PLOS ONE

RESEARCH ARTICLE

Reproducibility and repeatability of ^{18}F-(*2S*, *4R*)-4-fluoroglutamine PET imaging in preclinical oncology models

Gregory D. Ayers[1,2], Allison S. Cohen[3,4,5¤], Seong-Woo Bae[5], Xiaoxia Wen[5], Alyssa Pollard[5], Shilpa Sharma[5], Trey Claus[3,4], Adria Payne[3,4], Ling Geng[3,4†], Ping Zhao[3,4], Mohammed Noor Tantawy[4,6], Seth T. Gammon[5], H. Charles Manning[2,3,4,5,6]*

1 Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States of America, **2** Vanderbilt Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, United States of America, **3** Vanderbilt Center for Molecular Probes, Vanderbilt University Medical Center, Nashville, TN, United States of America, **4** Vanderbilt University Institute of Imaging Science, Vanderbilt University Medical Center, Nashville, TN, United States of America, **5** Department of Cancer Systems Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX, United States of America, **6** Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Medical Center North, Nashville, TN, United States of America

† Deceased.
¤ Current address: Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, United States of America
* hcmanning@mdanderson.org

Abstract

Introduction

Measurement of repeatability and reproducibility (R&R) is necessary to realize the full potential of positron emission tomography (PET). Several studies have evaluated the reproducibility of PET using ^{18}F-FDG, the most common PET tracer used in oncology, but similar studies using other PET tracers are scarce. Even fewer assess agreement and R&R with statistical methods designed explicitly for the task. ^{18}F-(*2S*, *4R*)-4-fluoro-glutamine (^{18}F-Gln) is a PET tracer designed for imaging glutamine uptake and metabolism. This study illustrates high reproducibility and repeatability with ^{18}F-Gln for *in vivo* research.

Methods

Twenty mice bearing colorectal cancer cell line xenografts were injected with ~9 MBq of ^{18}F-Gln and imaged in an Inveon microPET. Three individuals analyzed the tumor uptake of ^{18}F-Gln using the same set of images, the same image analysis software, and the same analysis method. Scans were randomly re-ordered for a second repeatability measurement 6 months later. Statistical analyses were performed using the methods of Bland and Altman (B&A), Gauge Reproducibility and Repeatability (Gauge R&R), and Lin's Concordance Correlation Coefficient. A comprehensive equivalency test, designed to reject a null hypothesis of non-equivalence, was also conducted.

OPEN ACCESS

Citation: Ayers GD, Cohen AS, Bae S-W, Wen X, Pollard A, Sharma S, et al. (2025) Reproducibility and repeatability of ^{18}F-(*2S*, *4R*)-4-fluoroglutamine PET imaging in preclinical oncology models. PLoS ONE 20(1): e0313123. https://doi.org/10.1371/journal.pone.0313123

Editor: Bing Xu, Brandeis University, UNITED STATES OF AMERICA

Received: August 28, 2024

Accepted: October 19, 2024

Published: January 9, 2025

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants from the NIH (P50 CA236733, U24 CA220325, S10 OD019963, and S10 OD016245) and CPRIT (RR200046). HCM is a Cancer Prevention Research Institute of Texas (CPRIT) Scholar in Cancer Research (CPRIT RR200046). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Results

In a two-way random effects Gauge R&R model, variance among mice and their measurement variance were 0.5717 and 0.024. Reproducibility and repeatability accounted for 31% and 69% of the total measurement error, respectively. B&A repeatability coefficients for analysts 1, 2, and 3 were 0.16, 0.35, and 0.49. One-half B&A agreement limits between analysts 1 and 2, 1 and 3, and 2 and 3 were 0.27, 0.47, and 0.47, respectively. The mean square deviation and total deviation index were lowest for analysts 1 and 2, while coverage probabilities and coefficients of the individual agreement were highest. Finally, the definitive agreement inference hypothesis test for equivalency demonstrated that all three confidence intervals for the average difference of means from repeated measures lie within our *a priori* limits of equivalence (i.e. ± 0.5%ID/g).

Conclusions

Our data indicate high individual analyst and laboratory-level reproducibility and repeatability. The assessment of R&R using the appropriate methods is critical and should be adopted by the broader imaging community.

Introduction

Imaging is useful for lesion detection, staging, and evaluation of treatment response and disease progression. The sensitive and quantitative nature of PET, coupled with the ability to produce targeted PET tracers, renders PET uniquely capable of detecting tumors and profiling their specific features. Importantly, PET provides a functional measure of tumor phenotype non-invasively *in vivo*, which allows for a quantitative assay of biological processes, such as the activity of transporters and enzymes. However, to realize the full potential of PET imaging, imaging must demonstrate agreement, which is dependent upon the measurement of repeatability and reproducibility (R&R) [1, 2].

Glutaminolysis is vital to tumor growth, progression, and survival [3–13]. To meet their demand for glutamine, tumor cells transport glutamine into the cell from the tumor microenvironment [3–9, 11–14]. Glutamine can then be used in various downstream processes, including the synthesis of proteins, nucleic acids, and hexosamines, or conversion to glutamate, which can then be used as a source of glutathione, α-ketoglutarate, or nonessential amino acids [3–11]. To study the uptake and metabolism of glutamine, syntheses of ^{18}F-labeled glutamine analogues for PET imaging have been reported [15–18]. ^{18}F-Gln has been studied preclinically [19–30] and clinically, including in brain, pancreas, breast, lung, and thyroid cancers [21, 26, 31–34]. Recent studies evaluated the reproducibility of ^{18}F-FDG PET imaging in phantoms [35] and preclinical models [36–38]. Similar studies using other PET tracers are scarce, and even fewer assess agreement and R&R with statistical methods designed explicitly for the task. These studies are needed for new tracers to support basic science research and, more importantly, are necessary for the clinical translation of these tracers and, ultimately, adoption of these methods as part of standard-of-care imaging. Here, we assess ^{18}F-Gln using data from an experiment explicitly designed to evaluate user agreement, reproducibility, and repeatability.

Materials and methods

Cell lines

HCT-116 cell lines were purchased from ATCC (American Type Culture Collection) and authenticated using a commercial vendor (Genetica). Cells were cultured in Dulbecco's

Modified Eagle Medium (DMEM) containing 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin (p/s). The cells were incubated in 5% CO_2 at 37˚C.

Animal models

All animal procedures complied with the Guide for the Care and Use of Laboratory Animal Resources (1996) and National Research Council and were approved by the Vanderbilt University Institutional Animal Care and Use Committee (Nashville, TN, USA). Animals were purchased from Envigo and used in accordance with Institutional and Federal guidelines. Female athymic nude mice (Hsd: Athymic Nude-*Foxn1^{nu}*, Envigo, #6903), 5-6-weeks old, were injected subcutaneously into the right flank with 8 x 10^6 HCT-116 cells. Twenty mice were used in this study. Mice were monitored daily and tumor size and body weight were measured three times per week. The tumor volume was calculated according to the formula WxLxH/2. Imaging was performed when the tumor volume reached ~250 mm^3 at days 19–22 post-tumor cell injection. Animals were anesthetized with 2% isofluorane prior to tracer administration and were kept warm using in their cages using circulating water bath until imaging. All efforts were made to minimize suffering, mice were kept warm during PET imaging via circulated heated water. None of the mice reached humane endpoints (tumor size greater than 1.5 cm in average diameter, body weight loss more than 20%, or state of moribundity) before completion of the experiment. At the end of the experiment, mice anesthetized by isoflurane gas were euthanized by carbon dioxide asphyxiation followed by cervical dislocation. Prior to asphyxiation and subsequent cervical dislocation, mice were palpated to ensure deep anesthesia and prevent suffering.

Radiochemistry

[^{18}F]-(*2S*, *4R*)-4-fluoro-glutamine was synthesized as previously described by our group using methodologies identical to those reported [15, 22].

PET imaging

PET imaging experiments were performed using HCT-116 tumor-bearing mice. Two sets of 10 mice each were imaged on consecutive days. Imaging conditions were kept as consistent as possible between days. Access to food and water was provided ad libitum. Animals were anesthetized with 2% isofluorane and administered 8.2–11.4 MBq of ^{18}F-Gln via retroorbital injection by highly trained personnel and as approved by the Vanderbilt University Institutional Animal Care and Use Committee (Protocol Numbers M1800041 and M1500003). The mice were returned to their cages and kept warm via a circulating water bath. Following 40 minutes of tracer uptake, mice were anesthetized with 2% isoflurane and static images were acquired for 20 minutes in an Inveon microPET (Siemens Preclinical Solutions). One mouse was accidentally imaged using the wrong imaging scanner protocol initially. This mouse was reimaged using the correct imaging scanner protocol (20-minute static images) at a later time point. Thus, the PET images for this mouse were acquired at a radiotracer uptake time of more than 2.5 hours post-injection instead of 40 minutes post-injection. The images for this mouse using the correct imaging scanner protocol at the later time point were included in the image analysis and statistical comparisons.

Image analysis

All data sets were reconstructed using the three-dimensional (3D) ordered subset expectation maximization/maximum a posteriori (OSEM3D/MAP) algorithm into 128 x 128 x 95 slices

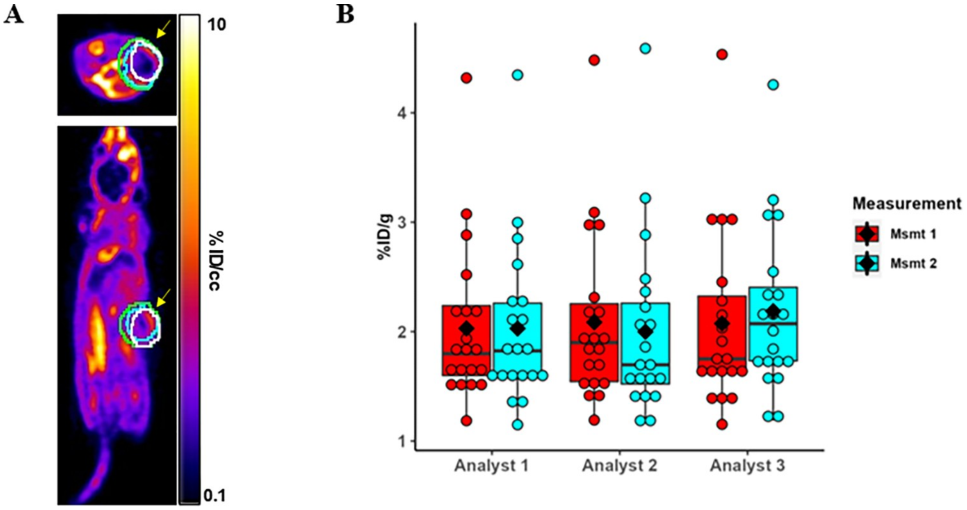

Fig 1. A) Representative axial and coronal images of %ID/g using [^{18}F]-(*2S*, *4R*)-4-fluoro-glutamine. Yellow arrows indicate tumor location. White (analyst 1), cyan (analyst 2), and green (analyst 3) volumes of interest (VOIs) on the tumor were captured by the 3D algorithm of auto iso-contour that each analyst drew. **B)** Dot/boxplot of %ID/g for twice replicated data of 20 mice among 3 analysts. Diamonds represent the mean value and horizontal bars the median. Bottom and tops of boxes represent the 25th and 75th percentiles. Vertical lines extending above and below the boxes are 1.5*the interquartile (IQR = 75th– 25th percentile) range. The data points that are greater than 4%ID/g are all from the same mouse. This mouse was imaged using a different protocol however despite this variation in protocol, repeat measurements by the three analysts result in similar uptake values.

https://doi.org/10.1371/journal.pone.0313123.g001

with a voxel size of 0.095 x 0.095 x 0.08 cm^3 at a beta value of 0.01. The PET images were loaded onto the image analysis tool Amide (www.souceforge.net) to export the Digital Imaging and Communications in Medicine (DICOM) format with the percentage of the injected dose per gram of tissue (%ID/g) unit. PMOD software (PMOD Technologies LLC, Zurich, Switzerland) was used to draw 3D volumes of interest (VOIs) around the tumors on the right flank and around the muscle on the contralateral left flank in the PET images. VOIs were captured by using the intrinsic 3D algorithm of the auto iso-contour tool based on spheres (analytic objects) of a certain size on the software. The measured counts were converted to the percentage of the injected dose per gram of tissue (%ID/g). Tumor/muscle ratio was calculated as the tumor %ID/g value divided by the muscle %ID/g value.

Three people analyzed the OSEM3D/MAP datasets from the HCT-116 test-retest study using the same image analysis software (PMOD) to evaluate the effect of user R&R on imaging results (Fig 1). Scans were randomly re-ordered for a second, repeatability measurement six months later. One person has more than 5 years of experience in analyzing preclinical PET data. The other two people have less than five years of experience each.

Statistical methods

Based on our reading, the methods of Bland and Altman, the intra class correlation coefficient (ICC), and Lin's concordance correlation coefficient (CCC) are the most cited methods for agreement. Gauge reproducibility and repeatability (Gauge R&R) and other useful methods based on mixed and random effect models are rare in the imaging literature. A definitive test for equivalency could not be found in the imaging literature. We provide, for the first time to our knowledge, direct comparison of these various metrics for imagers.

Bland and Altman (B&A) limits of agreement (LOA) for reproducibility and repeatability coefficients (RC) were estimated and graphically presented based on their landmark methods for estimands of R&R in medical applications [39–41] for average (over repeated measurements) and single measurements among reviewers. LOA is defined as LOA = $\bar{d} \pm 1.96*S_d$ where $\bar{d}$ and S_d are the average difference and standard deviation of the differences between reviewers. Repeatability coefficient (RC) is calculated according to the equation B&A RC = $1.96^{*}\sqrt{2}^{*}S_w$ where S_w is the within-subject standard deviation of the replicates from the same reviewer. The CCC has reached the imaging literature with CCC greater than 0.8 or 0.9 considered excellent [42–48]. Consequently, statistical tests comparing the CCC to zero concordance seem unwarranted (S1 File).

Gauge Reproducibility and Repeatability (Gauge R&R)

We also used Gauge Reproducibility and Repeatability (Gauge R&R) methods to estimate the capability of our measurement system as a whole for ^{18}F-Gln PET imaging [49, 50]. Using the 2-way crossed random effects model: $Y_{ijk} = \mu + M_i + O_j + (MO)_{ij} + E_{ijk}$, where i = 1, . . . 20 mice, j = 1, . . .,3 analysts, and k = 1, . . .,2 repeated measures we estimated σ^2_M, σ^2_O, σ^2_{MO}, and σ^2_E, respectively and define Repeatability = σ^2_E, reproducibility = $\sigma^2_O + \sigma^2_{MO}$, and the total variability of the measurement procedure = $\sigma^2_O + \sigma^2_{MO} + \sigma^2_E$. From these estimates arise several useful parameters [49] which can be extracted from the mean square estimates of random and mixed models in the EMSaov package or lme4 package found in the R software system [51–55]. See supplementary materials for more detail. A definitive comparison of agreement between reviewers was made using a test of equivalence. Novel to this paper, $\bar{d} \pm t_{df=N-1,\alpha=0.05/2}*S_d/\sqrt{N}$, is the 95% confidence interval for the average mean difference between two observers. A 95% confidence interval that lies within predetermined ±δ is equivalent to rejecting the null hypothesis ($p<0.05$) of non-equivalence, H_0: $\mu \leq$ -δ or $\mu \geq$ δ and declaring equivalence between observers. With 3 reviewers, we estimated a set of 3, 98.3% (1–0.05/3) confidence intervals, to control the experiment-wise error rate using a Bonferroni correction at 5% (S1 File).

Results

Summary of imaging measures

Tumor PET imaging data from 20 mice were analyzed by three different people at our institution, with replicate observations performed 6 months later, using the same set of images (OSEM3D/MAP reconstructed data), the same software (PMOD), and the same analysis method. Results are summarized in Table 1 (%ID/g) and S1 Table (tumor/muscle ratio) and depicted in Fig 1. Overall, mixed model-based estimates (95% CI) which incorporate repeated measures correlation, were 2.07 (1.71 to 2.42) and 1.13 (1.03 to 1.23), respectively, for %ID/g and tumor/muscle ratios. Raw data based average (SD) %ID/g ranged from 2.00 to 2.19 across all analysts and measurements, while the standard deviation ranged from 0.72 to 0.82. Coefficients of variation (CV) ranged from 0.34 to 0.41. Analyst 1 had the most consistent average

Table 1. Summary statistics of %ID/g by analyst and measurement.

Statistic	Analyst 1		Analyst 2		Analyst 3	
	Msmt 1*	Msmt 2	Msmt 1	Msmt 2	Msmt 1	Msmt 2
Mean (SD)	2.03 (0.72)	2.03 (0.74)	2.09 (0.78)	2.00 (0.82)	2.07 (0.81)	2.19 (0.75)
Median (range)	1.80 (1.19, 4.32)	1.82 (1.15, 4.35)	1.90 (1.19, 4.48)	1.70 (1.17, 4.59)	1.75 (1.15, 4.53)	2.07 (1.19, 4.26)

*Msmt = measurement

https://doi.org/10.1371/journal.pone.0313123.t001

Table 2. Results of Gauge reproducibility and repeatability on repeated measures data set.

Model Parameter	Analyst as a Fixed Effect	Analyst as a Random Effect
Mouse Variance	0.5716694	0.5716694
Analyst Variance	0.0015801	0.0023701
Interaction Variance	0.0052416	0.0052416
Error Variance	0.0166434	0.0166434
Mouse Variance	0.572 (0.143, 1.108)	0.572 (0.147, 1.114)
Measurement Variance	0.023 (0.013, 0.036)	0.024 (0.014, 0.037)
Mouse to Measurement Ratio	24.363 (5.530, 60.416)	23.569 (5.607, 58.479)
Repeatability Proportion	0.709 (0.493, 1)	0.686 (0.491, 0.969)
Reproducibility Proportion	0.291 (0, 0.507)	0.314 (0.031, 0.509)
Intraclass Correlation Coefficient	0.961 (0.847, 0.984)	0.959 (0.849, 0.983)

https://doi.org/10.1371/journal.pone.0313123.t002

and median values, the lowest standard deviations, and the lowest CV across their respective measurements. Fig 1B is a dot/boxplot that depicts the data for %ID/g.

Estimation of PET imaging system capability via Gauge R&R

Generally speaking, repeatability is a characteristic of dependent measurements by the same analyst, for the same tissue, or for the same mice. Reproducibility is a characteristic of independent entities as with analysts, operators, and devices.

Overall system reproducibility and repeatability metrics assuming a two-way random effects Gauge R&R model are presented in Table 2 (S2 Table for first measurement data). Variability among mouse tumors (mouse variance; (σ^2_M)) was 0.5717, analyst variance (σ^2_O) was 0.0024, process by imaging interaction (σ^2_{MO}) was 0.0052, and repeatability (σ^2_E) was 0.0166 for a total measurement error ($\sigma^2_O + \sigma^2_{MO} + \sigma^2_E$) of 0.0242. The animal variance was 24 times greater than measurement variation, indicating that biological variability is the main source of differences in PET uptake values. Reproducibility ($\sigma^2_O + \sigma^2_{MO}$) and repeatability accounted for 31% and 69% of the total measurement error, respectively. The intra-class correlation coefficient (ICC) is the proportion of total variation due to mice and represents the correlation between two measurements taken on the same mouse; the ICC in this study was 0.96. Confidence intervals are the 2.5 and 97.5 percentiles from bootstrap sampling based on 10,000 replicates.

Comparing analysts

The repeatability coefficients for analysts 1, 2, and 3 were 0.16, 0.35, and 0.49 (Table 3). Next, we compared data obtained between two analysts (Table 4 and S3 Table). One-half B&A LOA between analysts 1 and 2, 1 and 3, and 2 and 3 were 0.27%ID/g, 0.47%ID/g, and 0.47%ID/g,

Table 3. Bland-Altman repeatability index (RI–per analyst).

Analyst	Repeatability Index*
1	0.155
2	0.346
3	0.490

RI = 1.96 $\sqrt{2}$*S$_w$, where S$_w$ is the within-mouse (subject) standard deviation from repeated measures within analyst.

https://doi.org/10.1371/journal.pone.0313123.t003

Table 4. Bland-Altman Limits of Agreement (LOA—repeated observations model).

Analysts	1/2 Agreement Limit*	Equivalency Lower Confidence Limit**	Bias***	Equivalency Upper Confidence Limit**
1 vs 2	0.267	-0.078	-0.013	0.051
1 vs 3	0.465	-0.213	-0.101	0.01
2 vs 3	0.465	-0.213	-0.088	0.01

*LOA = 1.96*S_d, where S_d is the standard deviation of paired differences adjusted for repeated measures. **95% confidence intervals using t value with 19 degrees of freedom. ***Average difference of paired measurements.

https://doi.org/10.1371/journal.pone.0313123.t004

respectively. Bias between analyst pairs, were -0.013%ID/g, -0.101%ID/g, and -0.088%ID/g, respectively. B&A plots depict the ranges where 95% of differences between analysts will lie (Fig 2). Table 5 shows the agreement metrics based on a two-way mixed model with analysts as a fixed effect. The mean square deviation (MSD) and total deviation index (TDI) were lowest for analysts 1 and 2, while coverage probability (CP) and coefficient of individual agreement (CIA) were higher. The CIA was markedly different, reflecting the combined differences in S_w and bias between analysts. Analysts 1 and 2 exhibited the highest CIA, with lower average S_w and bias. Along with the ICC, analysts 1 and 3 had the worst CIA followed by analysts 2 and 3. Interestingly, analysts 2 and 3 had less bias and more narrow agreement limits on their first measurement than analyst 1 versus the other two analysts (S1–S3 Figs, S3 Table). The lower limits of the 95% confidence intervals for the CCC exceed 0.9 (S4 Table).

We tested the null hypothesis of non-equivalence with an experiment-wise error rate of 5% based on the set of Bonferroni corrected confidence intervals for the mean difference among

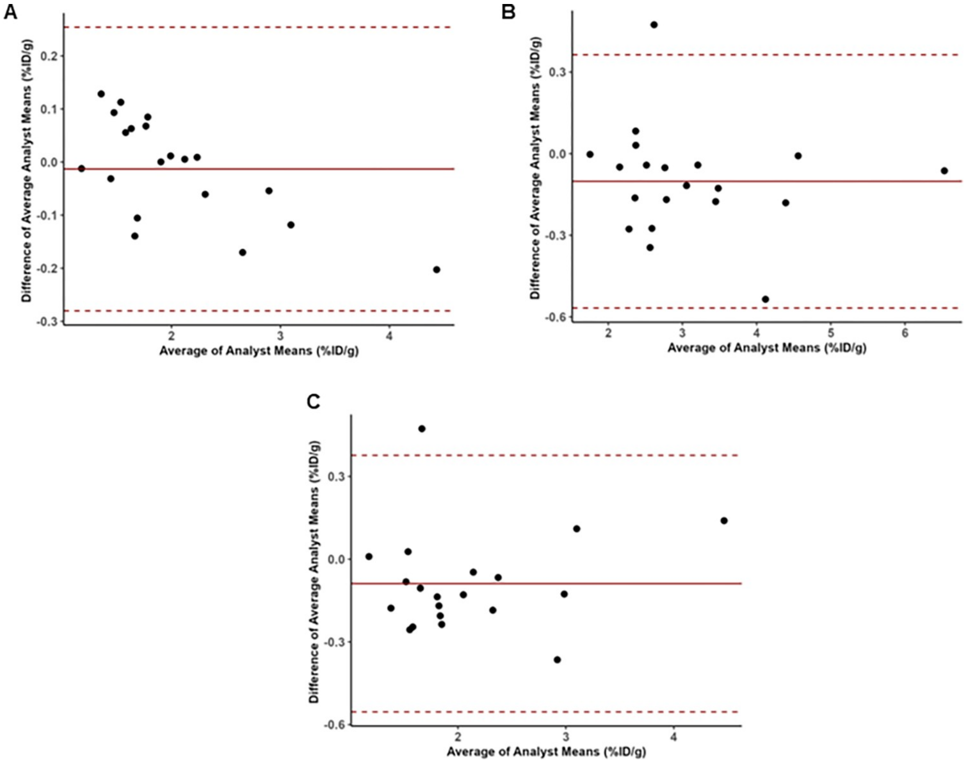

Fig 2. Bland-Altman plot comparing **A)** analysts 1 and 2, **B)** analysts 1 and 3, and **C)** analysts 2 and 3 from repeated measures (2 replicates) design.

https://doi.org/10.1371/journal.pone.0313123.g002

Table 5. Agreement metrics based on pairwise analyst mixed models*.

Metric	1 vs 2	1 vs 3	2 vs 3
1/2 Agreement Limits	0.267	0.465	0.465
Intraclass Correlation Coefficient	0.984 (0.934, 0.992)	0.934 (0.757, 0.974)	0.943 (0.782, 0.976)
Mean Square Deviation	0.019 (0.1, 0.32)	0.067 (0.035, 0.107)	0.064 (0.039, 0.096)
Total Deviation Index**	0.269 (0.193, 0.351)	0.506 (0.369, 0.640)	0.497 (0.388, 0.608)
Coverage Probability***	0.999 (0.955, 0.999)	0.947 (0.874, 0.992)	0.952 (0.893, 0.988)
Coefficient of Individual Agreement**	0.991 (0.597,0.999)	0.516 (0.282, 0.830)	0.729 (0.484, 0.883)

*Two-way mixed model (analyst as fixed effect). **Using mean square deviation with p = 0.95. ***Tolerance limits ± 0.5%ID/g. 95% bootstrap confidence intervals in parentheses from 10,000 replications.

https://doi.org/10.1371/journal.pone.0313123.t005

analyst pairs (Fig 3, S4 Fig and S4 Table). All three confidence limits lie within our *a priori* limits of equivalence (i.e., ± 0.5%ID/g).

Discussion

Limited R&R have been assessed by other labs for PET and MRI imaging. Savaikar et al. reported B&A limits of agreement for SUVmax (0.44) in ^{18}F-FDG-PET was approximately 3-fold higher than that for SUVmean (0.15), suggesting poor reproducibility for SUVmax, confirming the well-known phenomenon that variability among extreme (maximum,

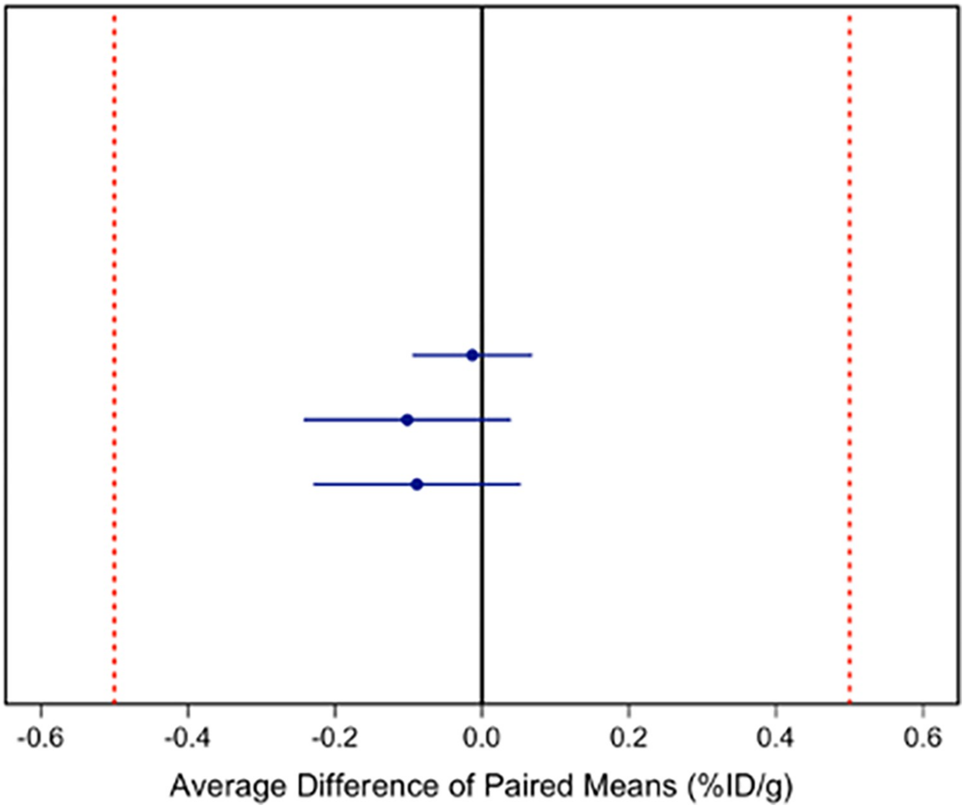

Fig 3. Hypothesis test to reject non-equivalency using confidence intervals for average difference of means from repeated measures (2 replicates) design. Shown are Bonferroni adjusted 98.3% (1–0.05/3) confidence intervals to control the experiment-wise type I error rate at 5%.

https://doi.org/10.1371/journal.pone.0313123.g003 minimum) values is greater than average values [38]. Whisenant et. al. used B&A methods to estimate repeatability of several imaging metrics in a murine model of HER2+ breast cancer [56, 57]. Limits of agreement were presented but no discussion of the adequacy of these limits or direct comparisons to other studies were made, likely due to the sparsity of such information in the literature.

We question a common hypothesis testing scenario for agreement. The first is testing for reproducibility using a null hypothesis of no difference between analysts, or devices, with the intent of declaring reproducibility following a $p > 0.05$ in comparative or pre-post comparisons. Such a comparison does not test for agreement. We cannot conclude, after failing to reject a null hypothesis, that the null hypothesis is true. We tested a non-equivalence null hypothesis for the average difference between analysts. We rejected the null hypothesis that bias (the average difference) in %ID/g among our analysts are greater than |0.5% ID/g| and conclude our analysts can be used interchangeably for the measurement of ^{18}F-Gln. Another weak approach for testing agreement measures (e.g., CCC, Kappa, ICC) is to test a null hypothesis of zero agreement. Statistically significant but unimportant agreement can be concluded simply by using a sufficiently large sample size alone if desired. The research community ultimately determines minimum values of agreement and tolerance. A meaningful statistical test for an agreement metric could reject a null hypothesis of |0.8| or lower, for example, among agreement metrics ranging from -1 to 1 or 0 to 1, respectively. For %ID/g in the first measurement analysis, the lower 95% CI for the CCC exceeded 0.9. For this reason, assessment of the adequacy of estimates of agreement metrics should be accompanied by confidence limits.

Pairwise comparisons of analysts, devices, and other factors implies the fixed effect setting since the interest lies in differences between specific devices or readers. Broadening inferences to exchangeability of devices, analysts, or reconstruction methods for an entire lab or across labs changes the indication to the use of the random effects model for estimation. Lab quality improvement benefits from pairwise comparisons. Analyst 1 elicited a markedly lower repeatability coefficient, narrower agreement limits, mean square deviation, and TDI with higher ICC, CP, and CIA. Technique improvement training following this highly reproducible user's methodology could be a source of training for current and future analysts in a lab.

Poor repeatability limits reproducibility between analysts and by extension, a lab's reliance on exchangeability when reporting experimental results in the literature. Measurement variability without replication per Gauge R&R was 0.013 (S2 Table) compared to 0.024 for the same analysis on replicated data (Table 2). Corresponding agreement limits for differences were lower for 2 of the 3 analysts accompanied by a different ordering of reproducibility: analysts 1, 2, 3 vs 3, 1, and 2, respectively. Clearly, an accurate assessment of reproducibility requires an assessment of repeatability.

Conclusion

While the panel of agreement metrics available is now extensive and we have added non-equivalence testing and Gauge R&R metrics to the evaluation set, the adoption of reproducibility and repeatability assessment in the imaging literature remains low. A subset of these metrics and methods may be sufficient by assessment of their intent and characteristics [55]. We suggest that correlation-based measures and statistical tests against null hypotheses of zero difference have lower utility. As agreement assessment suffuses the imaging research community, limits of agreement, equivalence confidence intervals, and confidence intervals of agreement statistics provide critical descriptive measures for improving reproducibility of experimental results for basic and clinical imaging. Research presented that cannot demonstrate adequate laboratory reproducibility and repeatability may be considered insufficiently rigorous. We defined tolerance limits for *in vivo* imaging of ^{18}F-Gln, based on experience prior to analysis, as ± 0.5% ID/g. Multiplicity-adjusted difference confidence intervals were well within tolerance limits. We conclude that we have high individual analyst and laboratory-level reproducibility and repeatability. Expanding the inference to the population of investigators in the research community at large was adjusted for by larger variability in the mixed models that treated investigators as random effects. Per usual, precision of such estimates improves to a limit with greater sample size. Such inference, however, would logically and statistically improve in an investigation of randomly sampled investigators across labs and institutions. Given the apparent plethora of metrics available, an equivalence test, reproducibility (e.g., biological variance to measurement variance ratio—ICC), and repeatability (e.g. ICC and coefficients of individual agreement), with 95% confidence intervals and supporting graphics, should be considered.

Supporting information

S1 File. Statistical methods for reproducibility and repeatability.
(DOCX)

S1 Fig. Bland-Altman plot comparing analysts 1 and 2 from first measurement data.
(TIF)

S2 Fig. Bland-Altman plot comparing analysts 1 and 3 from first measurement data.
(TIF)

S3 Fig. Bland-Altman plot comparing analysts 2 and 3 from first measurement data.
(TIF)

S4 Fig. Hypothesis test to reject non-equivalency using confidence intervals for mean difference from repeated measures (2 replicates) design. Shown are Bonferroni adjusted 98.3% (1–0.05/3) confidence intervals to control the experiment-wise type I error rate at 5%.
(TIF)

S1 Table. Summary statistics of tumor/muscle ratio by analyst and measurement.
(DOCX)

S2 Table. Results of Gauge reproducibility and repeatability on first measurement data set.
(DOCX)

S3 Table. Bland-Altman Limits of Agreement (LOA) for first measurement.
(DOCX)

S4 Table. Concordance correlation coefficient (95% confidence interval) for first measurement between analysts.
(DOCX)

S1 Graphical abstract.
(TIF)

Acknowledgments

The authors thank Karen B. Cotton-Jones and Dimitra K. Georgiou for critically assessing and editing the manuscript.

Author Contributions

Conceptualization: Gregory D. Ayers, Allison S. Cohen, H. Charles Manning.

Data curation: Seong-Woo Bae, Xiaoxia Wen, Alyssa Pollard, Trey Claus, Adria Payne, Seth T. Gammon.

Formal analysis: Seong-Woo Bae, Xiaoxia Wen, Alyssa Pollard, Trey Claus, Adria Payne, Mohammed Noor Tantawy, Seth T. Gammon.

Funding acquisition: H. Charles Manning.

Investigation: Gregory D. Ayers, Allison S. Cohen, Seong-Woo Bae, Xiaoxia Wen, Alyssa Pollard, Trey Claus, Adria Payne, Ling Geng, Ping Zhao, Mohammed Noor Tantawy, Seth T. Gammon, H. Charles Manning.

Methodology: Gregory D. Ayers, Mohammed Noor Tantawy, Seth T. Gammon, H. Charles Manning.

Project administration: H. Charles Manning.

Resources: H. Charles Manning.

Supervision: Gregory D. Ayers, H. Charles Manning.

Validation: Gregory D. Ayers, Allison S. Cohen, Seong-Woo Bae, Xiaoxia Wen, Alyssa Pollard, Shilpa Sharma, Trey Claus, Adria Payne, Ling Geng, Ping Zhao, Mohammed Noor Tantawy, Seth T. Gammon, H. Charles Manning.

Visualization: Allison S. Cohen.

Writing – original draft: Gregory D. Ayers, Allison S. Cohen, H. Charles Manning.

Writing – review & editing: Gregory D. Ayers, Shilpa Sharma, H. Charles Manning.

References

1. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet*. 1986; 1:307–310. PMID: 2868172
2. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999; 8:135–160. https://doi.org/10.1177/096228029900800204 PMID: 10501650
3. DeBerardinis RJ, Cheng T. Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. *Oncogene*. 2010; 29:313–324. https://doi.org/10.1038/onc.2009.358 PMID: 19881548
4. Rajagopalan KN, DeBerardinis RJ. Role of glutamine in cancer: therapeutic and imaging implications. *J Nucl Med*. 2011; 52:1005–1008. https://doi.org/10.2967/jnumed.110.084244 PMID: 21680688
5. Kishton RJ, Rathmell JC. Novel therapeutic targets of tumor metabolism. *Cancer J*. 2015; 21:62–69. https://doi.org/10.1097/PPO.0000000000000099 PMID: 25815845
6. Altman BJ, Stine ZE, Dang CV. From Krebs to clinic: glutamine metabolism to cancer therapy. *Nat Rev Cancer*. 2016; 16:619–634. https://doi.org/10.1038/nrc.2016.71 PMID: 27492215
7. Scalise M, Pochini L, Galluccio M, Console L, Indiveri C. Glutamine Transport and Mitochondrial Metabolism in Cancer Cell Growth. *Front Oncol*. 2017; 7:306. https://doi.org/10.3389/fonc.2017.00306 PMID: 29376023
8. Sai KKS, Zachar Z, Bingham PM, Mintz A. Metabolic PET Imaging in Oncology. *AJR Am J Roentgenol*. 2017; 209:270–276. https://doi.org/10.2214/AJR.17.18112 PMID: 28463521
9. Zhu L, Ploessl K, Zhou R, Mankoff D, Kung HF. Metabolic Imaging of Glutamine in Cancer. *J Nucl Med*. 2017; 58:533–537. https://doi.org/10.2967/jnumed.116.182345 PMID: 28232608
10. Pantel AR, Ackerman D, Lee SC, Mankoff DA, Gade TP. Imaging Cancer Metabolism: Underlying Biology and Emerging Strategies. *J Nucl Med*. 2018; 59:1340–1349. https://doi.org/10.2967/jnumed.117.199869 PMID: 30042161
11. Choi YK, Park KG. Targeting Glutamine Metabolism for Cancer Treatment. *Biomol Ther (Seoul)*. 2018; 26:19–28. https://doi.org/10.4062/biomolther.2017.178 PMID: 29212303

12. Liu Y, Zhao T, Li Z, Wang L, Yuan S, Sun L. The role of ASCT2 in cancer: A review. *Eur J Pharmacol*. 2018; 837:81–87. https://doi.org/10.1016/j.ejphar.2018.07.007 PMID: 30025811

13. Jiang H, Zhang N, Tang T, Feng F, Sun H, Qu W. Target the human Alanine/Serine/Cysteine Transporter 2(ASCT2): Achievement and Future for Novel Cancer Therapy. *Pharmacol Res*. 2020; 158:104844. https://doi.org/10.1016/j.phrs.2020.104844 PMID: 32438035

14. Schulte ML, Fu A, Zhao P, et al. Pharmacological blockade of ASCT2-dependent glutamine transport leads to antitumor efficacy in preclinical models. *Nat Med*. 2018; 24:194–202. https://doi.org/10.1038/nm.4464 PMID: 29334372

15. Qu W, Zha Z, Ploessl K, et al. Synthesis of optically pure 4-fluoro-glutamines as potential metabolic imaging agents for tumors. *J Am Chem Soc*. 2011; 133:1122–1133. https://doi.org/10.1021/ja109203d PMID: 21190335

16. Zhang X, Basuli F, Shi ZD, et al. Automated synthesis of [(18)F](2S,4R)-4-fluoroglutamine on a GE TRACERlab FX-N Pro module. *Appl Radiat Isot*. 2016; 112:110–114.

17. Li S, Schmitz A, Lee H, Mach RH. Automation of the Radiosynthesis of Six Different (18)F-labeled radiotracers on the AllinOne. *EJNMMI Radiopharm Chem*. 2017; 1:15. https://doi.org/10.1186/s41181-016-0018-0 PMID: 29564391

18. Zhang Y, Zhang L, Yang J, et al. Initial experience in synthesis of (2S,4R)-4-[(18) F]fluoroglutamine for clinical application. *J Labelled Comp Radiopharm*. 2019; 62:209–214. https://doi.org/10.1002/jlcr.3719 PMID: 30861162

19. Lieberman BP, Ploessl K, Wang L, et al. PET imaging of glutaminolysis in tumors by 18F-(2S,4R)4-fluoroglutamine. *J Nucl Med*. 2011; 52:1947–1955. https://doi.org/10.2967/jnumed.111.093815 PMID: 22095958

20. Ploessl K, Wang L, Lieberman BP, Qu W, Kung HF. Comparative evaluation of 18F-labeled glutamic acid and glutamine as tumor metabolic imaging agents. *J Nucl Med*. 2012; 53:1616–1624. https://doi.org/10.2967/jnumed.111.101279 PMID: 22935578

21. Venneti S, Dunphy MP, Zhang H, et al. Glutamine-based PET imaging facilitates enhanced metabolic evaluation of gliomas in vivo. *Sci Transl Med*. 2015; 7:274ra217. https://doi.org/10.1126/scitranslmed.aaa1009 PMID: 25673762

22. Hassanein M, Hight MR, Buck JR, et al. Preclinical Evaluation of 4-[18F]Fluoroglutamine PET to Assess ASCT2 Expression in Lung Cancer. *Mol Imaging Biol*. 2016; 18:18–23. https://doi.org/10.1007/s11307-015-0862-4 PMID: 25971659

23. Schulte ML, Hight MR, Ayers GD, et al. Non-Invasive Glutamine PET Reflects Pharmacological Inhibition of BRAF(V600E) In Vivo. *Mol Imaging Biol*. 2017; 19:421–428. https://doi.org/10.1007/s11307-016-1008-z PMID: 27770401

24. Zhou R, Pantel AR, Li S, et al. [(18)F](2S,4R)4-Fluoroglutamine PET Detects Glutamine Pool Size Changes in Triple-Negative Breast Cancer in Response to Glutaminase Inhibition. *Cancer Res*. 2017; 77:1476–1484. https://doi.org/10.1158/0008-5472.CAN-16-1945 PMID: 28202527

25. Abu Aboud O, Habib SL, Trott J, et al. Glutamine Addiction in Kidney Cancer Suppresses Oxidative Stress and Can Be Exploited for Real-Time Imaging. *Cancer Res*. 2017; 77:6746–6758. https://doi.org/10.1158/0008-5472.CAN-17-0930 PMID: 29021138

26. Liu F, Xu X, Zhu H, et al. PET Imaging of (18)F-(2 S,4 R)4-Fluoroglutamine Accumulation in Breast Cancer: From Xenografts to Patients. *Mol Pharm*. 2018; 15:3448–3455. https://doi.org/10.1021/acs.molpharmaceut.8b00430 PMID: 29985631

27. Li C, Huang S, Guo J, et al. Metabolic Evaluation of MYCN-Amplified Neuroblastoma by 4-[(18)F]FGln PET Imaging. *Mol Imaging Biol*. 2019; 21:1117–1126. https://doi.org/10.1007/s11307-019-01330-9 PMID: 30850970

28. Miner MW, Liljenback H, Virta J, et al. (2S, 4R)-4-[(18)F]Fluoroglutamine for In vivo PET Imaging of Glioma Xenografts in Mice: an Evaluation of Multiple Pharmacokinetic Models. *Mol Imaging Biol*. 2020; 22:969–978. https://doi.org/10.1007/s11307-020-01472-1 PMID: 31993927

29. Viswanath V, Zhou R, Lee H, et al. Kinetic Modeling of (18)F-(2S,4R)4-Fluoroglutamine in Mouse Models of Breast Cancer to Estimate Glutamine Pool Size as an Indicator of Tumor Glutamine Metabolism. *J Nucl Med*. 2021; 62(8):1154–1162. https://doi.org/10.2967/jnumed.120.250977 PMID: 33277391

30. Valtorta S, Toscani D, Chiu M, et al. [(18)F](2S,4R)-4-Fluoroglutamine as a New Positron Emission Tomography Tracer in Myeloma. *Front Oncol*. 2021; 11:760732. https://doi.org/10.3389/fonc.2021.760732 PMID: 34712616

31. Dunphy MPS, Harding JJ, Venneti S, et al. In Vivo PET Assay of Tumor Glutamine Flux and Metabolism: In-Human Trial of (18)F-(2S,4R)-4-Fluoroglutamine. *Radiology*. 2018; 287:667–675. https://doi.org/10.1148/radiol.2017162610 PMID: 29388903

32. Xu X, Zhu H, Liu F, et al. Imaging Brain Metastasis Patients With 18F-(2S,4R)-4-Fluoroglutamine. *Clin Nucl Med*. 2018; 43:e392–e399. https://doi.org/10.1097/RLU.0000000000002257 PMID: 30179907

33. Xu X, Zhu H, Liu F, et al. Dynamic PET/CT imaging of (18)F-(2S, 4R)4-fluoroglutamine in healthy volunteers and oncological patients. *Eur J Nucl Med Mol Imaging*. 2020; 47:2280–2292. https://doi.org/10.1007/s00259-019-04543-w PMID: 32166510

34. Grkovski M, Goel R, Krebs S, et al. Pharmacokinetic Assessment of (18)F-(2S,4R)-4-Fluoroglutamine in Patients with Cancer. *J Nucl Med*. 2020; 61:357–366. https://doi.org/10.2967/jnumed.119.229740 PMID: 31601700

35. McDougald W, Vanhove C, Lehnert A, et al. Standardization of Preclinical PET/CT Imaging to Improve Quantitative Accuracy, Precision, and Reproducibility: A Multicenter Study. *J Nucl Med*. 2020; 61:461–468. https://doi.org/10.2967/jnumed.119.231308 PMID: 31562220

36. Dandekar M, Tseng JR, Gambhir SS. Reproducibility of 18F-FDG microPET studies in mouse tumor xenografts. *J Nucl Med*. 2007; 48:602–607. https://doi.org/10.2967/jnumed.106.036608 PMID: 17401098

37. Mannheim JG, Mamach M, Reder S, et al. Reproducibility and Comparability of Preclinical PET Imaging Data: A Multicenter Small-Animal PET Study. *J Nucl Med*. 2019; 60:1483–1491. https://doi.org/10.2967/jnumed.118.221994 PMID: 30850496

38. Savaikar MA, Whitehead T, Roy S, et al. Preclinical PERCIST and 25% of SUVmax Threshold: Precision Imaging of Response to Therapy in Co-clinical (18)F-FDG PET Imaging of Triple-Negative Breast Cancer Patient-Derived Tumor Xenografts. *J Nucl Med*. 2020; 61:842–849. https://doi.org/10.2967/jnumed.119.234286 PMID: 31757841

39. Altman DG, Bland JM. Measurement in Medicine—the Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society Series D-the Statistician*. 1983; 32:307–317.

40. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. 1990; 20:337–340. https://doi.org/10.1016/0010-4825(90)90013-f PMID: 2257734

41. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat*. 2007; 17:571–582. https://doi.org/10.1080/10543400701329422 PMID: 17613642

42. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989; 45:255–268. PMID: 2720055

43. Carrasco JL, King TS, Chinchilli VM. The concordance correlation coefficient for repeated measures estimated by variance components. *J Biopharm Stat*. 2009; 19:90–105. https://doi.org/10.1080/10543400802527890 PMID: 19127469

44. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Stat Med*. 2007; 26:3095–3113. https://doi.org/10.1002/sim.2778 PMID: 17216594

45. Carrasco JL, Phillips BR, Puig-Martinez J, King TS, Chinchilli VM. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. *Computer Methods and Programs in Biomedicine*. 2013; 109:293–304. https://doi.org/10.1016/j.cmpb.2012.09.002 PMID: 23031487

46. Baessler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol*. 2019; 54:221–228. https://doi.org/10.1097/RLI.0000000000000530 PMID: 30433891

47. Buckler AJ, Danagoulian J, Johnson K, et al. Inter-Method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-Retest Data. *Acad Radiol*. 2015; 22:1393–1408. https://doi.org/10.1016/j.acra.2015.08.007 PMID: 26376841

48. Saltybaeva N, Tanadini-Lang S, Vuong D, et al. Robustness of radiomic features in magnetic resonance imaging for patients with glioblastoma: Multi-center study. *Phys Imaging Radiat Oncol*. 2022; 22:131–136. https://doi.org/10.1016/j.phro.2022.05.006 PMID: 35633866

49. Burdick RK, Borror CM, Montgomery DC. *Design and Analysis of Gauge R&R Studies*: *Making Decisions with Confidence Intervals in Random and Mixed Anova Models*. 2005; 17:1–201.

50. Woodall WH, Borror CM. Some relationships between gage R&R criteria. *Quality and Reliability Engineering International*. 2008; 24:99–106.

51. Choe HM, Kim M, Lee EK. EMSaov: An R Package for the Analysis of Variance with the Expected Mean Squares and its Shiny Application. *R Journal*. 2017; 9:252–261.

52. Bates D MM, Bolker B, Walker S lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0–6. x. http://CRAN.R-project.org/package=lme4.

53. Parker RA, Scott C, Inacio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: a tutorial for practitioners. *Bmc Medical Research Methodology*. 2020; 20:154. https://doi.org/10.1186/s12874-020-01022-x PMID: 32532218

54. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*. 2002; 97:257–270.

55. Barnhart HX, Yow E, Crowley AL, et al. Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Statistical Methods in Medical Research*. 2016; 25:2939–2958. https://doi.org/10.1177/0962280214534651 PMID: 24831133

56. Whisenant JG, Ayers GD, Loveless ME, Barnes SL, Colvin DC, Yankeelov TE. Assessing reproducibility of diffusion-weighted magnetic resonance imaging studies in a murine model of HER2+ breast cancer. *Magn Reson Imaging*. 2014; 32:245–249. https://doi.org/10.1016/j.mri.2013.10.013 PMID: 24433723

57. Whisenant JG, Peterson TE, Fluckiger JU, Tantawy MN, Ayers GD, Yankeelov TE. Reproducibility of static and dynamic (18)F-FDG, (18)F-FLT, and (18)F-FMISO MicroPET studies in a murine model of HER2+ breast cancer. *Mol Imaging Biol*. 2013; 15:87–96. https://doi.org/10.1007/s11307-012-0564-0 PMID: 22644988