating confidence intervals for the mean of a lognormal distribution, it has been shown in simulation studies [29, 30] that the usual 95% normal confidence interval can have a very large degree of coverage error depending upon the sample size and severly underestimate the true coverage probability. While the LoA are not confidence intervals, the same lessons apply. Incorrectly applying the LoA equations based on the normality assumption to lognormal data will result in intervals that do not accurately capture where 95% of the differences will lie.

A key point to interpretting 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. In our experience as statisticians, we have frequently been asked to analyze data evaluating agreement between two anatomic or nonanatomic measurements. We can easily produce Bland-Altman plots, but when we show the plots to our radiology collaborators we find they often look to us to determine whether the level of agreement is acceptable. As Bland and Altman very poignantly point out [21], this is not a statistical question but a clinical one. Statisticians cannot answer this question in a vacuum and may be hard-pressed to provide guidance to their clinical collaborators particularly when talking about differences in tumor measurements on the absolute scale. In this situtation, we are fortunate to have a difference metric that is intimately familiar to all radiologists who image cancer patients. It is a metric that can be directly related to patient care and placed in the context of how the lack of agreement could affect patient management. We suggest that basing Bland-Altman plots and correctly constructed LoA on this metric leads to evaluating agreement on a naturally intuitive scale and helps to faciliate interpretation.

References

[1] Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**:307–317.

[2] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods for clinical measurement. *Lancet* 1986; **327**(8476):307–310.

[3] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumors: Revised recist guideline (version 1.1). *European Journal of Cancer* 2009; **45**:228–247.

[4] Miller AB, Hoogstraten B, Staquest M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981; **47**:207–214.

[5] Kelloff GJ, Hoffman JM, Johnson B, Scher HI, Siegel B, Cheng EY, Cheson BD, O'Shaughnessy J, Guyton KZ, Mankoff DA, Shankar L, Larson SM, Sigman CC, Schilsky RL, Sullivan DC. Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. *Clinical Cancer Research* 2005; **11**(8):2785–2808.

[6] Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *The Journal of Nuclear Medicine* 2009; **50**(5 (Suppl)):122S–150S.

[7] Lavin PT. An alternative model for the evaluation of antitumor activity. *Cancer Clinical Trials* 1981; **4**:451–457.

[8] Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: Application to a study of Sorafenic and Erlotinib in non-small-cell lunch cancer. *Journal of the National Cancer Institute* 2007; **99**:1455–1461.

[9] Dhani N, Tu D, Sargent DJ, Seymour L, Moore MJ. Alternate endpoints for screening phase II studies. *Clinical Cancer Research* 2009; **15**:1873–1882.

[10] Wason JMS, Mander AP, Eisen TG. Reducing sample size in two-stage phase II cancer trials by using continuous tumour shrinkage end-points. *European Journal of Cancer* 2011; **47**:983–989.

[11] Gietema HA, Schaefer-Prokop CM, Mali WPTM, Groenewegen G, Prokop M. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection ct - influence of inspiration level, nodule size, and segmentation performance. *Radiology* 2007; **245**(3):888–894.

[12] Wormanns D, Kohl G, Klotz E, Marheine A, Beyer F, Heindel W, Diederich S. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility. *European Radiology* 2004; **14**:86–92.

[13] Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetic measurements. *AJR* 2006; **186**:989–994.

[14] Sensakovi WF, Armato SG, Straus C, Roberts RY, Caligiuri P, Starkey A, Kindler HL. Computerized segmentation and measurement of malignant pleural mesothelioma. *Medical Physics* 2011; **38**(1):238–244.

[15] Bauknecht HC, Romano VC, Rogalla P, Klingbiel R, Wolf C, Bornemann L, Hamm B, Hein PA. Intra- and interobserver variability of linear and volumetric meausrements of brain metastases using contrast-enhanced magnetic resonance imaging. *Investigative Radiology* 2010; **45**(1):49–56.

[16] Dubus L, Gayet M, Zappa M, Abaleo L, De Cooman A, Orieux G, Vilgrain V. Comparison of semi-automated and manual methods to measure the volume of liver tumours on MDCT images. *European Radiology* 2011; **21**:996–1003.

[17] Rominger MB, Fournell D, Nadar BT, Behrens SNM, Figiel JH, Heverhagen JT. Comparison of semi-automated and manual methods to measure the volume of liver tumours on MDCT images. *European Radiology* 2009; **19**:1097–1107.

[18] Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, Qin Y, Riely GJ, Kris LG, Schwartz LH. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* 2009; **252**(1):263–272.

[19] Oxnard GR, Zhao B, Sima CS, Ginsberg M, James LP, Lefkowitz RA, Gua P, Kris MG, Schwartz LH, Riely GJ. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *Journal of Clinical Oncology* 2011; **29**(23):3114–3119.

[20] Bland JM, Altman DG. Agreement between methods of measurements with multiple observations per indvidual. *Journal of Biopharmaceutical Statistics* 2007; **17**:571–582.

[21] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999.

[22] Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia* 2007; **99**(3):309–311.

[23] Borovkova S, Permana F. A closed form approach to valuing and hedging basket options. In *Computing in Economics and Finance 2006* number 54. Society for Computation Economics 2006.

[24] Johnson NL, Kotz S. *Continuous Univariate Distributions*. John Wiley and Sons: New York first edition edition, 1970.

[25] Fenton LF. The sum of log-normal probability distibutions in scattered transmission systems. *IRE Trans. Commun. Systems* 1960; **8**:57–67.

[26] Abu-Dayya AA, Beaulieu NC. Outage probabilities in the presence of correlated lognormal interferers. *IEEE Transactions on Vehicular Technology* 1994; **1**(326):164–173.

[27] Ligeti A. Outage probabilities in the presence of correlated lognormal useful and interfering components. *IEEE Communication Letters* 2000; **4**(1):15–17.

[28] Fox JJ, Autran-Blanc E, Morris MJ, Gavane S, Nehmeh S, Van Nuffel A, Gönen M, Schöder H, Humm JL, Scher HI, Larson SM. Practical approach for comparative analysis of multi-lesion molecular imaging using a semi-automated program for PET/CT. *Journal of Nuclear Medicine* in press.

[29] Zhou XH, Gao S. Confidence intervals for the log-normal mean. *Statistics in Medicine* 1997.

[30] Olsson U. Confidence intervals for the mean of a log-normal distribution. *Journal of Statistics Education* 2005.

5. Figure Legends

Figure 1. Histograms of $\mathrm{SUV}_{\mathrm{max}}$ measurements for FDG-PET and FDHT-PET

Figure 2. Histograms showing the distribution of the relative change between $\mathrm{SUV}_{\mathrm{max}}$ measurements for FDG-PET and FDHT-PET at (a) the lesion level in terms of RC and (b) the patient level in terms of $\mathrm{RC}_{\mathrm{total}}$

Figure 3. Bland-Altman plot evaluating the agreement between $\mathrm{SUV}_{\mathrm{max}}$ measurements from FDG-PET and FDHT-PET for individual lesions. The dotted line shows the mean RC of -77.0% between the two measurements. The black dashed lines show the 95% LoA, (-485.3%, 69.6%), calculated assuming the RC follows a negative shifted lognormal distribution. The grey dashed-dotted lines demonstrate estimated 95% LoA incorrectly calculated assuming the RC is normally distributed.

Figure 4. Bland-Altman plot evaluating the agreement between $\mathrm{SUV}_{\mathrm{max}}$ measurements from FDG-PET and FDHT-PET for total patient tumor burden. The dotted line shows the mean $\mathrm{RC}_{\mathrm{total}}$ of -52.8% between the two measurements. The black dashed lines show the 95% LoA, (-382.1%, 68.7%), calculated assuming the RC_{total} follows a negative shifted lognormal distribution. The grey dashed-dotted lines demonstrate estimated 95% LoA incorrectly calculated assuming the $\mathrm{RC}_{\mathrm{total}}$ is normally distributed.

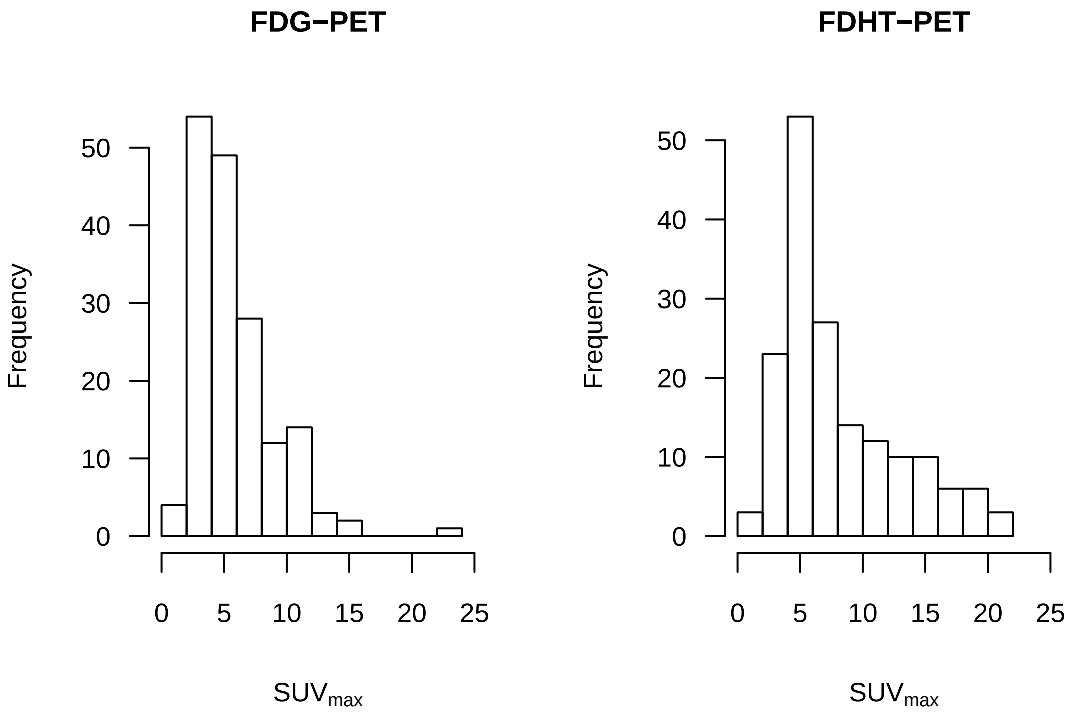

FDG−PET
FDHT−PET
Frequency
0
10
20
30
40
50
0
5
10
15
20
25
SUV_{max}
Frequency
0
10
20
30
40
50
0
5
10
15
20
25
SUV_{max}

Figure 1

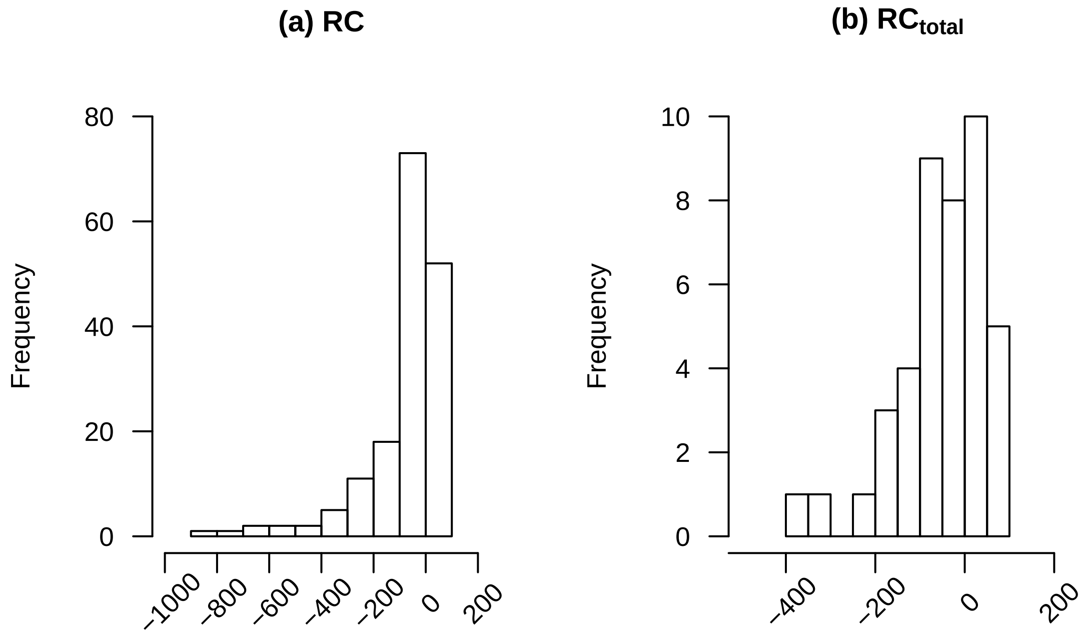
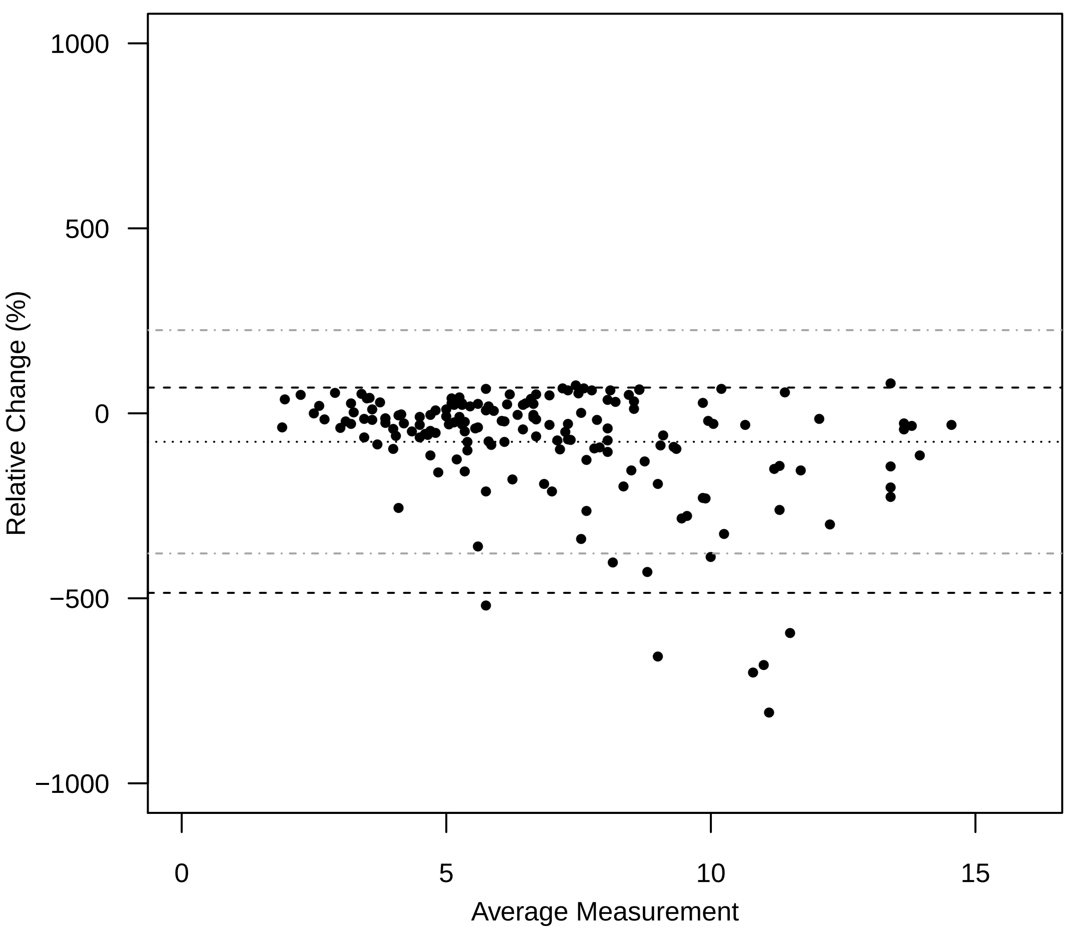

(a) RC
(b) RC_{total}
Frequency
80
60
40
20
0
–1000
–800
–600
–400
–200
0
200
Frequency
10
8
6
4
2
0
–400
–200
0
200

Figure 2



1000
500
0
−500
−1000
Relative Change (%)
0
5
10
15
Average Measurement

Figure 3

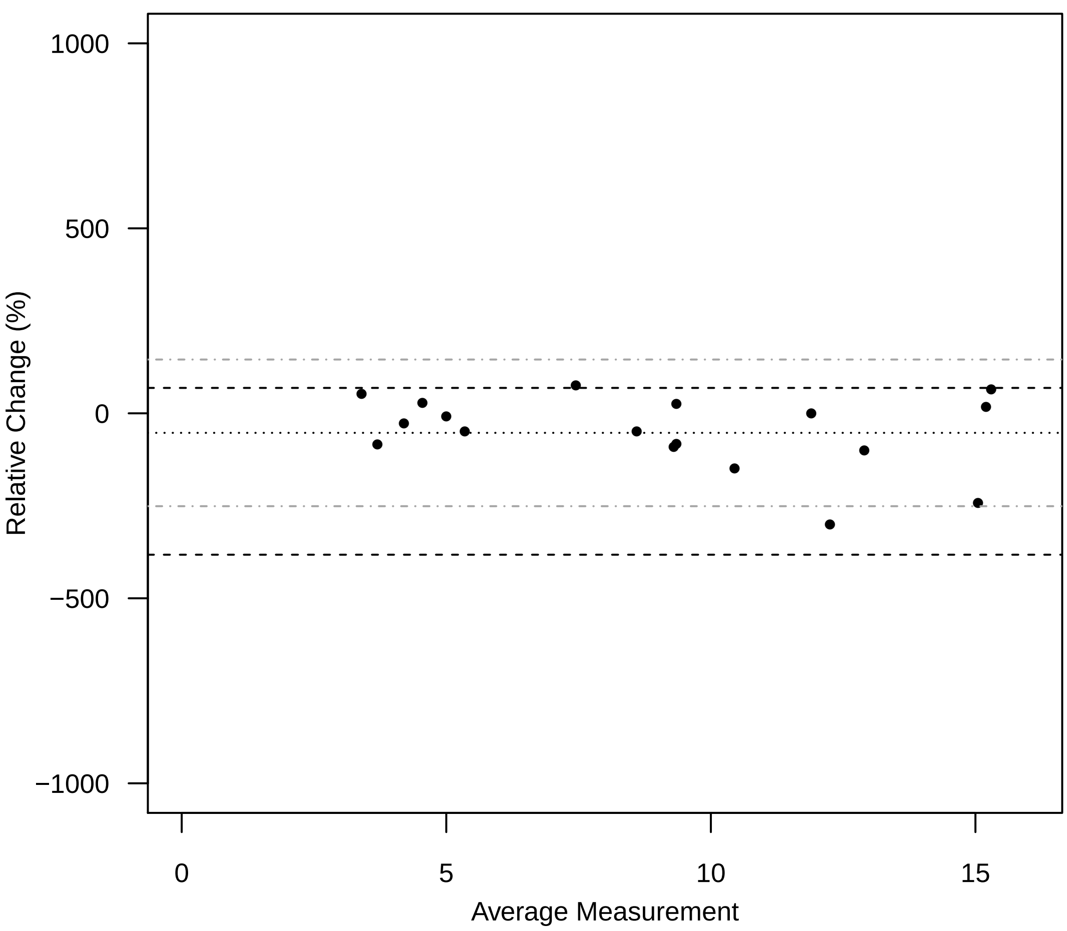

1000
500
0
−500
−1000
Relative Change (%)
0
5
10
15
Average Measurement

Figure 4