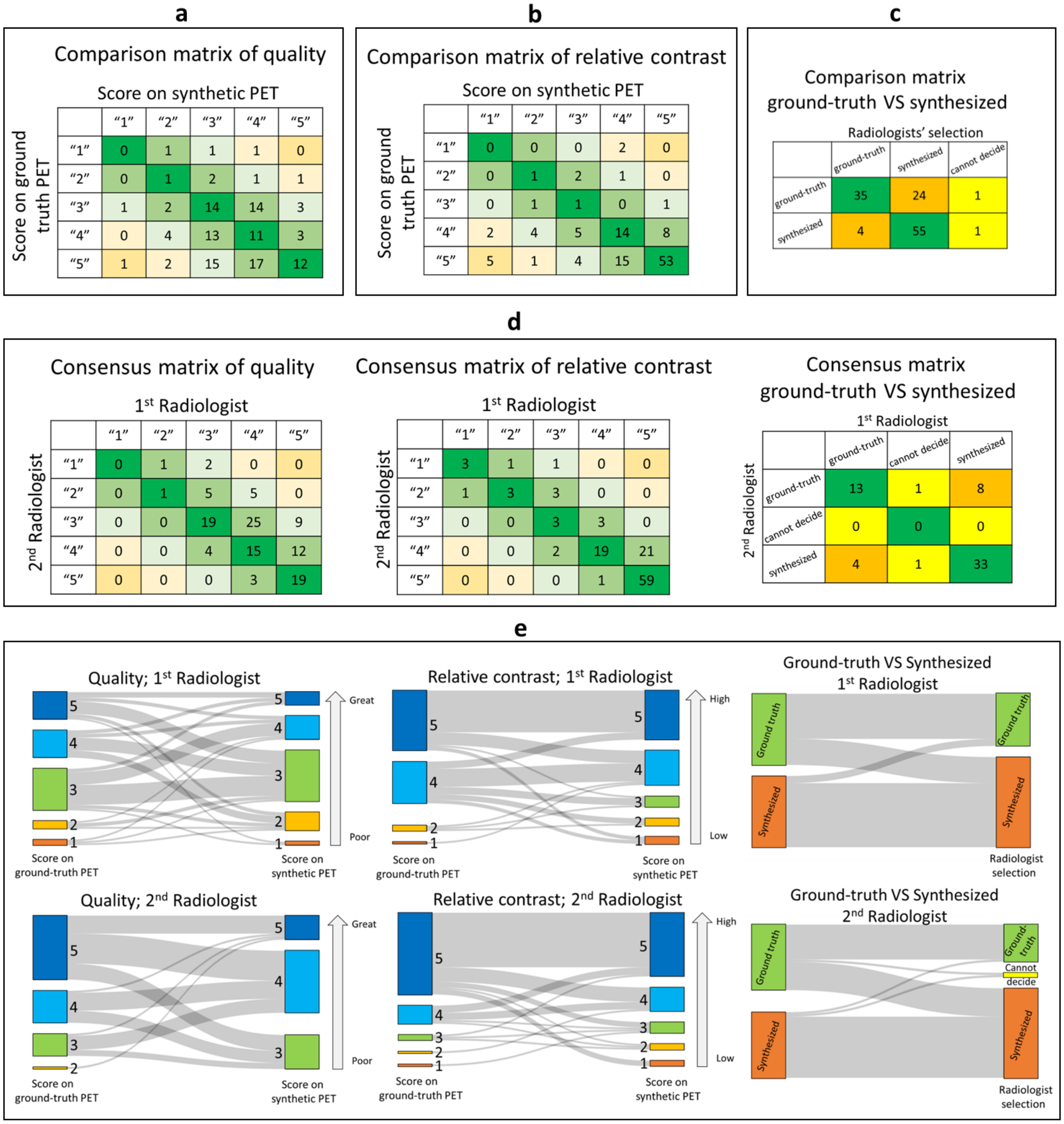
Cell Reports Medicine, Volume *5*

Supplemental information

Synthetic PET from CT improves diagnosis and prognosis for lung cancer: Proof of concept

Morteza Salehjahromi, Tatiana V. Karpinets, Sheeba J. Sujit, Mohamed Qayati, Pingjun Chen, Muhammad Aminu, Maliazurina B. Saad, Rukhmini Bandyopadhyay, Lingzhi Hong, Ajay Sheshadri, Julie Lin, Mara B. Antonoff, Boris Sepesi, Edwin J. Ostrin, Iakovos Toumazis, Peng Huang, Chao Cheng, Tina Cascone, Natalie I. Vokes, Carmen Behrens, Jeffrey H. Siewerdsen, John D. Hazle, Joe Y. Chang, Jianhua Zhang, Yang Lu, Myrna C.B. Godoy, Caroline Chung, David Jaffray, Ignacio Wistuba, J. Jack Lee, Ara A. Vaporciyan, Don L. Gibbons, Gregory Gladish, John V. Heymach, Carol C. Wu, Jianjun Zhang, and Jia Wu

Supplementary figures



a Comparison matrix of quality

Score on synthetic PET (columns); Score on ground truth PET (rows)

	"1"	"2"	"3"	"4"	"5"
"1"	0	1	1	1	0
"2"	0	1	2	1	1
"3"	1	2	14	14	3
"4"	0	4	13	11	3
"5"	1	2	15	17	12

b Comparison matrix of relative contrast

Score on synthetic PET (columns); Score on ground truth PET (rows)

	"1"	"2"	"3"	"4"	"5"
"1"	0	0	0	2	0
"2"	0	1	2	1	0
"3"	0	1	1	0	1
"4"	2	4	5	14	8
"5"	5	1	4	15	53

c Comparison matrix ground-truth VS synthesized

Radiologists' selection

	ground-truth	synthesized	cannot decide
ground-truth	35	24	1
synthesized	4	55	1

d Consensus matrix of quality

1st Radiologist (columns); 2nd Radiologist (rows)

	"1"	"2"	"3"	"4"	"5"
"1"	0	1	2	0	0
"2"	0	1	5	5	0
"3"	0	0	19	25	9
"4"	0	0	4	15	12
"5"	0	0	0	3	19

Consensus matrix of relative contrast

1st Radiologist (columns); 2nd Radiologist (rows)

	"1"	"2"	"3"	"4"	"5"
"1"	3	1	1	0	0
"2"	1	3	3	0	0
"3"	0	0	3	3	0
"4"	0	0	2	19	21
"5"	0	0	0	1	59

Consensus matrix ground-truth VS synthesized

1st Radiologist (columns); 2nd Radiologist (rows)

	ground-truth	cannot decide	synthesized
ground-truth	13	1	8
cannot decide	0	0	0
synthesized	4	1	33

Figure S1. Radiologists assessment on synthetic PET. Related to Figure 2. (a) Comparison matrix of imaging quality. (b) Comparison matrix of lesion contrast evaluation. (c) Comparison matrix of Turing test. (d) From left to right, consensus matrices between 2 radiologists for imaging quality, lesion contrast evaluation, and catching the synthetic scans, respectively. (e) Individual radiologists' performance for individual task.

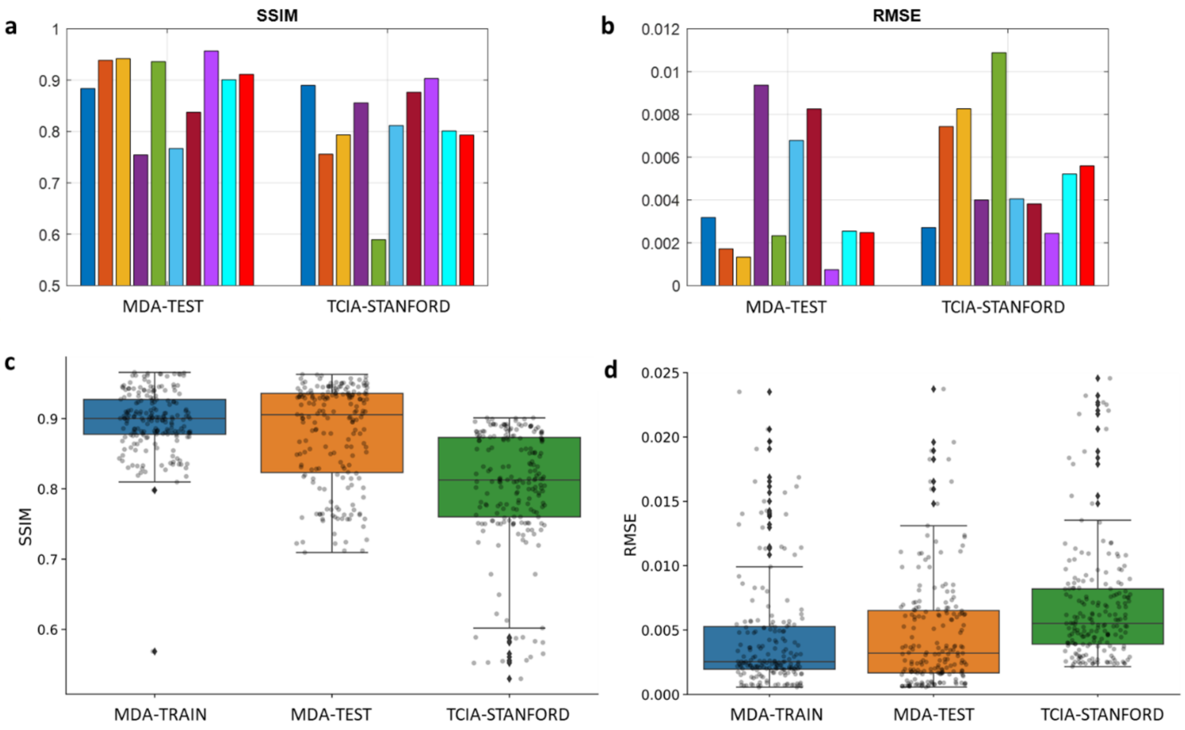


Figure S2. Quantitative assessment of signal fidelity of synthetic PET. Related to Figure 2. The SSIM (a) and RMSE (b) for true and synthetic PET images in Fig 2a. From left to right, the ten measurements for each cohort correspond to the ten pair of true and synthetic PET images from top to bottom i.e. i-x in figure 2a. The SSIM (c) and RMSE (d) for 195 sampled lung region slices of true and synthetic PET images for the MDA-TRAIN, MDA-TEST and TCIA-STANFORD cohorts.

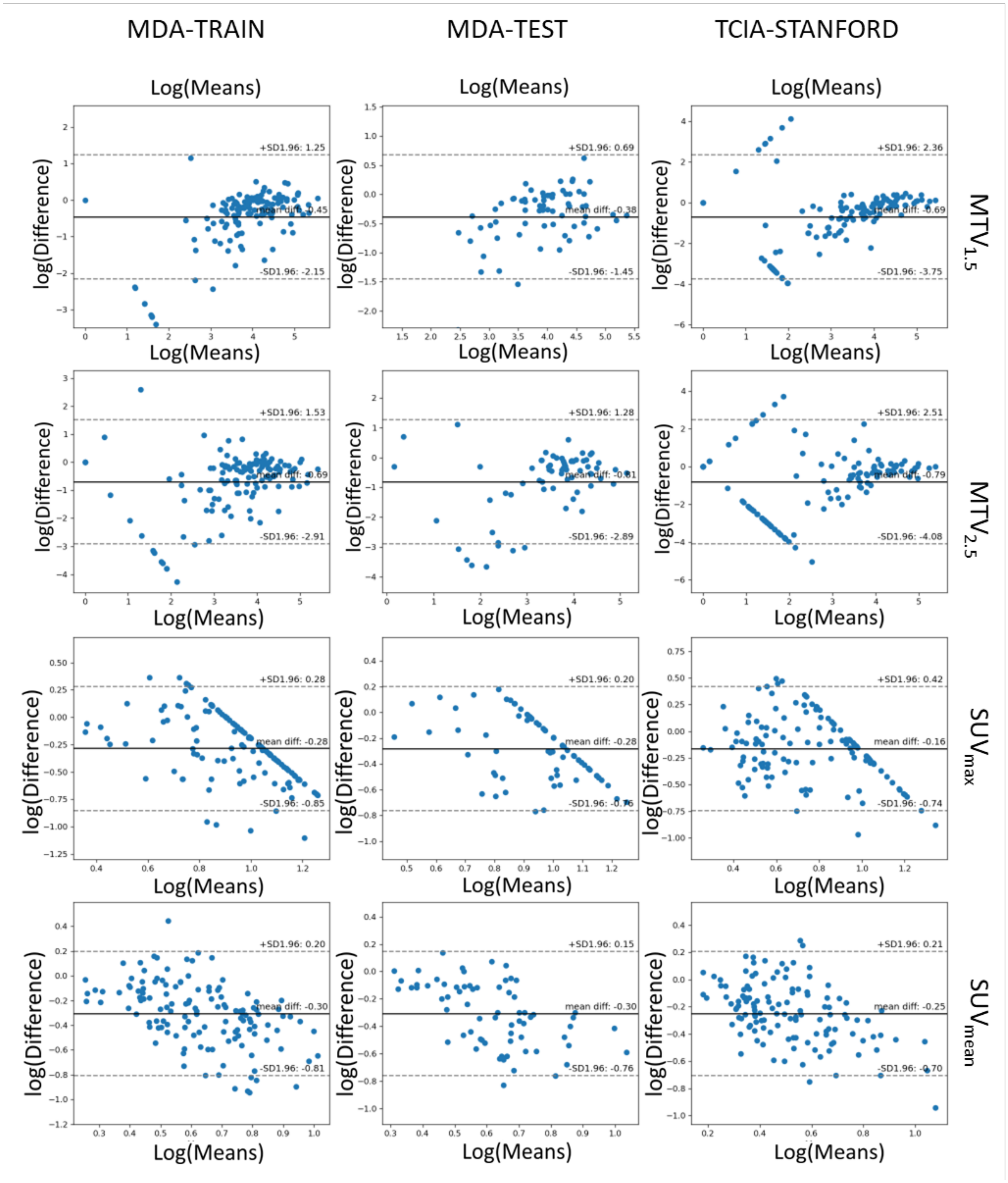


Figure S3. Assess imaging feature fidelity on synthetic PET. Related to Figure 3. Bland-Altman plots for agreement analysis between the ground-truth and synthetic PET features.

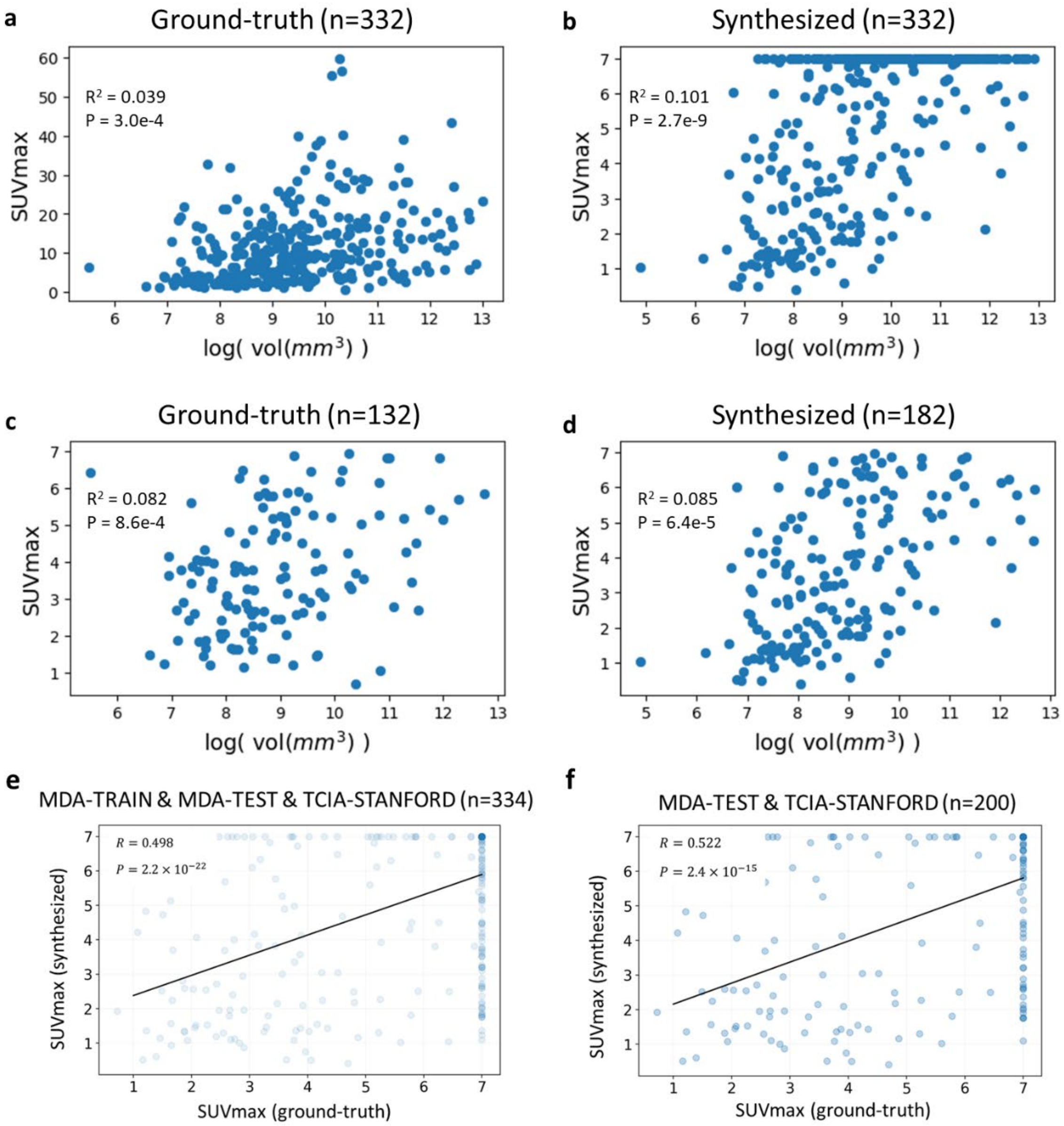


Figure S4. Assessment correlation between tumor volume and SUVmax, as well as synthetic SUVmax and ground-truth SUVmax. Related to Figure 3. (a) and (b) are the scatter plots of SUVmax vs. volume for the ground-truth and synthetic PET on the combined cohorts of MDA-TRAIN, MDA-TEST and TCIA-STANFORD. (c) and (d) are the scatter plots of SUVmax vs. volume on cases with SUVmax<7 for ground-truth and synthetic PET, respectively. (e) and (f) depict the scatter plots between ground-truth and synthetic SUVmax values in the combined cohorts (MDA-TRAIN, MDA-TEST, and TCIA-STANFORD) and the test cohorts (MDA-TEST and TCIA-STANFORD) where transparent dots represent fewer points, while solid dots means more data points close together.

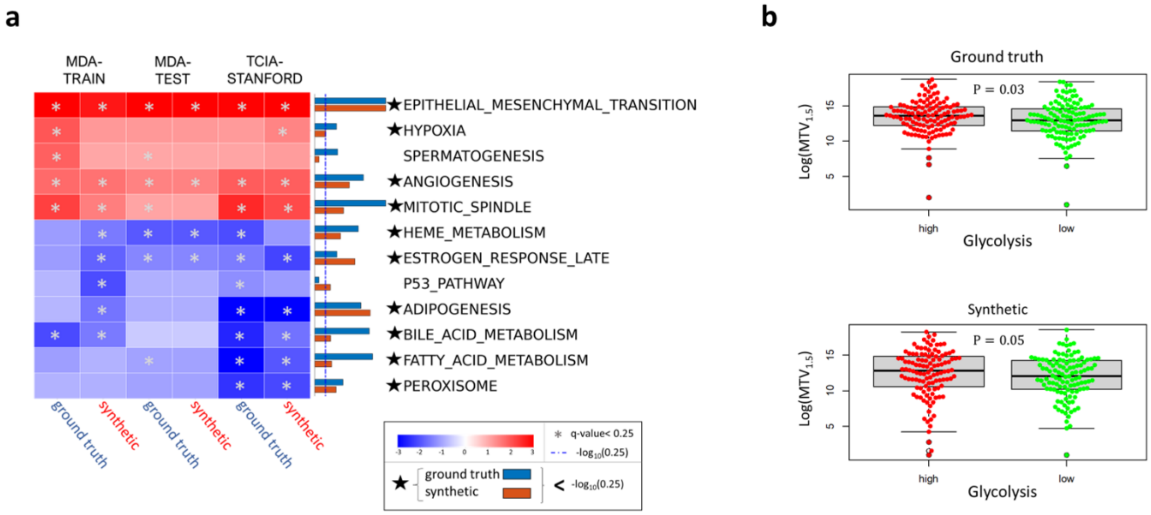


Figure S5. Biological correlates of imaging feature with Cancer Hallmark pathways and glycolytic score. Related to Figure 4. (a) The first column shows the unsupervised hierarchical clustering heatmap of up- and down-regulated Hallmark pathways normalized enrichment score (NES) for correlation of each pathway with MTV feature from true and synthetic PET across MDA-TRAIN, TCIA-STANFORD and MDA-TEST cohorts, where the * represent the significant FDR q-value>0.25. The second column barplot is the -$\log_{10}$() transform of combined q-values obtained by using Fisher's method from all three true and predicted q-values in MDA-TRAIN, MDA-TEST and TCIA-STANFORD cohorts. (b) boxplots show the MTV from synthetic and true PETs distributed for glycolysis high versus low groups.

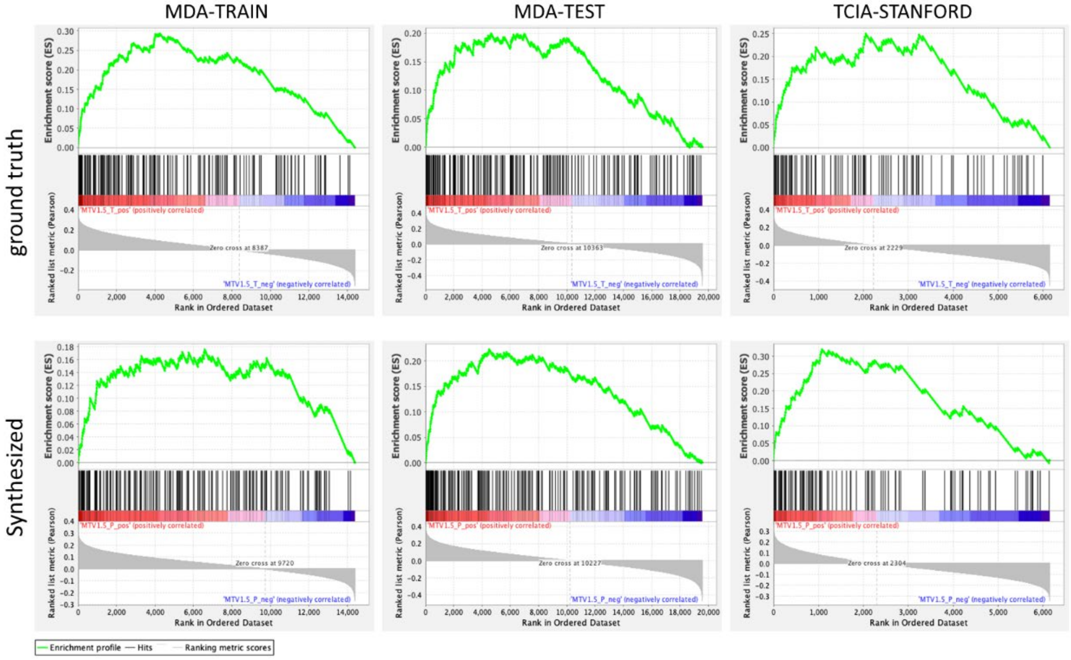


Figure S6. Biological correlates of imaging feature with EMT pathway. Related to Figure 4. The enrichment plots of EMT hallmark based on synthetic and true $MTV_{1.5}$ feature for MDA-TRAIN, MDA-TEST and TCIA-STANFORD cohorts.

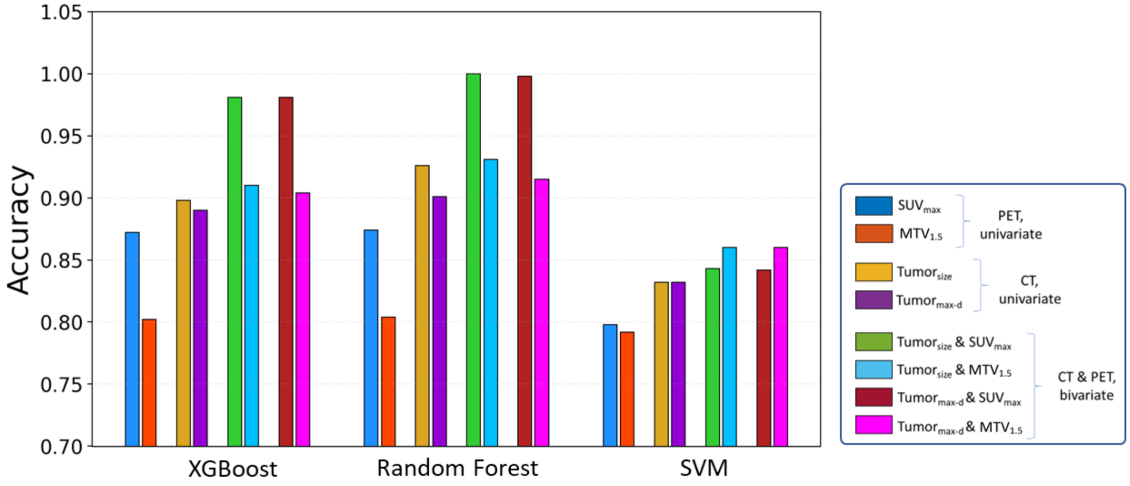


Figure S7. Clinical value by diagnosing malignant versus benign from indeterminant pulmonary nodules during model training from synthetic PET. Related to Figure 5. Model accuracy in the training cohort (n=1048) correspond to Fig. 5a.

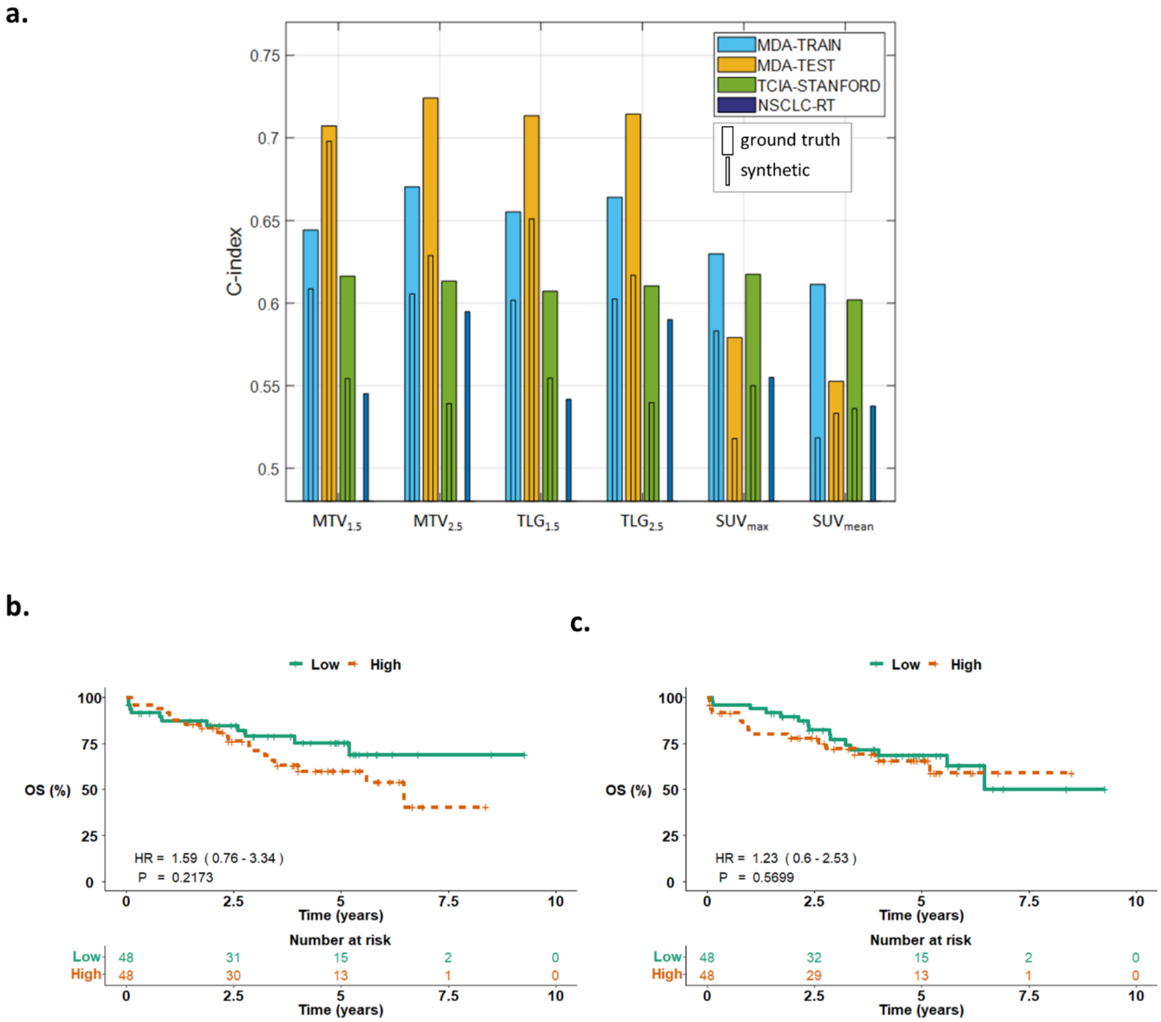


Figure S8. prognostic value of individual features and Kaplan-Meier curves in patients with low synthetic PET correlation. Related to Figure 7. (a) Comparison of prognostic value for individual features between ground-truth PET and synthetic PET. The C-index of overall survival (OS) stratified by different features obtained from the true and synthetic PETs on MDA-TRAIN, MDA-TEST, TCIA-STANFORD and NSCLC-RT cohorts. Of note, the NSCLC-RT cohort does not possess the true PET. (b-c) The Kaplan-Meier curves of patients' overall survival (OS) on the combined MDA Test and TCIA-STANFORD datasets stratified by $MTV_{1.5}$ and SUVmax features. This subset of patients was selected from the lower half of the group showing lower correlation between their predicted SUVmax values and the corresponding ground-truth ones.

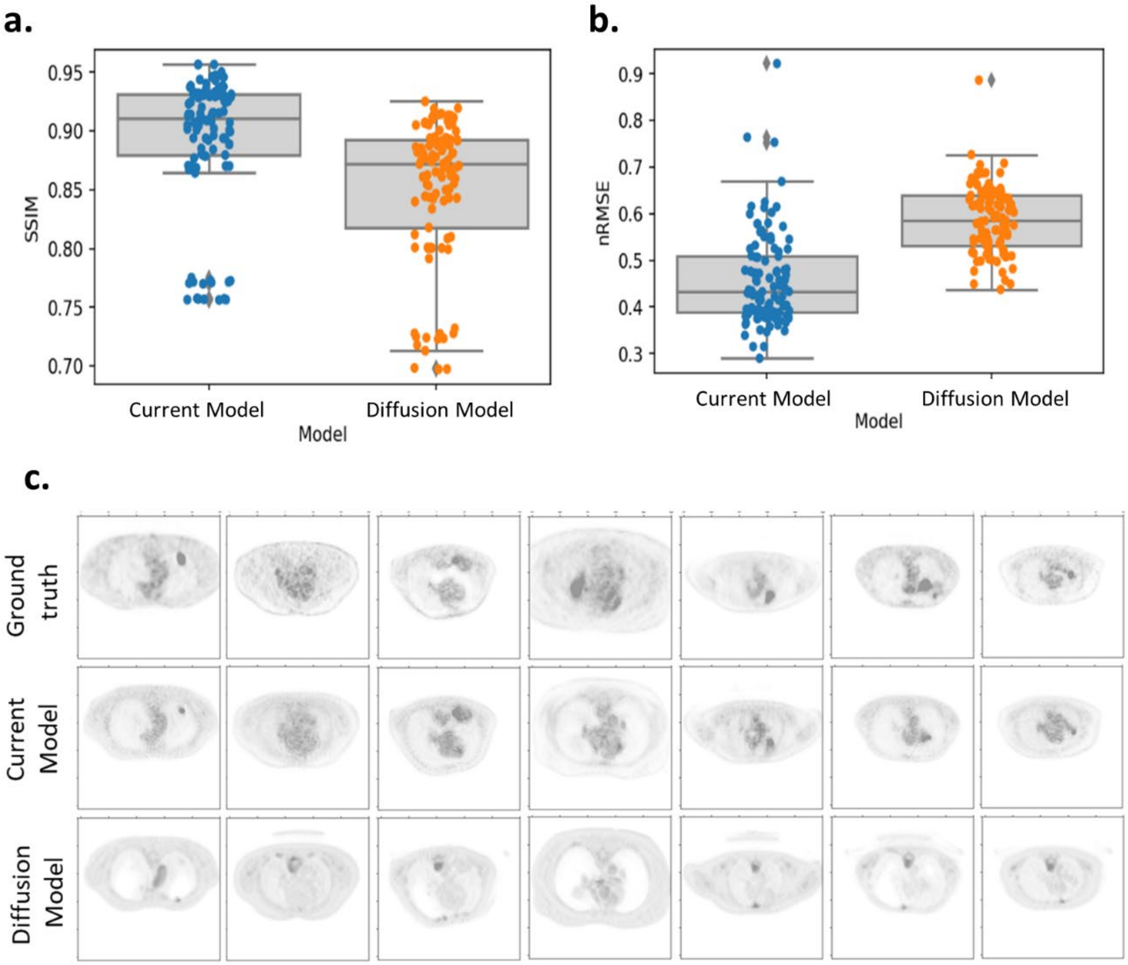
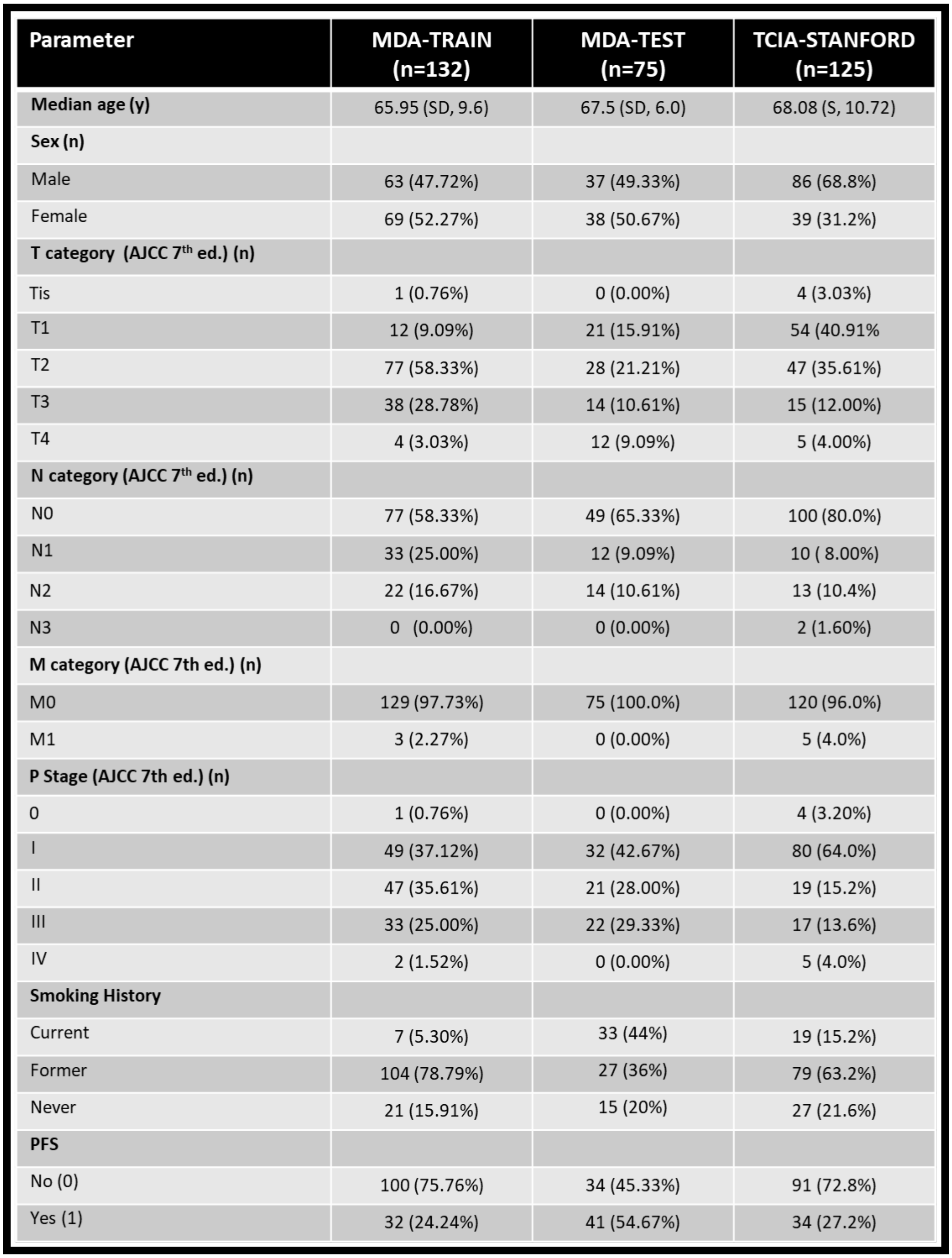


Figure S9. Comparative performance of cGAN and diffusion models: SSIM and nRMSE Metrics Analysis. Related to Figure 2. SSIM (a) and nRMSE(b) indices in comparing the current cGAN model and a diffusion model. (C) Visual comparison of between the current and the diffusion model.

Supplementary tables

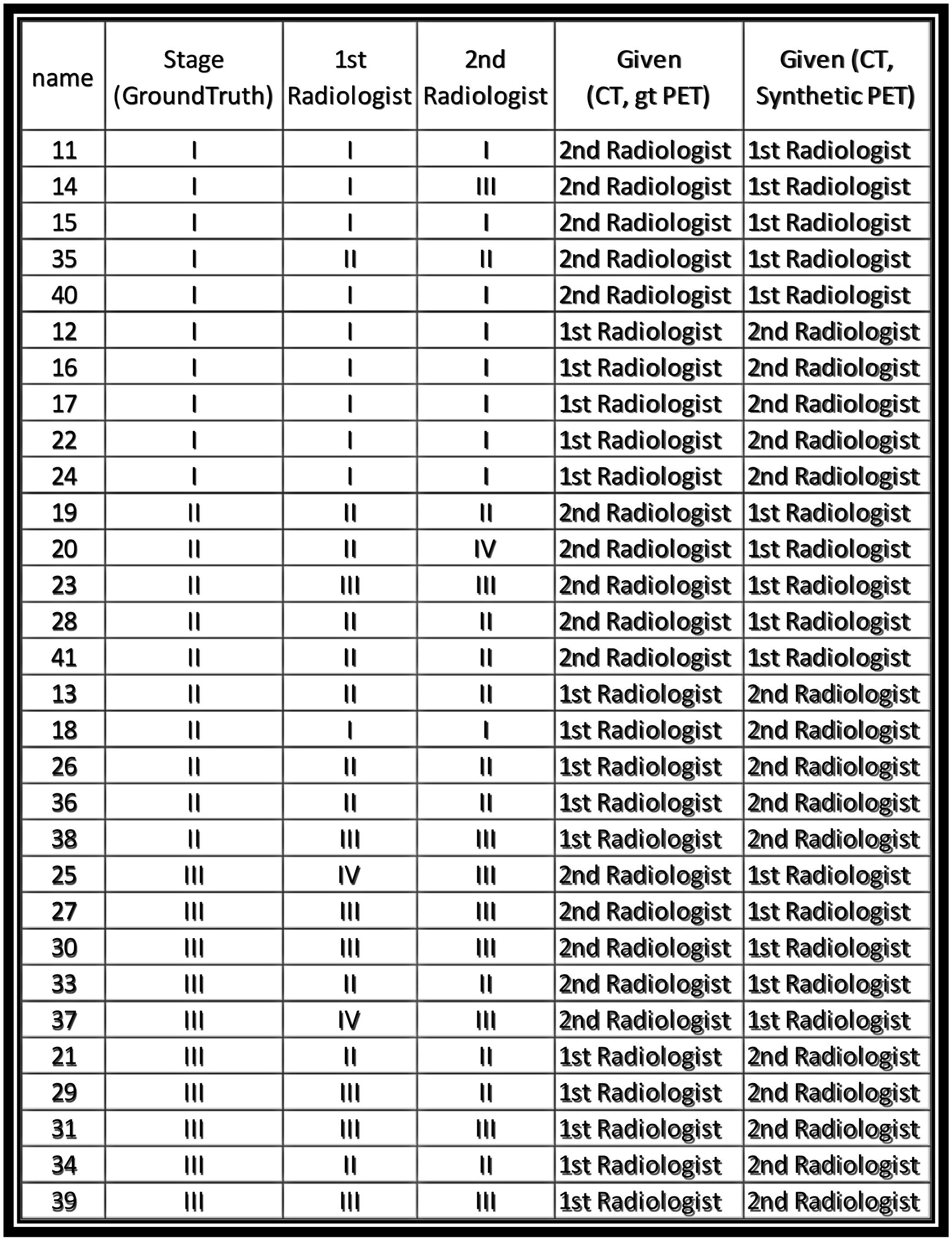
Table S1. (a) Clinical characteristics of MDA-TRAN, MDA-TEST and TCIA-STANFORD cohorts. (b) Clinical pathological staging of 30 selected patients from MDA-TEST along with the radiologists' prediction. Related to Figure 2.

a



Parameter	MDA-TRAIN (n=132)	MDA-TEST (n=75)	TCIA-STANFORD (n=125)
Median age (y)	65.95 (SD, 9.6)	67.5 (SD, 6.0)	68.08 (S, 10.72)
Sex (n)			
Male	63 (47.72%)	37 (49.33%)	86 (68.8%)
Female	69 (52.27%)	38 (50.67%)	39 (31.2%)
T category (AJCC 7th ed.) (n)			
Tis	1 (0.76%)	0 (0.00%)	4 (3.03%)
T1	12 (9.09%)	21 (15.91%)	54 (40.91%
T2	77 (58.33%)	28 (21.21%)	47 (35.61%)
T3	38 (28.78%)	14 (10.61%)	15 (12.00%)
T4	4 (3.03%)	12 (9.09%)	5 (4.00%)
N category (AJCC 7th ed.) (n)			
N0	77 (58.33%)	49 (65.33%)	100 (80.0%)
N1	33 (25.00%)	12 (9.09%)	10 (8.00%)
N2	22 (16.67%)	14 (10.61%)	13 (10.4%)
N3	0 (0.00%)	0 (0.00%)	2 (1.60%)
M category (AJCC 7th ed.) (n)			
M0	129 (97.73%)	75 (100.0%)	120 (96.0%)
M1	3 (2.27%)	0 (0.00%)	5 (4.0%)
P Stage (AJCC 7th ed.) (n)			
0	1 (0.76%)	0 (0.00%)	4 (3.20%)
I	49 (37.12%)	32 (42.67%)	80 (64.0%)
II	47 (35.61%)	21 (28.00%)	19 (15.2%)
III	33 (25.00%)	22 (29.33%)	17 (13.6%)
IV	2 (1.52%)	0 (0.00%)	5 (4.0%)
Smoking History			
Current	7 (5.30%)	33 (44%)	19 (15.2%)
Former	104 (78.79%)	27 (36%)	79 (63.2%)
Never	21 (15.91%)	15 (20%)	27 (21.6%)
PFS			
No (0)	100 (75.76%)	34 (45.33%)	91 (72.8%)
Yes (1)	32 (24.24%)	41 (54.67%)	34 (27.2%)

b



name	Stage (GroundTruth)	1st Radiologist	2nd Radiologist	Given (CT, gt PET)	Given (CT, Synthetic PET)
11	I	I	I	2nd Radiologist	1st Radiologist
14	I	I	III	2nd Radiologist	1st Radiologist
15	I	I	I	2nd Radiologist	1st Radiologist
35	I	II	II	2nd Radiologist	1st Radiologist
40	I	I	I	2nd Radiologist	1st Radiologist
12	I	I	I	1st Radiologist	2nd Radiologist
16	I	I	I	1st Radiologist	2nd Radiologist
17	I	I	I	1st Radiologist	2nd Radiologist
22	I	I	I	1st Radiologist	2nd Radiologist
24	I	I	I	1st Radiologist	2nd Radiologist
19	II	II	II	2nd Radiologist	1st Radiologist
20	II	II	IV	2nd Radiologist	1st Radiologist
23	II	III	III	2nd Radiologist	1st Radiologist
28	II	II	II	2nd Radiologist	1st Radiologist
41	II	II	II	2nd Radiologist	1st Radiologist
13	II	II	II	1st Radiologist	2nd Radiologist
18	II	I	I	1st Radiologist	2nd Radiologist
26	II	II	II	1st Radiologist	2nd Radiologist
36	II	II	II	1st Radiologist	2nd Radiologist
38	II	III	III	1st Radiologist	2nd Radiologist
25	III	IV	III	2nd Radiologist	1st Radiologist
27	III	III	III	2nd Radiologist	1st Radiologist
30	III	III	III	2nd Radiologist	1st Radiologist
33	III	II	II	2nd Radiologist	1st Radiologist
37	III	IV	III	2nd Radiologist	1st Radiologist
21	III	II	II	1st Radiologist	2nd Radiologist
29	III	III	II	1st Radiologist	2nd Radiologist
31	III	III	III	1st Radiologist	2nd Radiologist
34	III	II	II	1st Radiologist	2nd Radiologist
39	III	III	III	1st Radiologist	2nd Radiologist

Table S2. Score our study based on AI-based algorithms development criteria. Related to Figure 1.

Category	Score on Topics	More explanation
Study design	#Task definition ✓ #Study type ✓	#Multi-disciplinary team including radiologists, oncologists and computational scientists collaborated. #Related publications and studies were identified.
Data collection	#Bias anticipation ✓ #Data labeling ✓	#Different cohorts were gathered which are vulnerable to bias. #Tumor or nodule segmentation was carried out by radiologists within our institution when required.
Model design, training and testing	#Cross validation ✓ #Model comparison ✓ #Model selection ✓ #Data leakage ✓ #Use of external datasets ✓ #Evaluation metric ✓	#The study utilized a 5-fold cross-validation, involving 120 patients for training and 12 patients for validation. #Model selection and the final hyperparameters of the GAN model were provided. #No information leaks from test sets during model training. #There are 2 and 5 different external cohorts used for imaging and clinical validations (n=200 in imaging test) and (n=1346 in clinical test). #Different evaluation metrics have been utilized such as imaging quality indices, radiologists' validation, radiogenomics and clinical validations.
Reporting and dissemination	# Reproducibility, accessibility of code, models✓ #Transparency ✓	#The corresponding codes are shared. #The limitation of our study has been widely discussed in different parts of the paper by comparing to the predictions made by ground-truth PET.

Table S3. Score our study based on the evaluation criteria of AI-based algorithms. Related to Figure 1.

Class of evaluation	**Score on Topics**	**More explanation**
Proof of concept evaluation	#Ensure no overlap between development & testing cohort. ✓ #Check that ground-truth quality is reasonable. ✓ #Provide comparison with state-of-the-art methods. ✓ #Choose figures of merit that motivate further evaluation. ✓	#There were 5 different external cohorts used for testing (n=1346). #The ground-truth predictions were checked to be reasonable. #The difference in performance of the AI-based method with ground-truth was demonstrated and its limitations were discussed.
Technical task-specific evaluation	#Choose clinically relevant tasks and determine the right clinical study type. ✓ #Testing cohort should be external. ✓ #Reference standard should be high quality and correspond to the task. ✓ #Use a reliable strategy to extract task-specific information. ✓ #Choose figures of merit that quantify task performance. ✓	#Our method yields reasonable and correlated MTV, TLG and SUVmax values compared with ground truth. #Imaging quality indices along with radiologist assessment were utilized. #Radiogenomic analysis found reasonable association of cancer Hallmarks with extracted MTV features from ground-truth and synthetic PET scans. #Reference standards are based on the ground truth PET images. #There were 2 different external cohorts used for technical test evaluation (n=200).
Clinical evaluation	#Efficiency in making clinical predictions. ✓ #Testing cohort must be external. ✓ #Reference standard should be high quality and be representative of those used for clinical decision making. ✓ #Figure of merit should reflect performance on clinical decision making. ✓	#MTV values from synthetic PET can predict overall survival. #Features obtained from synthetic PET can improve the prediction of lung cancer development. #There were 5 different external cohorts used for testing (n=1346). #Reference standards are based on the derived features from the ground truth PET images. #The difference in performance of the AI-based method with ground-truth was demonstrated and its limitations were discussed.