Journal of Nuclear Medicine, published on November 22, 2019 as doi:10.2967/jnumed.119.234286

SUV_{25} and µPERCIST: Precision Imaging of Response to Therapy in Co-Clinical FDG-PET Imaging of Triple Negative Breast Cancer (TNBC) Patient-Derived Tumor Xenografts (PDX)

Madhusudan A. Savaikar[1], Timothy Whitehead[1], Sudipta Roy[1], Lori Strong[1], Nicole Fettig[1], Tina Prmeau[2], Jingqin Luo[3], Shunqiang Li[2], Richard L. Wahl[1], Kooresh I. Shoghi[1,4]

[1]Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA; [2]Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, USA; [3]Department of Surgery, Washington University School of Medicine, St. Louis, USA; [4]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA

Running title: Co-clinical imaging

Keywords: Co-clinical trials, triple negative breast cancer, patient-derived tumor xenografts (PDX), quantitative imaging, response to therapy, reproducibility.

First author (in training):
Madhusudan A. Savaikar, Ph.D.
Department of Radiology
510 S. Kingshighway Blvd.
Campus Box 8225
St. Louis, MO 63110
Email: msavaikar@wustl.edu

Corresponding author:
Kooresh Isaac Shoghi, Ph.D.
Department of Radiology
510 S. Kingshighway Blvd.
Campus Box 8225
St. Louis, MO 63110
Tel.: 314-362-8990
Email: shoghik@wustl.edu

Word count: 4908

Financial support: This work was supported by NCI grants U24CA209837; U54CA224083 and U54CA199092; NIBIB grant P41EB025815; Siteman Cancer Center (SCC) Support Grant P30CA091842; and Internal funds provided by Mallinckrodt Institute of Radiology.

Disclosure of potential conflicts of interest: None

ABSTRACT

Numerous recent works highlight the limited utility of established tumor cell lines in recapitulating the heterogeneity of tumors in patients. More realistic preclinical cancer models are thought to be provided by transplantable, patient-derived tumor xenografts (PDX). Inter- and intra-tumor heterogeneity of PDX, however, present several challenges in developing optimal quantitative pipelines to assess response to therapy. The objective of this work was to develop and optimize image metrics of FDG-PET to assess response to combination docetaxel/carboplatin therapy in a co-clinical trial involving triple negative breast cancer (TNBC) PDX. We characterize the reproducibility of SUV metrics to assess response to therapy and optimize a preclinical PERCIST (μPERCIST) paradigm to complement clinical standards. Considerations in this effort included variability in tumor growth rate and tumor size; solid tumor vs. tumor heterogeneity and necrotic phenotype; and optimal selection of tumor slices versus whole tumor. A test-retest protocol was implemented to optimize the reproducibility of FDG-PET SUV thresholds, SUV_{peak} metrics, and μPERCIST parameters. In assessing response to therapy, FDG-PET imaging was performed at baseline and +4 days following therapy. The reproducibility, accuracy, variability, and performance of imaging metrics to assess response to therapy were determined. We defined an index—Quantitative Response Assessment Score (QRAS)—to integrate parameters of prediction and precision, and thus aid in selecting the optimal image metrics of response to therapy. Our data suggests that a threshold value of 25% (SUV_{25}) of SUV_{max} was highly reproducible (<9% variability). Concordance and reproducibility of μPERCIST were maximized at α=0.7 and β=2.8 and exhibited high correlation to SUV_{25} measures of tumor uptake, which in turn correlated to SUV of metabolic tumor. QRAS favors SUV_{25} followed by SUV_{P14} as optimal metrics of response to therapy. Additional studies are warranted to fully characterize the utility of SUV_{25} and μPERCIST SUV_{P14} as image metrics of response to therapy across a wide range of therapeutic regiments and PDX models.

INTRODUCTION

Co-clinical trials are an emerging area of investigation in which a clinical trial is coupled with a corresponding preclinical trial to inform the corresponding clinical trial (*1-7*). The preclinical arm of the co-clinical trial generally uses genetically engineered mouse models (GEMMs) of human cancer or patient-derived tumor xenografts (PDXs) to aid in assessing therapeutic efficacy, patient stratification, and to design optimal treatment strategies (*8,9*). The emergence of GEMMs and PDXs as co-clinical platforms is largely motivated by the realization that established cell-lines do not recapitulate the heterogeneity of human tumors and the diversity of tumor phenotypes (*10*), and that better oncology models are needed to support high-impact translational cancer research. To that end, the National Cancer Institute's (NCI) Patient-Derived Models Repository (https://pdmr.cancer.gov), EuroPDX (https://www.europdx.eu), academic institutions, and numerous commercial entities have launched wide-ranging PDX and GEMMs repositories to advance the biological and molecular basis for cancer prevention and treatment towards realization of precision medicine. Importantly, the NCI has recently launched the Co-Clinical Imaging Research Resource Program (CIRP) (https://nciphub.org/groups/cirphub) to advance the utility of oncology models of human cancers in preclinical imaging.

The use of PDX in preclinical imaging offers numerous advantages in translational imaging research. Chief among them is retention of human tumor heterogeneity which can be exploited to develop image metrics of heterogeneity and response to therapy. Unlike established tumor cell lines, PDXs also exhibit significant variability in growth profiles both, within patient-generated PDX and between patient-generated PDX. In addition to biological variability (due to genotypic variability), the gross phenotype of PDX tumors is also highly variable with some exhibiting a necrotic phenotype. Clinically, patients with TNBC have shown high sensitivity to the addition of carboplatin to anthracycline and taxane-based neoadjuvant chemotherapy (NAC) (*11*). With that in mind, we designed a co-clinical trial to assess the efficacy of FDG-PET in predicting response to docetaxel/carboplatin therapy in the context of a co-clinical trial (Figure 1A, ClinicalTrial.gov ID # NCT02124902). The clinical arm of the co-clinical trial aims to predict response to combination of docetaxel/carboplatin therapy using FDG-PET. The preclinical arm of the co-clinical trial uses tumor biopsies derived from patients in the co-

clinical trial to generate PDX which are then used, among other objectives, to optimize FDG-PET imaging biomarkers of response to therapy.

Through this framework, we identified six TNBC subtypes including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem–like (MSL), and a luminal androgen receptor (LAR) subtype (Figure S1). A subset of these PDX were used to develop optimal quantitative imaging strategies to assess response to combination docetaxel/carboplatin therapy in triple negative breast cancer (TNBC) PDX. We characterize the reproducibility and precision of SUV metrics to assess response to therapy, and optimize a preclinical Positron Emission Tomography Response Criteria in Solid Tumors (μPERCIST) paradigm to complement clinical PERCIST standards (*12*). The performance of SUV quantiles for whole tumor, single high-intensity slice (SS), maximum uptake (SUV_{max}), and SUV peak (SUV_{peak}) measures to assess response to therapy is determined. The work addresses a central effort within the imaging community and the NCI to reach a consensus on reproducibility and utility of imaging metrics of response to therapy in oncology animal models.

METHODS

Generation of TNBC PDX

Gene expression analyses of 93 TNBC PDXs (29657 unique genes/probes) was performed to identify six TNBC subtypes including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem–like (MSL), and a luminal androgen receptor (LAR) subtype (Figure S1) as described previously (*13*). Details regarding animals, surgeries, and tumor xenografts were reported previously (*14*) and in Supplementary Methods. All animal experiments were conducted in compliance with the Guidelines for the Care and Use of Research Animals established by Washington University's Animal Studies Committee.

Characterization of PDX Tumor Growth

After inoculation, mice were examined three times per week for palpable tumors. When a palpable tumor was observed, caliper measurements were made of the major (L) and minor (W) axes bi-weekly. Tumor volume was calculated using the formula volume=$1/6*L*W^2$. The natural growth curves were constructed for each PDX

subtype using the daily average and standard deviation of all mice. To determine tumor growth doubling times (T_{2x}) the exponential growth region was determined for each mouse individually. Time scale was shifted to the start of exponential growth and the tumor volumes were normalized to the volume at this time. The mean *ln*(fold change) was plotted against time in exponential growth and the doubling time calculated from the slope.

Preclinical Studies

Three distinct experiments were carried out: 1) test-retest studies were performed on consecutive days (Day 1 vs Day 2) to assess the reproducibility of PET image metrics (total of 34 PDX; 13 solid, 21 necrotic phenotype); 2) an experiment to assess the impact of animal handling/imaging on survival (N=16 PDX); and 3) a therapeutic study with imaging to assess response to therapy (total of 29 PDX; 13 solid, 16 necrotic phenotype). The second experiment suggested that repeat imaging impacted survival (Supplemental Results, Figure S2), and thus, we kept imaging to a minimum. In all therapeutic studies, docetaxel (20mg/kg I.P.)/carboplatin (50mg/kg I.P) was administered at baseline (following imaging) and weekly for a period of four weeks. Preclinical imaging was performed at baseline and +4 days following therapy (Figure 1A). Tumor volumes were measured bi-weekly as surrogate measures of response to therapy.

Preclinical Imaging and Image Analysis

Preclinical PET/CT imaging is described in detail in Supplemental Methods. Data from 50-60min post injection of FDG were used in the analysis. The PET/CT image data from all the mice were processed in two-steps. In the first step, the co-registered PET/CT images were analyzed using the Inveon Research Workplace (IRW) software (Siemens Healthcare). Regions of interest (ROIs) were manually drawn on co-registered PET/CT images. The corresponding voxels were further processed in MATLAB (Mathworks Inc) as detailed in subsequent sections. ROIs and individual voxels were normalized to standardized uptake value (SUV) using the relation: SUV = [activity (Bq / mL)] x [animal weight (g)] / [injected dose (Bq)]. Multiple analytic pipelines were pursued including 1) Intensity histogram reproducibility analysis (IHRA) to compute tumor thresholds. At each percent threshold (Th), the SUV_{Th} is calculated as percent of SUV_{max}, i.e., $SUV_{Th}=Th*SUV_{max}/100$. SUV_{th} represents the mean of the voxels with SUV greater than SUV_{Th}; 2) SUV_{max} and SUV_{peak} of three distinct volumes

centered at SUV_{max}; 3) whole tumor and single slice (SS) analyses; and 4) optimization and evaluation of preclinical PERCIST (μPERCIST) as detailed in Supplemental Methods.

Statistical Analysis

The reproducibility analysis included image data from IM and BL1 and BL2 PDX. Optimization of μPERCIST, assessment of response to therapy, and performance of image metrics in assessing response to therapy included image data from IM, LAR, M, BL1, and BL2 PDX.

Growth profile of PDX. Coincidence tests (*15*) were used to compare the slopes between passages within a PDX subtype and between PDX subtypes. GraphPad Prism ver.7 was used to perform these tests.

Reproducibility Statistics. PDX were imaged on consecutive days to assess reproducibility. Two methods for assessing reproducibility were used, Lin's concordance correlation coefficient (LCC) (*16*) and Bland-Atlman plots (BA) (*17*). The "Day 1" vs. "Day 2" absolute differences were shown to be independent of the means using Kendal' tau test for correlation (*18*) using Stata version 12.1. The limits of agreement (LOA) provided by the repeatability coefficients (RC) (see Supplemental Methods) of the test-retest data were later used to investigate the utility of metrics to assess response to therapy.

Performance Assessment of Response to Therapy. Decrease in tumor volume of greater than 20% was considered response to therapy; no change or increase in tumor volume as non-response. The change in image metrics between +4d post-treatment and baseline scan was used as the predictive criterion. To assess the applicability of these parameters, the difference between the baseline and post treatment values were plotted against the mean of the two values on the BA plot for all PDX tumors. If image metrics of response to therapy were within the LOA, it was considered indistinguishable from metric variability and prediction was not evaluated. The two class labels are response and no-response, which were used to assess response to the therapy (*19,20*). End-point caliper measured volume changes were considered as binary indicators of response to therapy. Additional details are provided in Supplementary Methods section. Finally, we defined an index which we coined

"Quantitative Response Assessment Score (QRAS)" to integrate parameters of prediction performance and precision, and thus aid in selecting optimal image metrics. QRAS is defined as QRAS=(RC)*(Uncertainty)/(F-Score) with lower scores favorable.

RESULTS

Variability in PDX tumor growth

The caliper volume growth curves for IM, BL1, BL2, M and LAR PDX tumors are depicted in Figure 1B, and the average logarithmic growth curves in Figure 1C. Coincidence tests for the slopes of the logarithmic growth curves indicate that, IM equal to BL1, and BL2 equal to M, and that these groups differ from each other and from LAR ($p<0.0001$ for all comparisons). The average doubling times calculated is similarly depicted in Figure 1C. The day of scan distribution of PET tumor volumes used for test-retest studies in depicted in Figure 1D.

IHRA to optimize image metrics of reproducibility

Select tumor phenotypes are depicted in Figure 2 for, BL1 and BL2 tumors. Normalized line-intensity profiles across individual slices from distinct PDX tumors (Figure 2B) and when centered at "zero" (Figure 2C) illustrate the heterogeneity in the tumors. The minima along the line profiles in Figure 2C vary from 0 to ~ 25 % of the hottest voxel. Representative samples of H&E staining for each of the PDX is depicted in Figure 2D.

The test-retest PET-derived volume measures are depicted in Figure 3A. There is excellent agreement between Day 1 and Day 2 volumes measures (R^2=0.92). The IHRA for solid tumors is depicted in Figure 3B. For SUV_{max} (i.e., 100% of SUV_{max}) reproducibility is low (LCC=~0.58); with increasing quantiles LCC saturates at 25% tumor quantile (LCC=0.77, PCC=0.80, and BCF=0.97). Thus, metabolic tumor volume defined by 25% of SUV_{max} is an inflection point at which point the PCC, BCF, and LCC saturate and show negligible change thereafter. These observations are better reflected by the BA plots of the quantile boundaries shown in Figures 3C and Figure 3D. The 95 % confidence limits of agreement (solid red line) of SUV_{mean} (RC=0.20) is significantly tighter than the confidence limit of SUV_{max} (RC=0.34).

Overall, 21 PDX tumors exhibited a necrotic core phenotype (low FDG uptake at the core) with varying tumor dimensions. In contrast, solid tumors (N=13) are defined as having no visual necrotic phenotype. PET-derived volumes ranged from ~ 85 to 1400 mm^3. Despite the range, there is excellent agreement in Day 1 and Day 2 concordance plot (Figure 3E), with a slight bias at high tumor volumes due to a large tumor at 1400mm^3 which skews the linearity. The IHRA for these tumors is depicted in Figure 3F. SUV_{max} image metrics exhibited poor concordance (LCC=0.52) with high RC (approaching 0.50). As the intensity quantile reaches 25%, both the PCC and the LCC achieve ~ 0.79. At peak LCC, LCC=0.81 for SUV_{mean} (at 0% IHRA). The BA plots for the SUV_{mean} and SUV_{max} are shown in Figure 3G-3H. The 95 % CL range of SUV_{max} (RC=0.44) is approximately 3-fold higher than that of SUV_{mean} (RC=0.15), suggesting poor reproducibility of SUV_{max}.

High-intensity slice vs. whole tumor analysis

IHRA implemented on a single slice (SS) and the total tumor volume is depicted in Figure 4. All three metrics of performance, PCC, BCF, and LCC peak at 25% of intensity voxels with LCC=0.88 for SS, and LCC=0.93 for the total tumor volume. There is negligible change at quantiles < 25% as remaining low intensity voxels in the ROI are included in the analysis (Panel A). The BA plots of SS and whole tumor show similar statistics (Panel B). There is excellent correlation between Day 1 vs Day 2 measures, as indicated in Figure 4 panel C.

Optimization of preclinical PERCIST—μPERCIST

The IHRA plot of Figure 5A depicts LCC and RC as a function of α and select β values. Figure 5B depicts the surface plot of the objective function LCC/RC which is maximized at α=0.7 (denoted by dash line in Figure 5A) and β=2.8. Tumor SUV BA plot for optimized liver threshold is depicted in Figure S3. Figure 5C depicts the correlation between mean tumor SUV with liver threshold defined by μPERCIST parameters (α=0.7, β=2.8) (SUV_{mLTh}) and SUV_{25} while Figure 5D depicts the correlation between SUV_{25} and SUV of metabolic tumor ($SUV_{metabolic}$). There is excellent correlation between SUV_{mLTh} and SUV_{25} (R^2=0.98) and between SUV_{25} and $SUV_{metabolic}$ (R^2=0.98) with slope not significantly different than identity (see Figure S4 for correlation between

SUV_{mean} and SUV_{max} to $SUV_{metabolic}$). The BA plots corresponding to SUV_{max} and the three distinct SUV_{peak} are depicted in Figures 5E-H. With increased peak ROI volumes, there is less variability in test-retest measures as denoted by reduced RC.

Predicting response to therapy

In Figure 6, we depict the BA plots of response to therapy for SUV_{mean} and 25% threshold using whole tumor (Figure 6A) and SS (Figure 6B) while Figure 6C depicts the BA plots of response to therapy using SUV_{peak} metrics. The performance of imaging metrics to predict response to therapy for data points outside the LOA is summarized in Table S1 along with percent of datapoints within the LOA. Importantly, the accuracy of predicting response conditioned by subtype is tabulated in Table S2. Figure 7 depicts the percent relative difference (%RD) between image metrics SUV_{25} and SUV peak measures relative to SUV_{mean} in assessing response to therapy. The latter figure suggests that all metrics have higher dynamic range than that of SUV_{mean} to assess response to therapy on the order of 2-4 fold. Table 1 tabulates performance parameter for SUV_{max}, SUV_{25}, and SUV peak measures and the calculated QRAS. Measures of SUV_{25} scored the lowest (best) followed by SUV_{P14}.

DISCUSSION

We generated five PDX subtypes of TNBC as a preclinical platform to develop and optimize image metrics of response to therapy. PDX provided a wide range of phenotypes which we exploited to develop and test image metrics of response to therapy. In light of the heterogeneity of tumors, we took a top-down image-data-centric approach in optimizing image metrics of reproducibility and response to therapy. We stratified PDX tumors to those exhibiting solid phenotype and to those exhibiting a necrotic phenotype and implemented IHRA in each group and the combined groups to define optimal measures of FDG-PET uptake in PDX. For both solid tumors and tumors exhibiting a necrotic phenotype, reproducibility peaked at 25% of SUV_{max}. Similarly, in the combined dataset, measures of reproducibility plateau at 25%. Thus, a threshold of 25% of SUV_{max}, referred to as SUV_{25}, was optimal to maximize reproducibility. Wu et al. (*21*) performed extensive histological analysis of co-registered to preclinical FDG-PET slices in an effort to define tumor boundaries. In agreement with our

findings, Wu et al. (*21*) concluded a minimum threshold (cut-off) value of up to 30% of C_{max} to define viable tumors.

Clinically, the PERCIST criteria (*12*) is widely used to assess response to therapy (*22-25*). In the above-mentioned work, Wu et al. (*21*) additionally optimized a PERCIST-motivated cut-off value of *α**[mean value of liver ROI]+*β**[standard deviation of liver ROI] with α=6 and β=2 to define viable tumors *in vivo*. In an effort to harmonize preclinical efforts with clinical standards of response to therapy assessment, we similarly optimized μPERCIST parameters to maximize concordance and reproducibility. Our data suggest that LCC/RC is maximized at α=0.7 and β=2.8. Liver-threshold tumor uptake values (SUV_{mLTh}) exhibited high correlation to SUV_{25} which in turn highly correlated to $SUV_{metabolic}$, suggesting that SUV_{mLTh} and SUV_{25} provide measures of viable metabolic tumor. The clinical utility of PERCIST is that it provides an internal patient-specific reference across diverse subjects. Herein, we used homogeneous population of mice (all NSG of same approximate weight), thus variability across species/strains needs to be explored.

In assessing response to therapy, the optimal imaging metric needs to take into account the reproducibility, extent of uncertainty in predicting response to therapy, and performance of an imaging metric in assessing response to therapy which led to the development of QRAS. QRAS analyses suggests that SUV_{25} measures (whole tumor and SS) followed by SUV_{P14} are an optimal metrics to assess response to therapy. When using whole-tumor or single slice measures of uptake, inclusion of low intensity voxels attributed to necrotic phenotype are expected to lower image metrics of FDG-PET uptake and bias against measures of tumor response to therapy. Thus, it is expected that exclusion of tumor voxels attributed to necrotic phenotype will improve the sensitivity of imaging biomarkers in assessing response to therapy. Indeed, the above-mentioned metrics have a wider dynamic range than SUV_{mean} (approximately 2-4 fold) in predicting response to therapy. The choice between SUV_{25} and SUV_{P14} may depend on dynamic range of response assessment but will ultimately require additional validation in other animal models, therapeutic interventions, and considerations of confounding factors (e.g., anesthesia). Finally, we note that the accuracy in predicting response to therapy is dependent on the PDX subtype, suggesting that with *a priori* knowledge of TNBC subtype, one can define

confidence in predicting response to therapy. The timing of response assessment may be a function of subtype as well (which we were unable to investigate in this work). This notion underscores the premise of precision medicine and precision imaging, i.e., integrating genomic signatures (defining a subtype in this case) to enhance prediction of response to therapy.

CONCLUSION

The work addresses a central effort within the imaging community and the NCI to reach a consensus on reproducibility and utility of imaging metrics to assess response to therapy in more realistic models of human cancers (e.g., PDX, GEMMs), thus enhancing the translational impact of preclinical imaging studies. In a co-clinical study design using patient-derived tumors, our data suggests that SUV_{25} FDG-PET measures are highly reproducible. Importantly, QRAS scores favor SUV_{25} followed by SUV_{P14} as optimal metrics of response to therapy. The choice between SUV_{25} and SUV_{P14} may depend on dynamic range of response assessment. Additionally, SUV_{25} correlated to optimized μPERCIST measures of tumor uptake and $SUV_{metabolic}$, suggesting that both may provide image metrics of viable tumor. Further studies are warranted to fully characterize the utility of SUV_{25} and optimized implementation of μPERCIST via SUV_{P14} as image metrics of response to therapy across a wide range of therapeutic regiments and animal models of human cancer.

KEY POINTS

QUESTION: What is the optimal FDG-PET SUV image metric of response to therapy in TNBC PDX?

PERTINENT FINDINGS: In a co-clinical study design using PDX, our data suggests that image metric threshold of 25% of SUV_{max} (SUV_{25}) are highly reproducible. QRAS scores favor SUV_{25} followed by SUV_{P14} (for implementation of μPERCIST) as optimal metrics of response to therapy. The choice between SUV_{25} and SUV_{P14} may depend on dynamic range of response assessment.

IMPLICATIONS FOR PATIENT CARE: The work addresses a central effort within the imaging community and the NCI to reach a consensus on reproducibility and utility of imaging metrics to assess response to therapy in

more realistic models of human cancers (e.g., PDX, GEMMs), thus enhancing the translational impact of preclinical imaging studies.

REFERENCES

1. Chen Z, Akbay E, Mikse O, et al. Co-clinical trials demonstrate superiority of crizotinib to chemotherapy in ALK-rearranged non-small cell lung cancer and predict strategies to overcome resistance. *Clin Cancer Res.* 2014;20:1204-1211.

2. Kim HR, Kang HN, Shim HS, et al. Co-clinical trials demonstrate predictive biomarkers for dovitinib, an FGFR inhibitor, in lung squamous cell carcinoma. *Ann Oncol.* 2017;28:1250-1259.

3. Kwong LN, Boland GM, Frederick DT, et al. Co-clinical assessment identifies patterns of BRAF inhibitor resistance in melanoma. *J Clin Invest.* 2015;125:1459-1470.

4. Lunardi A, Ala U, Epping MT, et al. A co-clinical approach identifies mechanisms and potential therapies for androgen deprivation resistance in prostate cancer. *Nat Genet.* 2013;45:747-755.

5. Nishino M, Sacher AG, Gandhi L, et al. Co-clinical quantitative tumor volume imaging in ALK-rearranged NSCLC treated with crizotinib. *Eur J Radiol.* 2017;88:15-20.

6. Owonikoko TK, Zhang G, Kim HS, et al. Patient-derived xenografts faithfully replicated clinical outcome in a phase II co-clinical trial of arsenic trioxide in relapsed small cell lung cancer. *J Transl Med.* 2016;14:111.

7. Sia D, Moeini A, Labgaa I, Villanueva A. The future of patient-derived tumor xenografts in cancer treatment. *Pharmacogenomics.* 2015;16:1671-1683.

8. Cho SY, Kang W, Han JY, et al. An Integrative Approach to Precision Cancer Medicine Using Patient-Derived Xenografts. *Mol Cells.* 2016;39:77-86.

9. Clohessy JG, Pandolfi PP. Mouse hospital and co-clinical trial project--from bench to bedside. *Nat Rev Clin Oncol.* 2015;12:491-498.

10. Sulaiman A, Wang L. Bridging the divide: preclinical research discrepancies between triple-negative breast cancer cell lines and patient tumors. *Oncotarget.* 2017;8:113269-113281.

11. Sharma P, Lopez-Tarruella S, Garcia-Saenz JA, et al. Efficacy of Neoadjuvant Carboplatin plus Docetaxel in Triple-Negative Breast Cancer: Combined Analysis of Two Cohorts. *Clin Cancer Res.* 2017;23:649-657.

12. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50 Suppl 1:122S-150S.

13. Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011;121:2750-2767.

14. Li S, Shen D, Shao J, et al. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* 2013;4:1116-1130.

15. Glantz SA. Primer of biostatistics. 2012.

16. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255-268.

17. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135-160.

18. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR Biomed.* 2002;15:132-142.

19. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology.* 2016;4:320-330.

20. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10:e0118432.

21. Wu I, Wang H, Huso D, Wahl RL. Optimal definition of biological tumor volume using positron emission tomography in an animal model. *EJNMMI Res.* 2015;5:58.

22. Cho SY, Lipson EJ, Im HJ, et al. Prediction of Response to Immune Checkpoint Inhibitor Therapy Using Early-Time-Point (18)F-FDG PET/CT Imaging in Patients with Advanced Melanoma. *J Nucl Med.* 2017;58:1421-1428.

23. Hyun OJ, Luber BS, Leal JP, et al. Response to Early Treatment Evaluated with 18F-FDG PET and PERCIST 1.0 Predicts Survival in Patients with Ewing Sarcoma Family of Tumors Treated with a Monoclonal Antibody to the Insulinlike Growth Factor 1 Receptor. *J Nucl Med.* 2016;57:735-740.

24. Kairemo K, Rohren EM, Anderson PM, et al. Development of sodium fluoride PET response criteria for solid tumours (NAFCIST) in a clinical trial of radium-223 in osteosarcoma: from RECIST to PERCIST to NAFCIST. *ESMO Open.* 2019;4:e000439.

25. Kim JE, Chae SY, Kim JH, et al. 3'-Deoxy-3'-(18)F-Fluorothymidine and (18)F-Fluorodeoxyglucose positron emission tomography for the early prediction of response to Regorafenib in patients with metastatic colorectal cancer refractory to all standard therapies. *Eur J Nucl Med Mol Imaging.* 2019.

TABLE 1. Parameters in selecting optimal image metrics of response to therapy.

SUV metric	RC	F-Score	Uncertain Fraction	QRAS
ΔSUV_{max}	0.73	0.73	0.45	0.45
ΔSUV_{25}	0.28	0.72	0.31	0.12
ΔSUV_{25} (SS)	0.33	0.74	0.34	0.15
ΔSUV_{P4}	0.59	0.77	0.48	0.37
ΔSUV_{P14}	0.47	0.74	0.34	0.22
ΔSUV_{P33}	0.45	0.69	0.41	0.27

RC, Repeatability Coefficient; F-Score is derived from the performance evaluation of response to therapy. "Uncertain Fraction" is the fraction of all data points within the LOA for a given image metric (see Supplemental Results) which were not used to predict response. Quantitative Response Assessment Score QRAS=(RC)*(Uncertain Fraction)/(F-Score) with lower scores as favorable.

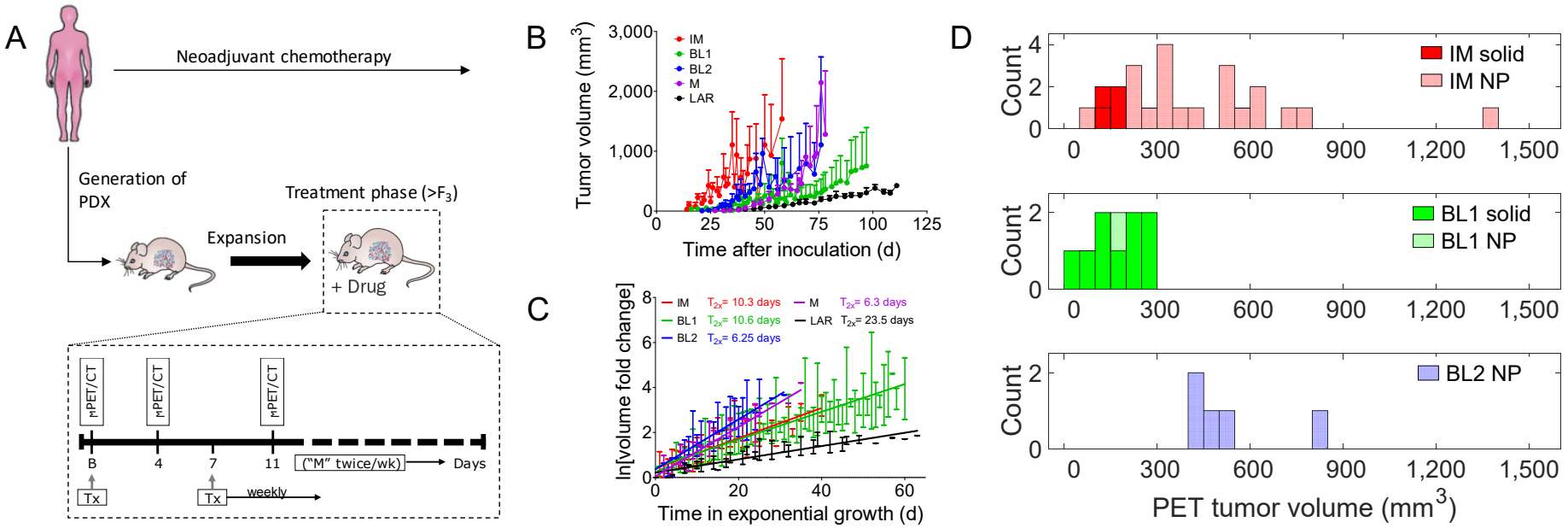


Figure 1. Co-clinical study design and heterogeneity of PDX. (A) Co-clinical study design. PDX were generated from patient biopsies derived at baseline; Preclinical study design. Following baseline ("B") imaging, PDX were treated ("Tx") weekly for 4 weeks. Tumor volumes were measured ("M") by caliper bi-weekly. Initially we tested mid-therapy imaging at 4 days and 11 days post-baseline. In the therapy arm, only 4-day time point was used; **(B)** Growth profile of basal-like (BL1 and BL2), and immunomodulatory (IM), mesenchymal (M), luminal androgen receptor (LAR) TNBC subtype PDX. **(C)** Log-normalized growth profile with offset time-scale to start of exponential growth. **(D)** Histogram of tumor volumes for PDX subtypes used in test-retest studies with designation of solid tumor and necrotic phenotype (NP).

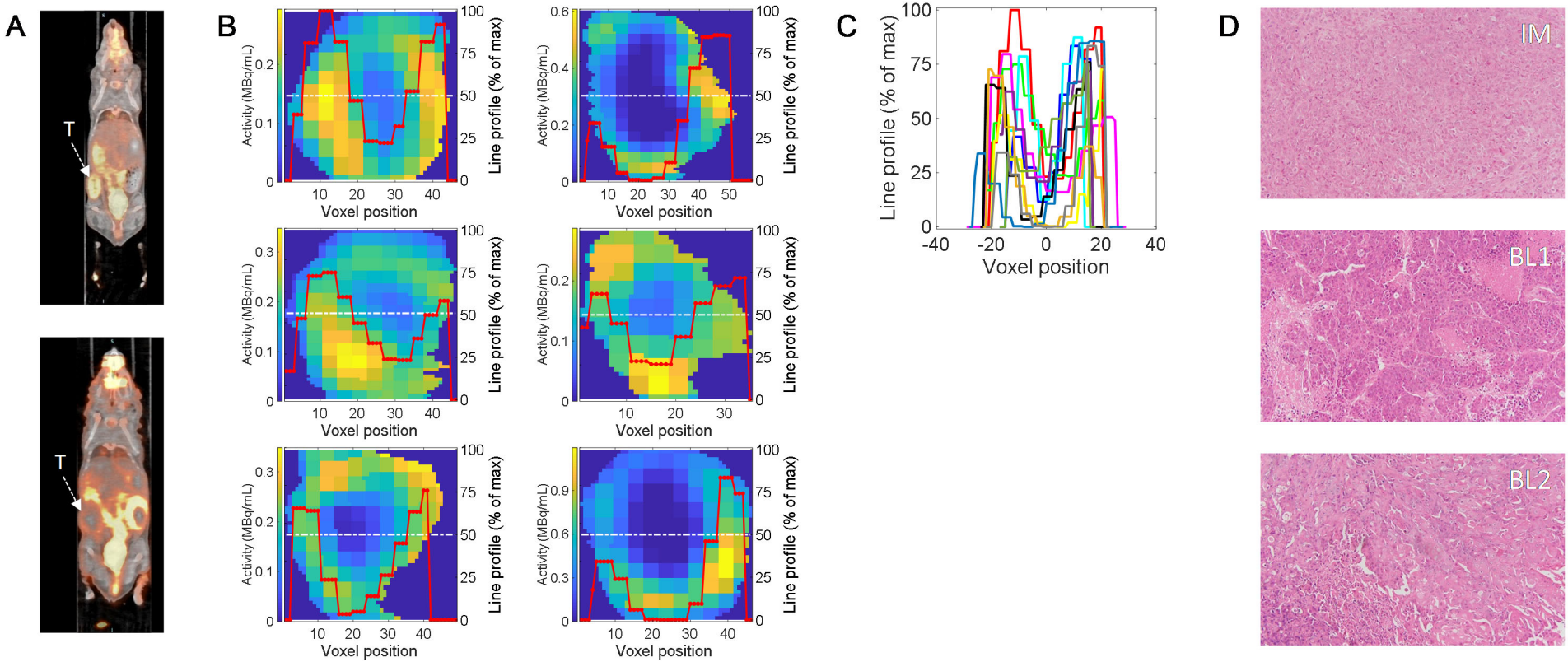


Figure 2. **Image Analytics of TNBC subtypes and select H&E pathology. (A)** Representative FDG-PET coronal slices of PDX (Tumor denoted by "T") **(B)** Representative coronal tumor slices at the center coronal plane from PDX of all three subtypes. The red lines denote the intensity line profiles along the slice center (white dotted line) normalized to the maximum intensity in the respective slice (displayed as percent of max intensity). **(C)** Intensity line profiles of all tumors in (A) with their minima centered at zero position to highlight variability in threshold of necrotic phenotype. **(D)** Representative H&E staining of IM, BL1, and BL2 PDX.

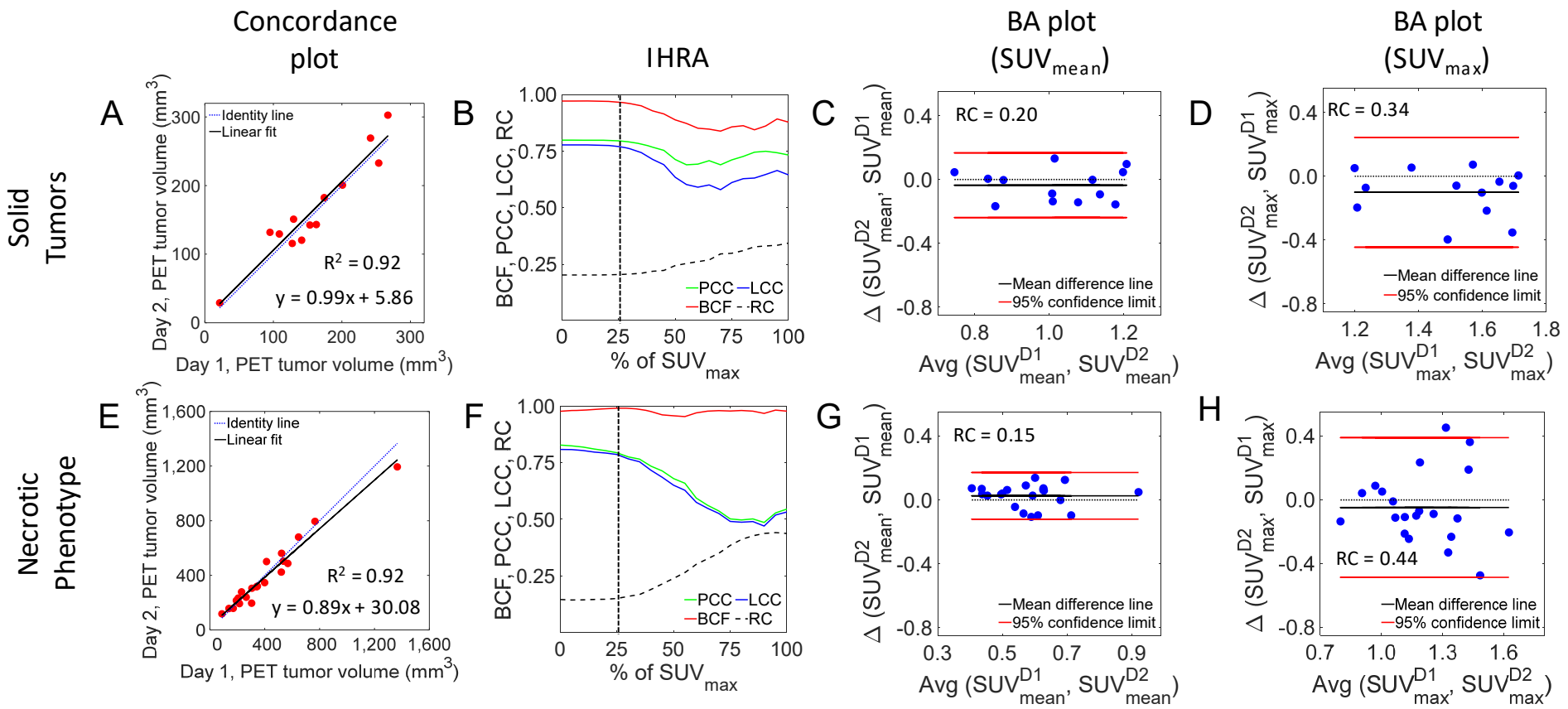


Figure 3. Concordance and IHRA for solid and necrotic tumor phenotypes. (A) Day 1 vs Day 2 metabolic (PET) tumor volume concordance plot for solid tumors. The blue dotted line represents the identity curve, the black solid line is the linear least squares line. **(B)** IHRA depicting PCC, BCF, LCC, and RC as function of percent (threshold) of SUV_{max}. **(C)** and **(D)** depict the BA plots for the SUV_{mean} and SUV_{max}. The red lines represent the 95 % CL and the black solid line is the mean difference line. **(E) - (H)** depict similar parameters for tumors with necrotic phenotype.

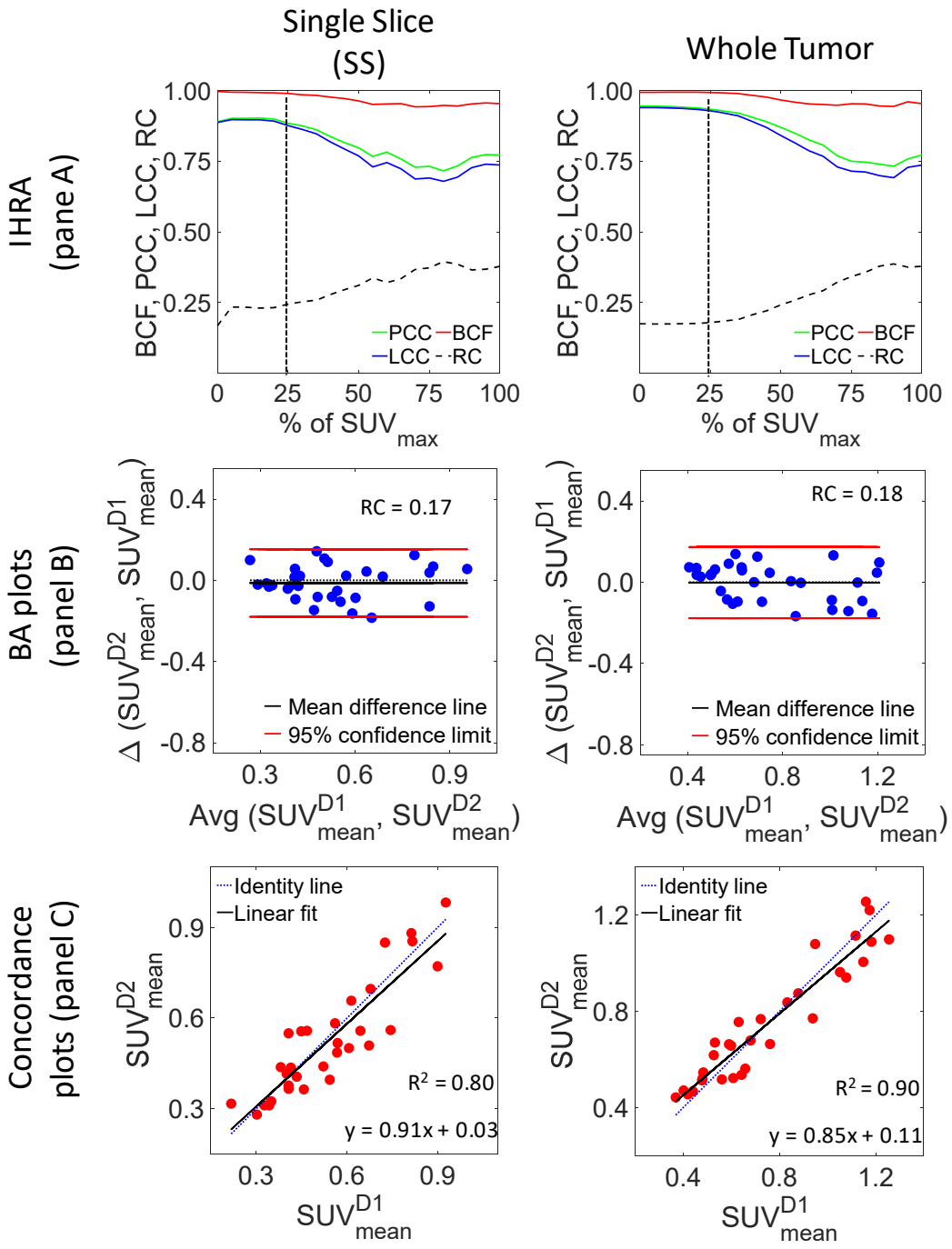


Figure 4. Concordance and reproducibility analysis of all PDX in test-retest cohort. Each Panel depicts **(A)** IHRA **(B)** BA plots and **(C)** concordance plot for single slice (SS) and whole tumor. In all three cases, the PCC, BCF, and LCC show similar trends and LCC approaches a plateau at ~ 25% of SUV_{max}, or SUV_{25}.

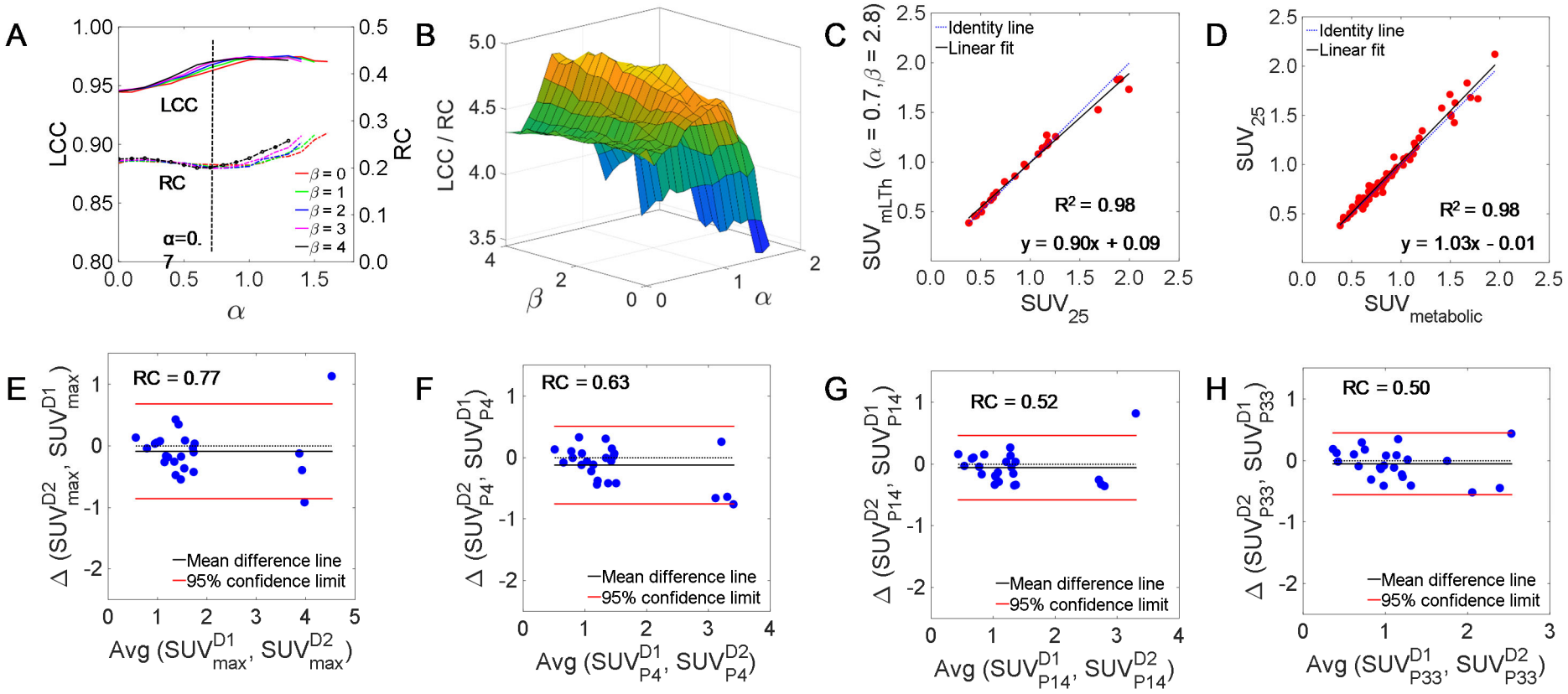


Figure 5. **Optimization of preclinical μPERCIST. (A)** IHRA of μPERCIST depicting LCC and RC as a function of α and select β values. **(B)** Surface plot of the objective function LCC/RC which is maximized at α= 0.7 (denoted by dash line in Figure 5A) and β= 2.8. **(C)** Correlation between optimized liver threshold for μPERCIST and SUV_{25}. **(D)** Correlation between SUV_{25} and SUV of metabolic tumor ($SUV_{metabolic}$). **(E)** through **(H)** depict BA plots for SUV_{max} and SUV_{peak} with spherical volumes of 4mm^3 (SUV_{P4}), 14mm^3 (SUV_{P14}), and 33mm^3 (SUV_{P33}) centered at SUV_{max}.

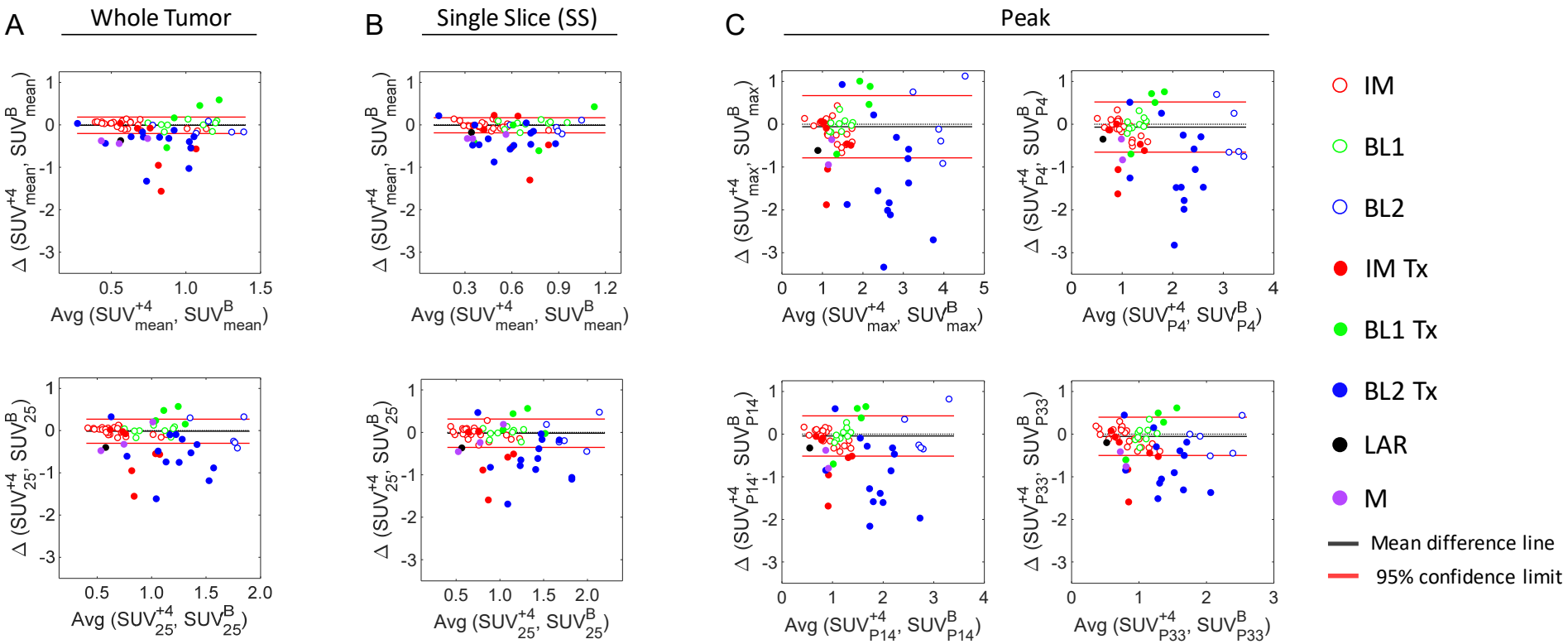


Figure 6. BA plots of image metrics of response assessment for whole tumor measures, single slice, and peak metrics. (A) BA plots of response to therapy for SUV_{mean} for whole tumor (top) and 25% tumor threshold (SUV_{25}) (bottom). **(B)** Similar measures as in (A) but for single high intensity slice. **(C)** BA plots for SUV_{max}, SUV_{P4}, SUV_{P14}, and SUV_{P33}. The red lines represent the 95 % CL and the black solid line is the mean difference line computed from Day 1 Day 2 test-retest data. Open circles represent test-retest data points while filled circles are post-therapy data points. In general, metrics of response to therapy are outside the limits of agreement (test-retest) denoted by LOA.

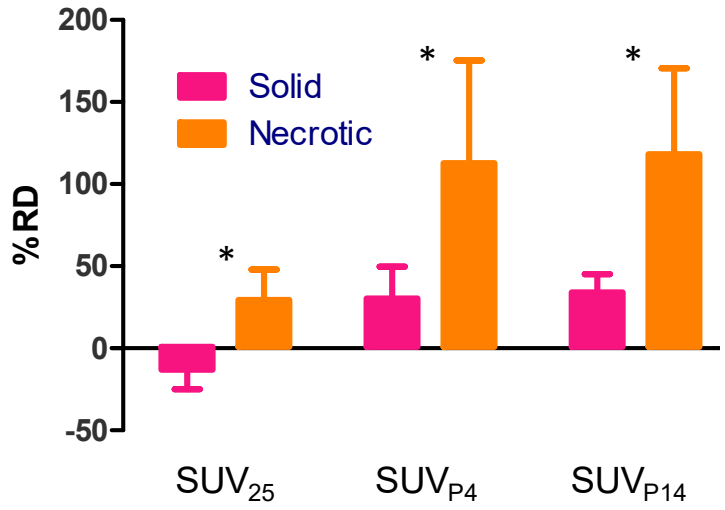


Figure 7. Dynamic range of SUV_{25}, SUV_{P4}, and SUV_{P14} relative to SUV_{mean}. Let *D* denote the difference in SUV between +4d and baseline, i.e., $D = SUV^{+4d} - SUV^{B}$, the percent relative difference $\%RD = 100 * [D_{\#} - D_{mean}]/D_{mean}$ where $D_{\#}$ represents D_{25}, D_{P4}, or D_{P14}. Data represent mean±SEM. *denotes significantly different (ANOVA).

SUPPLEMENTAL METHODS

Generation of TNBC PDX

Gene expression analyses of 93 TNBC PDXs (29657 unique genes/probes) was performed to identify six TNBC subtypes including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem–like (MSL), and a luminal androgen receptor (LAR) subtype (Figure S1) as described previously (*1*). Six to ten-week-old female NSG mice were obtained from The Jacksons Laboratory (https://www.jax.org) and were used for engraftment of human tissue. Mice were anesthetized with isoflurane. An inverted Y−shaped incision was made along the thoracic−inguinal region to expose the mammary glands. Two-to-four million tumor cells mixed with Matrigel in a volume of 30 µl were injected into the 4th inguinal mammary fat pad. The skin was gathered, and the incision closed with wound clips. Following engraftment, tumor growth in PDX mice was monitored.

Figure S1 is on separate page at end of Supplemental document.

Figure S1. Heatmap of the correlation matrix among the 93 PDXs. The heatmap of the correlation matrix among the 93 PDXs was generated with row side color bar indicating subtype (light blue: BL1, dark blue: BL2, red: IM, green: LAR, orange: M, black: MSL, grey: UNS). The column side color bar indicating the PDX lines. The result showed that the PDX of the same lines are highly correlated with each other and mostly belonged to the same TNBC subtype.

Preclinical Studies

Three distinct experiments were carried out. In the first experiment, test-retest studies were performed on consecutive days (Day 1 vs Day 2) to assess the reproducibility of PET image

metrics. Typically, N=8-12 PDX mice for each TNBC subtypes were used in the study. PDX mice were imaged as per the imaging protocol described below. Care was taken to repeat the exact conditions on Day 1 and Day 2 including scanner utilized. Total of 46 PDX mice were used in this cohort. In the second experiment, a separate cohort was used to assess the impact of animal handling/imaging on survival using the study design depicted in Figure 1A. To that end, a separate cohort (N=8; N=16 total) of PDX was administered treatment weekly, but no imaging was performed. Our results suggested that repeat imaging impacted survival (Figure S2), and for that reason we excluded +11d imaging time point from the study design. Previous studies have reported that animal handling has dramatic effects on biodistribution and image metrics of FDG uptake (*2*). This observation has broad implications in developing best practices for therapeutic imaging studies, as it suggests that in designing preclinical therapeutic-imaging protocols, the complexity of a combined therapeutic-imaging study should be kept minimal as to not impact the overall objectives of a given investigation.

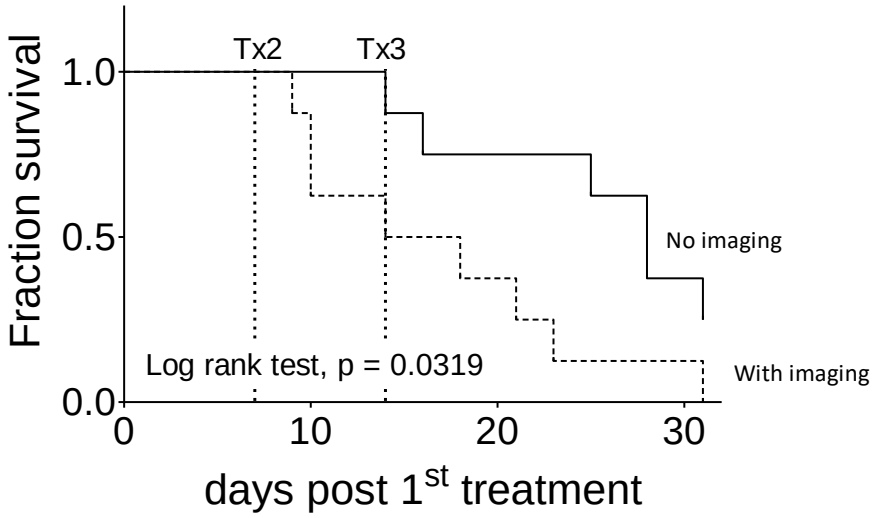


Figure S2. Kaplan–Meier survival curves for two cohort of PDX mice (N=8-10 per group) with weekly combined therapy Docetaxel (20mg/kg I.P.)/Carboplatin

(50mg/kg I.P). One group (dashed line) was imaged at baseline, +4d, and +11d post baseline (per imaging protocol depicted in Scheme 1C). The second cohort (solid line) was not imaged. Caliper measurements were performed bi-weekly. We observed significant differences in the survival of PDX. For this reason, we dropped the +11d imaging time point from our study design.

The third experiment involved a therapeutic arm with imaging. The study design of the therapeutic arm is depicted in Figure 1A. Preclinical imaging was performed at baseline and +4 days following therapy. In all therapeutic studies, the docetaxel (20mg/kg I.P.)/carboplatin (50mg/kg I.P) was administered at baseline (following imaging) and weekly for a period of four weeks. Tumor volumes were measured bi-weekly as a surrogate measure of response to therapy.

Preclinical PET/CT Imaging

Four hours prior to imaging session, food was removed from metabolism cages while water was given ad libitum. Mice were anesthetized with 2-2.5% isoflurane by inhalation via an induction chamber. Anesthesia was maintained throughout the imaging session by delivering 1%–1.5% isoflurane via a custom-designed nose cone. A heat lamp was used to maintain body temperature. Mice were injected with ^{18}FDG (6.66 – 8.14 MBq) by tail vein immediately before a 0-60 min dynamic small animal PET acquisition. Small animal PET images were acquired on the microPET Focus 220 scanner (Concorde Microsystems Inc., Knoxville, TN) or on the Inveon microPET/CT scanner (Siemens Medical Solutions, Washington D.C.), while the CT images were acquired with the Inveon. CT-based attenuation correction was used. PET scanners are cross-calibrated as per the established standard operating procedures outlined at https://c2ir2.wustl.edu/.

Image Analysis

We evaluated thresholds of SUV_{max} ranging from 100% to 0% in 5% increments for each tumor (i_{th}). Please note, threshold of 100% of SUV_{max} amounts to SUV_{max}, and at the limit of threshold of 0%, the resulting image metrics defines SUV_{mean}. Thus, we evaluated 21 image metrics. These 21 image metrics were evaluated for the whole tumor and using a single highest intensity slice of the tumor. In addition, we calculated peak measures of $4mm^3$, $14mm^3$, and $33mm^3$. Thus, overall 43 imaging metrics were evaluated.

Image Histogram Reproducibility Analysis (IHRA). IHRA was performed as percent threshold of SUV_{max}. At 100% threshold, SUV_{100} corresponds to high intensity voxels (or SUV_{max}). At the limit, as the threshold reaches 0%, SUV_0 is identical to SUV_{mean}. Image voxels were used to compute mean SUV above a given threshold. At each percent threshold (Th varies from 100% to 0%), the SUV_{Th} is calculated as percent of SUV_{max}, i.e., $SUV_{Th}=Th*SUV_{max}/100$. SUV_{th} represents the mean of the voxels with SUV greater than SUV_{Th}. At Th=25% for example, the mean of voxels >= $SUV_{25}=0.25*SUV_{max}$ is calculated. Therefore, as the %Th decreases, the volume of the tumor region under consideration increases with the addition of lower intensity voxels. At each threshold, the mean of the voxels at the threshold is computed by taking the average over all the voxels in the defined tumor region/threshold. This process is repeated for the whole tumor as well as for the metabolically active tumor region in each mouse that is being investigated for tumor reproducibility studies.

Analysis of single slice. In an effort to facilitate analysis, results obtained from whole tumor analysis were compared to those obtained from single slice (SS). A single slice (SS) with the maximum mean activity over the slice (the hottest slice) was selected for processing to investigate the reproducibility of the data. Here also, I_{th} was used to define the tumor region and the hottest

slice data was processed following the same procedure as discussed earlier in the case of whole tumor volume data to compute different thresholds of interest.

SUV_{peak} analyses. SUV_{peak} denotes the mean of all the voxels of in a sphere centered at the hottest voxel. Three different spherical volumes of ~ 4 mm^3 (SUV_{P4}), 14 mm^3 (SUV_{P14}), and 33 mm^3 (SUV_{P33}) were considered corresponding to spheres of radius of 1, 2, 3 voxels. The SUV_{Peak} values were further investigated in the reproducibility and treatment response studies; first to compute the limits of agreement (LOA) and later to evaluate their performance in assessing the response to therapy.

Evaluation of preclinical PERCIST (µPERCIST). The tumor threshold based on the PERCIST criteria (*3*) is provided by Th=*α*[mean concentration of liver ROI] + β*[standard deviation of liver ROI]*. Liver ROIs were determined 50-60min post injection of FDG. Optimization of α and β entails maximizing Lin's concordance correlation coefficient (LCC) while minimizing the repeatability coefficient (RC) (which would minimize the 95% CI, hence maximize reproducibility); thus the objective function to maximize is the ratio of LCC (*4*) to the RC (defined in Supplementary Statistics section). A range of values for α and β were evaluated and optimized. Implementation of *µPERCIST* relies on evaluation of SUV_{peak} which was described earler.

Statistical Analysis

Reproducibility Statistics. Let Δ denote the within mouse difference between the measurements, and N denote the number of paired measurements. The standard deviation for the mean difference is calculated using Eq. 1, and the within-mouse standard deviation (wSD), using Eq. 2.

$$dsd = \sqrt{\frac{\sum(\Delta_i)^2}{N}} \qquad \text{Eq. 1}$$

$$wSD = \frac{dsd}{\sqrt{2}} \quad \text{Eq. 2}$$

The 95% confidence limits (95% CL) in the BA plots are the limits of agreement (LOA) defined as the mean difference ± the repeatability coefficient (RC), defined in Eq.3. These limits are independent of the sample size so that the results from an individual test-retest experiment is expected to fall within these boundaries 95% of the time.

$$RC = 1.97 \times \sqrt{2} \times wSD = 2.77 \times wSD \quad \text{Eq. 3}$$

Two methods for assessing reproducibility were used, Lin's concordance correlation coefficient (LCC) (*4*) and Bland-Atlman plots (BA) (*5*). The LCC, being the product of the Pearson correlation coefficient (PCC) and the bias correction factor (BCF), accounts for both precision and accuracy. The method outlined in Watson and Petrie (*6*) was followed to calculate these metrics. The procedure used to calculate the statistical parameter for the BA plots are summarize in Galbraith (*7*) and Raunig (*8*).

Performance analysis of image metrics response to therapy. Sensitivity, the number of positive responses that are correctly classified as positive; Specificity, the number of negative responses that are correctly classified as negative; Precision, the probability that a prediction of positive is actually positive; Negative predictive value (NPV), the probability that a prediction of negative is actually negative; Accuracy, the fraction of correct prediction to the total number of observation; and F-score, the harmonic mean of precision and sensitivity, are the standard performance binary classification metrics used to assess the response to the therapy (*9,10*). The evaluations were categorized as; True Positive (TP) when outcome was a positive response (True) and SUV change also predicted a positive response (True). False Negative (FN) when outcome truth was

a positive result (True), but SUV change predicted a non-response (False). True Negative (TN) when outcome showed a nonresponse (False), and the SUV change also predicted a nonresponse. False Positive (FP) when outcome is a nonresponse, but the SUV change predicts positive response (True). If an image metrics of response to therapy was within the LOA, it was considered indistinguishable from metric variability. In that scenario, the image metric was not used in calculating performance.

SUPPLEMENTAL RESULTS

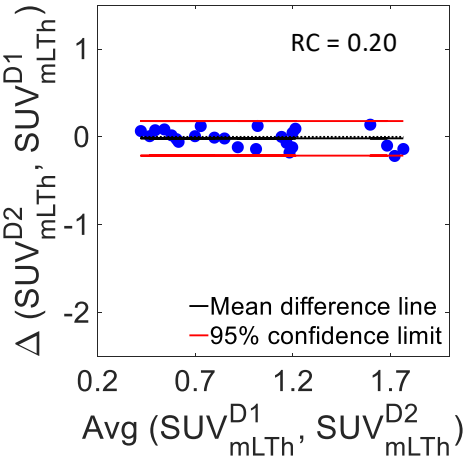


Figure S3. Tumor SUV BA plot for optimized liver threshold.

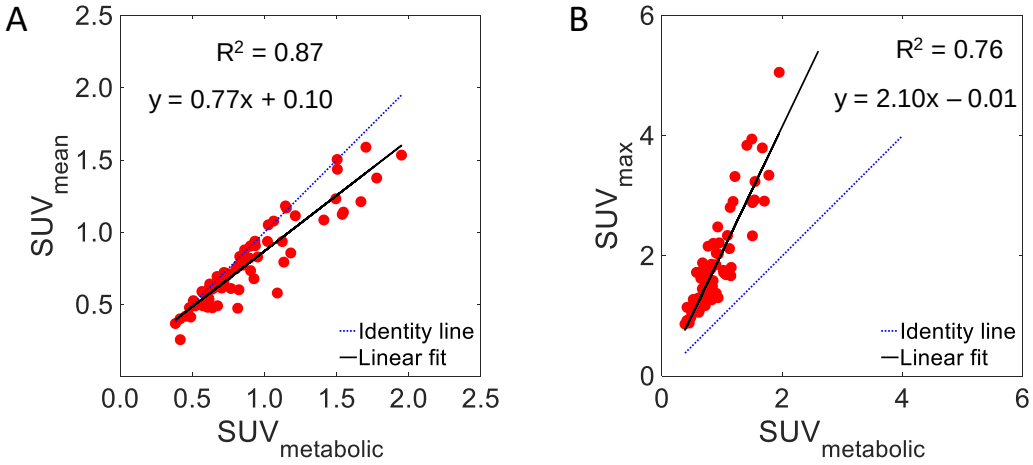


Figure S4. Correlation between SUV_{mean} and SUV_{max} to SUV of metabolic tumor ($SUV_{metabolic}$).

The performance of image metrics is tabulated in Supplementary Table S1. The accuracy of image metrics by subtype is tabulated in Table S2.

TABLE S1A. Performance of imaging metrics to predict response to therapy for data points outside the LOA.

SUV metric	Sensitivity	Specificity	Precision	NPV	Accuracy	F-Score	Uncertain Fraction (%)
ΔSUV_{max}	1.00	0.25	0.57	1.00	0.63	0.73	45
ΔSUV_{mean}	0.92	0.22	0.63	0.67	0.64	0.75	24
ΔSUV_{25}	0.91	0.22	0.59	0.67	0.60	0.72	31
ΔSUV_{mean} (SS)	0.71	0.11	0.56	0.20	0.48	0.63	21
ΔSUV_{25} (SS)	0.91	0.25	0.63	0.67	0.63	0.74	34
ΔSUV_{P4}	1.00	0.29	0.62	1.00	0.67	0.77	48
ΔSUV_{P14}	0.91	0.25	0.63	0.67	0.63	0.74	34
ΔSUV_{P33}	0.89	0.25	0.57	0.67	0.59	0.69	41

TABLE **S1B**. Performance of imaging metrics to predict response to therapy for all data points (29 samples)

SUV metric	Sensitivity	Specificity	Precision	NPV	Accuracy	F score
ΔSUV_{max}	0.74	0.30	0.67	0.38	0.59	0.70
ΔSUV_{mean}	0.84	0.30	0.70	0.50	0.66	0.76
ΔSUV_{25}	0.79	0.30	0.68	0.43	0.62	0.73
ΔSUV_{mean} (SS)	0.74	0.10	0.61	0.17	0.52	0.67
ΔSUV_{25} (SS)	0.74	0.30	0.67	0.38	0.59	0.70
ΔSUV_{P4}	0.79	0.30	0.68	0.43	0.62	0.73
ΔSUV_{P14}	0.89	0.30	0.71	0.60	0.69	0.79
ΔSUV_{P33}	0.79	0.30	0.68	0.43	0.62	0.73

All performance metrics range from 0 (no prediction) to 1 (high). Data points within the LOA were penalized for uncertainly by exclusion from the analysis. Refer to supplementary Table S1 for performance analysis inclusive of data points within the LOA. Uncertain Fraction is the percent of studies within the LOA which were not used in prediction (due to uncertainty) for each image metric.

TABLE S2A: Accuracy of prediction by PDX subtype* (for data points outside the LOA)

WHIM	SUV_{mean}	SUV_{25}	SUV_{max}	SUV_{mean} (SS)	SUV_{25} (SS)	SUV_{P4}	SUV_{P14}	SUV_{P33}	SUV_{P64}
IM	1.00	1.00	1.00	0.50	1.00	1.00	1.00	1.00	1.00
BL1	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00
BL2	0.55	0.50	0.50	0.50	0.50	0.44	0.44	0.38	0.38
M	0.33	0.33	0.50	0.33	0.50	1.00	0.50	0.50	0.50

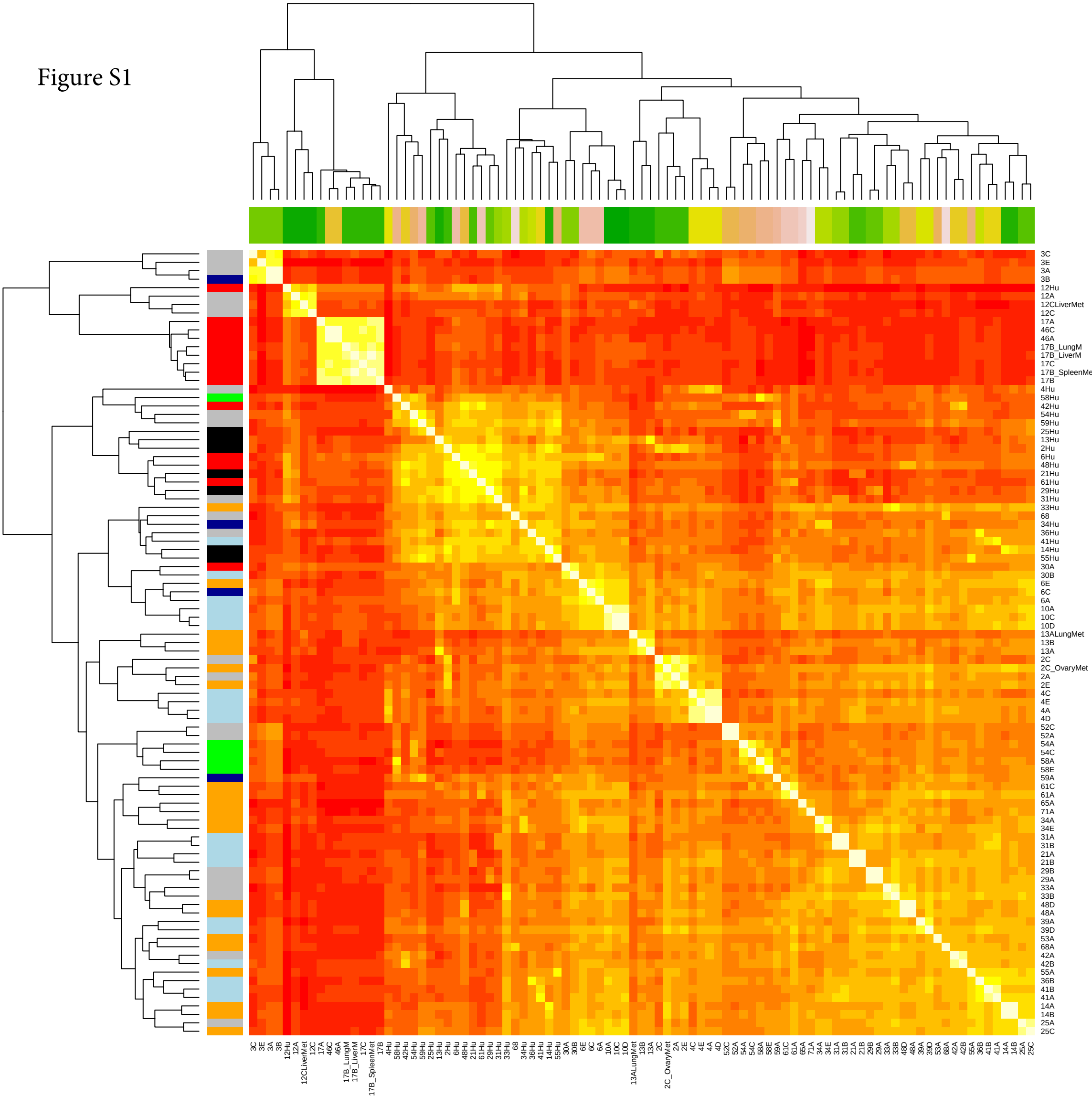
*Accuracy = (TP + TN)/(P+N); LAR PDX are not included due to low sample count (N=4).

TABLE S2B: Accuracy for each PDX* (for data points inclusive of within LOA)

PDX	SUV_{mean}	SUV_{25}	SUV_{max}	SUV_{mean} (SS)	SUV_{25} (SS)	SUV_{P4}	SUV_{P14}	SUV_{P33}	SUV_{P64}
IM	0.75	0.75	0.63	0.63	0.50	0.75	0.88	0.75	0.75
BL1	1.00	0.75	1.00	0.50	1.00	1.00	1.00	1.00	1.00
BL2	0.62	0.62	0.54	0.54	0.62	0.54	0.62	0.54	0.54
M	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33

*Accuracy = (TP + TN)/(P+N); LAR PDX are not included due to low sample count (N=4).

Figure S1


3C
3E
3A
3B
12Hu
12A
12CLiverMet
12C
17A
46C
46A
17B_LungM
17B_LiverM
17C
17B_SpleenMet
17B
4Hu
58Hu
42Hu
54Hu
59Hu
25Hu
13Hu
2Hu
6Hu
48Hu
21Hu
61Hu
29Hu
31Hu
33Hu
68
34Hu
36Hu
41Hu
14Hu
55Hu
30A
30B
6E
6C
6A
10A
10C
10D
13ALungMet
13B
13A
2C
2C_OvaryMet
2A
2E
4C
4E
4A
4D
52C
52A
54A
54C
58A
58E
59A
61C
61A
65A
71A
34A
34E
31A
31B
21A
21B
29B
29A
33A
33B
48D
48A
39A
39D
53A
68A
42A
42B
55A
36B
41B
41A
14A
14B
25A
25C

References

1. Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011;121:2750-2767.

2. Fueger BJ, Czernin J, Hildebrandt I, et al. Impact of animal handling on the results of 18F-FDG PET studies in mice. *J Nucl Med.* 2006;47:999-1006.

3. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50 Suppl 1:122S-150S.

4. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255-268.

5. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135-160.

6. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology.* 2010;73:1167-1179.

7. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR Biomed.* 2002;15:132-142.

8. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res.* 2015;24:27-67.

9. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology.* 2016;4:320-330.

10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10:e0118432.