

Interobserver and Intraobserver Variability among Measurements of FDG PET/CT Parameters in Pulmonary Tumors

Gülgün Büyükdereli¹, Mehtap Güler², Gülşah Şeydaoğlu³

¹Department of Nuclear Medicine, Çukurova University School of Medicine, Adana, Turkey

²Department of Nuclear Medicine, State Hospital, Niğde, Turkey

³Department of Biostatistics, Çukurova University School of Medicine, Adana, Turkey

Background: 18F-fluorodeoxyglucose (FDG) positron emission tomography computed tomography (PET/CT) provides information about metabolic and morphologic status of malignancies. Tumor size and standardized uptake value (SUV) measurements are crucial for cancer treatment monitoring.

Aims: The purpose of our study was to assess the variability of these measurements performed by observers evaluating lung tumors.

Study Design: Retrospective cross-sectional study.

Methods: FDG PET/CT images of 97 patients with pulmonary tumors were independently evaluated by two experienced nuclear medicine physicians. Primary tumor size (UDCT), maximum SUV (SUVmax), mean SUV (SUVmean) and maximum SUV normalized to liver mean SUV (SUVnliv max) were measured by each observer at two different times with an interval of at least 2 weeks. In-

terobserver and intraobserver variabilities of measurements were evaluated through statistical methods.

Results: Size of the lesions varied from 0.81 to 13.6 cm (mean 4.29±2.24 cm). Very good agreement was shown with correlation, Bland-Altman and regression analysis for all measured PET/CT parameters. In the interobserver and intraobserver variability analysis, the Pearson correlation coefficients were greater than 0.96 and 0.98, respectively.

Conclusion: Semi-quantitative measurements of pulmonary tumors were highly reproducible when determined by experienced physicians with clinically available software for routine FDG PET/CT evaluation. Consistency may be improved if the same observer performs serial measurements for any one patient.

Keywords: FDG PET/CT, pulmonary tumors, standardized uptake value, variability

The integrated positron emission tomography/computed tomography (PET/CT) allows the precise localization of the abnormal isotope uptake. PET/CT imaging using glucose analogue, 18F-fluorodeoxyglucose (FDG) provides valuable information about differential diagnosis, staging and treatment response of malignant tumors (1). Many malignant neoplasms and their metastases are characterized by enhanced glucose utilization and therefore increased FDG uptake. Besides a qualitative evaluation, different quantitative measurements of FDG uptake can be obtained from PET/CT scans. Standardized uptake value (SUV), which is a measurement of activity per unit volume of tissue (MBq/mL) adjusted for administered activity

per unit of body weight (MBq/g), is the preferred index used as a semi-quantitative measurement of glucose intake.

Various approaches for SUV determination have been used (2-13). Among them, maximum SUV within the slice with highest radioactivity concentration is commonly used (2-6). Quantitation requires delineation of the tumor tissue by regions of interest (ROIs). ROI definition is not fully automated in most of the approaches. Observers manually selects the region to be measured.

SUV has been found helpful for differentiation between benign and malignant pulmonary lesions. SUV more than 2.5 usually has a positive predictive value (PPV) of approximately

Address for Correspondence: Dr. Gülgün Büyükdereli, Department of Nuclear Medicine, Çukurova University School of Medicine, Adana, Turkey

Phone: +90 532 430 51 18 e-mail: gulgunbuyukdereli@hotmail.com

Received: 11 July 2014

Accepted: 26 June 2015 • DOI: 10.5152/balkanmedj.2016.140530

Available at www.balkanmedicaljournal.org

Cite this article as:

Büyükdereli G, Güler M, Şeydaoğlu G. Interobserver and Intraobserver Variability among Measurements of FDG PET/CT Parameters in Pulmonary Tumors. *Balkan Med J* 2016;33:308-15



80% when indicating malignancy and an SUV above 4.0 has a PPV of about 90% (14). SUV is also preferable to visual assessment when evaluating the effects of therapy in lung cancer (13,15). Response of lung cancers to treatment is determined by serial size and SUV measurements of tumor on PET/CT scans. The percentage of change in the measurements between a baseline scan and a second scan obtained during treatment or after the end of treatment is used to monitor response. Therefore, variability among measurements must be known. In the present study, our purpose was to assess the interobserver and intraobserver variability of size and SUV measurements of primary lung tumors with clinically available software used in the evaluation of routine FDG PET/CT.

MATERIALS AND METHODS

We retrospectively analyzed FDG PET/CT scans of 97 consecutive patients who had pulmonary tumors and were referred for diagnosis or initial staging between January 2011 and December 2012 obtained from the database of our institute. Only lesions which were visually identified in both FDG PET and CT images were included in the study. If multiple lesions were observed, the dominant pulmonary lesions were selected for measurements. We obtained informed consent from the patients. This study was approved by institutional ethics committee.

FDG PET/CT Imaging

FDG PET/CT was performed using an integrated PET/CT scanner which consisted of a full-ring high-resolution LSO PET and a six-slice CT (Siemens Biograph 6; Knoxville, USA). All patients fasted for at least 6 hours before undergoing PET/CT. Serum glucose levels were measured to ensure

that the results were <200 mg/dl. Whole-body images were acquired 60 minutes after intravenous injection of FDG, and images were obtained from the level of vertex to that of the proximal thigh region.

FDG PET/CT Image analysis

Two nuclear medicine physicians who have faculty experience in reading PET/CT (Reader 1 has 4 years of experience and had read at least 5000 scans; Reader 2 has 8 years of experience and had read at least 10,000 scans) evaluated all PET/CT images on an E-soft workstation, independently. Images were analyzed semi-quantitatively by use of the SUV as indices of FDG uptake. Each observer measured primary tumor size on the axial slice that showed the largest tumor dimension on the CT component of the PET/CT (UD_{CT}). Maximum (SUV_{max}) and mean (SUV_{mean}) SUVs were calculated by manually drawing ROIs over the primary tumor. ROI was placed around the most intense slice of the tumor, which was identified by defining ROIs over every axial image plane of the whole tumor. In order to measure the mean standard uptake value of the liver, a standard ROI with a diameter of 30 mm was drawn on the right lobe of the liver. Then, the ratio of primary tumor SUV_{max} to the liver SUV_{mean} was calculated to obtain SUV_{max} normalized to liver SUV_{mean} ($SUV_{nliv max}$) (Figure 1, 2). Each observer measured each PET/CT parameter at two different times. There was an interval of at least 2 weeks between the first and second image analysis. The observers were blinded to the measurements performed before.

Statistical analysis

Interobserver and intraobserver variabilities between 2 observers were determined with different statistical analyses evaluating different aspects of agreement. Hence, correlation, regression and Bland-Altman methods were used. The 8 pairs of measure-

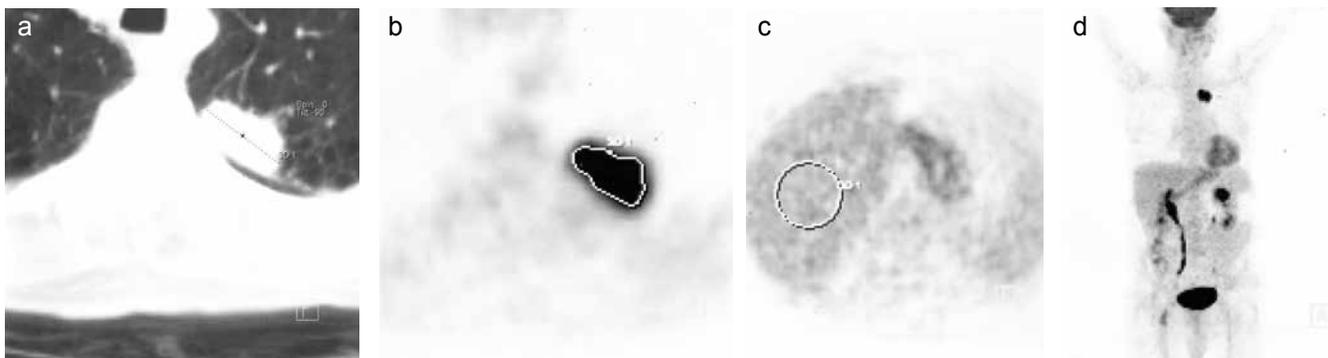


FIG. 1. a-d. FDG PET/CT of a 69 year-old man with lung cancer in the left upper lobe. Here, unidimensional size measurements on CT scan (a), SUV_{max} and SUV_{mean} measurements of tumour (yada lung lesion) on PET scan (b), SUV_{mean} measurement of normal liver paranchyma on PET scan (c), and whole-body PET scan (d) are shown. For observer 1, UD_{CT} , SUV_{mean} , SUV_{max} and $SUV_{nliv max}$ values were 3.58, 10.01, 15.91, 6.51 at time 1 and 3.71, 10.13, 16.13, 6.45 at time 2, respectively. For observer 2, these values were 3.80, 9.96, 16.08, 6.86 at time 1 and 3.73, 10.44, 16.11, 6.74 at time 2, respectively.

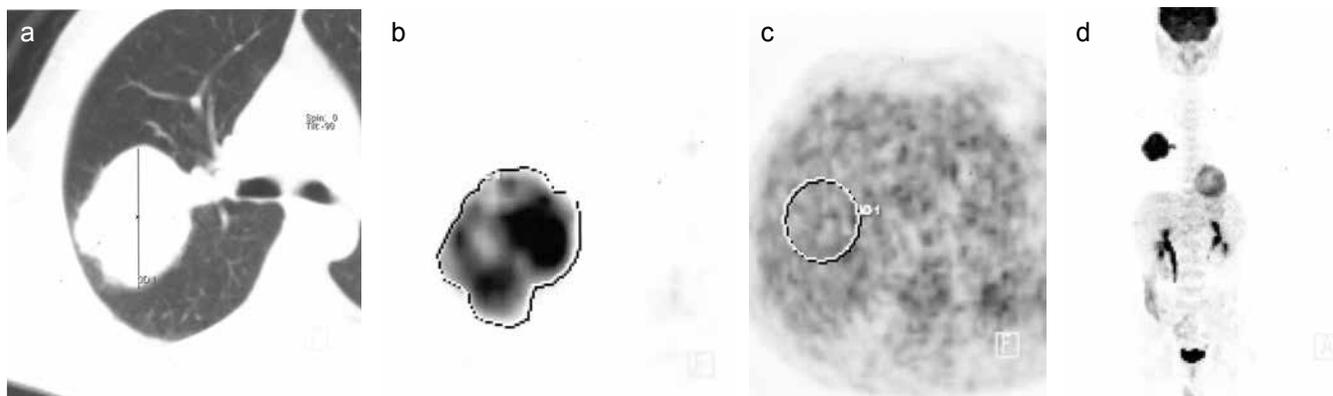


FIG. 2. a-d. FDG PET/CT of a 47 year-old man with lung cancer in the right upper lobe. Here, unidimensional size measurement on CT scan (a), SUV_{max} and SUV_{mean} measurements of tumour (yada lung lesion) on PET scan (b), SUV_{mean} measurement of normal liver paranchyma on PET scan (c), and whole-body PET scan (d) are shown. For observer 1, UD_{CT} , SUV_{mean} , SUV_{max} and $SUV_{nliv,max}$ values were 6.89, 5.10, 11.27, 8.80 at time 1 and 6.92, 5.02, 10.95, 9.04 at time 2, respectively. For observer 2, these values were 7.05, 5.03, 11.03, 8.96 at time 1 and 6.90, 4.93, 11.27, 8.87 at time 2, respectively.

ments by the 2 observers were evaluated in this way. Interobserver variability of the 4 parameters was defined with the use of an intraclass correlation coefficient (ICC). When interobserver agreement for the parameters measured is perfect the ICC will approach 1. Landis and Koch (16) classified the interpretation of ICC as follows: 0.0-0.20 slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.00 almost excellent agreement. The interobserver and intraobserver correlation of the measurements were calculated by the Pearson correlation analysis. The null hypothesis for analysis of correlation was that the correlation was 0; thus, excellent agreement would be 1. In the analysis of regression for intraobserver variability, the SUVs at time 1 were subjected to regression on the SUVs at time 2 and for interobserver agreement, the SUVs for observer 1 (the dependent variable) were subjected to regression on those of observer 2 (the independent variable). Good agreement was assumed when the regression line went through the origin with the slope of 1. Coefficient of variation (COV) was used to analyze the interobserver variability. COV was obtained for every parameter for each patient by dividing the SD by means which were calculated from two readings. In order to acquire the whole interobserver agreement of the 4 parameters, root-mean-square value of the 97 COVs was obtained. All of the measured parameters were expressed as mean±standard deviation (SD). To carry out statistical analysis, SPSS version 20 (SPSS Inc., Chicago, IL, USA) was used. Statistical significance was defined as $p < 0.05$.

RESULTS

The study group was composed of 97 patients (77 men, 20 women; average age, 58.2 ± 9.8 years; range 28-81 years) who had pulmonary tumors (59 malignant and 38 benign). The

mean±SD and COV of the SUV_{max} , SUV_{mean} , $SUV_{nliv,max}$ and the UD_{CT} measurements of the pulmonary tumors obtained by the 2 observers from the 97 patients' scans at 2 different times are presented in Table 1. The size of the primary tumor varied from 0.81 to 13.6 cm (mean 4.29 ± 2.24 cm). Root-mean-square value of these COVs was lowest for UD_{CT} (7.23%) and highest for SUV_{max} (9.11%).

Intraobserver variability

Correlation analysis was performed to assess intraobserver variability. Very high correlation was found between 2 reading times for all the parameters. With p values set at < 0.05 , the correlation was significantly different from 0 for all cases. The Pearson correlation coefficients varied from 0.985 to 0.998 for observer 1 and from 0.979 to 0.999 for observer 2 (Table 2).

Regression analysis also indicated a very good intraobserver reproducibility for every observer at 2 reading times, with an intercept of 0 and a slope of 1 for almost overall parameters. However, intercepts were 0 for all measurements (confidence interval [CI] of the intercepts included 0) only for SUV_{mean} and for UD_{CT} , slope wasn't 1 (95% CI of the slope did not include value 1). The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero. The p values were > 0.05 . The alternative hypothesis for the intercept was that the intercept is 0, and slope is 1. The fact that the null hypotheses could be rejected for both intercept and slope, indicated good intraobserver reproducibility for both observers.

Bland-Altman Analysis was also carried out in order to determine intraobserver agreement. Difference (the dependent variable) between the two reading times was subjected to regression on the sum (the independent variable) of the readings at times 1 and 2. No association was found between the dif-

TABLE 1. The SUV_{max}, SUV_{mean}, SUV_{niv max} and the UDCT measurements and their respective COV of patients with pulmonary tumors

	Time 1 Mean±SD		Time 2 Mean±SD		Average of 4 measurements Mean±SD	COV %
	Observer 1	Observer 2	Observer 1	Observer 2		
SUV _{max}	10.29±8.53	10.48±8.69	10.29±8.59	10.48±8.69	10.38±8.62	83.0 (9.11)
SUV _{mean}	4.36±3.58	4.42±3.66	4.43±3.68	4.51±3.68	4.43±3.60	81.5 (9.03)
SUV _{niv max}	5.25±4.39	5.40±4.45	5.25±4.34	5.41±4.48	5.32±4.41	82.7 (9.10)
UDCT	4.24±2.21	4.27±2.26	4.30±2.23	4.36±2.35	4.29±2.25	52.3 (7.23)

COV: coefficient of variation; SUV_{max}: maximum standardized uptake value; SUV_{mean}: mean standardized uptake value; SUV_{niv max}: SUV_{max} normalized to liver SUV_{mean}; UDCT: unidimensional tumor size measurement on CT scan

TABLE 2. Intraobserver agreement of the SUV_{max}, SUV_{mean}, SUV_{niv max} and UDCT measurements by regression and correlation analysis

	Reader	Intercept	95% CI	Slope	95% CI	r
SUV _{max}	Observer 1	0.06	-0.07-0.18	0.99	0.98-1.00	0.998
	Observer 2	0.00	-0.10-0.09	1.00	0.99-1.00	0.999
SUV _{mean}	Observer 1	0.10	-0.01-0.30	0.96	0.89-0.97	0.985
	Observer 2	0.04	-0.20-0.27	0.97	0.93-1.01	0.979
SUV _{niv max}	Observer 1	0.05	-0.19-0.10	1.00	0.98-1.02	0.994
	Observer 2	0.04	-0.06-0.14	0.99	0.97-1.07	0.997
UDCT	Observer 1	0.14	-0.02-0.29	0.94	0.91-0.98	0.991
	Observer 2	0.12	-0.06-0.15	0.97	0.95-1.00	0.992

Regression analysis conducted by subjecting measurements at time 1 to regression on measurements at time 2 for each observer. CI: Confidence Interval, r: Pearsons Correlation Coefficient

ference and sums for both observers and for all parameters, showing a very good intraobserver reproducibility.

Interobserver variability

Table 3 shows the results for the interobserver agreement at time 1 and at time 2. Concerning interobserver agreement, the correlation between observer 1 and observer 2 was very high. Utilizing the Fisher test and with p values set at <0.05, the correlation was significantly different from 0, in all cases. The Pearson correlation coefficients between observers ranged from 0.965 to 0.999 at time 1 and from 0.959 to 0.999 at time 2, respectively, for all parameters. The ICC between the two observers was found to be over 0.98 for all parameters at both times, indicating a very good agreement between observers (ICC approaches 1.0 if there is an excellent agreement between the observers for the measured parameters).

TABLE 3. Interobserver agreement of the SUV_{max}, SUV_{mean}, SUV_{niv max} and UDCT measurements by regression and correlation analysis

	SUV _{max}	SUV _{mean}	SUV _{niv max}	UD _{CT}
Time 1				
Intercept (95% CI)	0.01 (-0.11-0.13)	0.17 (-0.13-0.47)	-0.05 (-0.19-0.08)	0.14 (-0.04-0.33)
Slope (95% CI)	0.98 (0.97-1.00)	0.95 (0.89-1.00)	0.98 (0.96-1.00)	0.96 (0.92-1.00)
ICC (95% CI)	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.98 (0.98-0.99)
R	0.999	0.965	0.995	0.991
Time 2				
Intercept (95% CI)	-0.05 (-0.20-0.04)	0.11 (-0.23-0.43)	0.06 (-0.10-0.17)	0.23 (-0.04-0.41)
Slope (95% CI)	0.99 (0.98-1.00)	0.96 (0.90-1.02)	0.97 (0.95-0.98)	0.93 (0.89-0.97)
ICC (95% CI)	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.98-0.99)
R	0.999	0.959	0.997	0.983

ICC: intraclass correlation coefficient; CI: confidence interval; r: Pearson's correlation coefficient

Agreement between the observers was also evaluated by regression of the first observer and that of second observer for all measurements. Overall, there was a very good agreement between observers with the intercept of 0 and the slope of 1. The exception was between observer 1 and 2 at time 2, for SUV_{niv max} and for UD_{CT} where the slope was not 1 (95% CI of the slope did not include value 1). However, intercepts were 0 for all measurements (CI of the intercept included 0), which indicated that the null hypothesis could be rejected for both intercept and slope and also indicated good agreement between observers at both times.

Variability according to tumor size and malignancy status

We also analyzed the results according to tumor size and pathology. All pulmonary tumors were classified into two groups according to their size (<=30 mm, >30mm) and pathology (malign or benign lesions). The correlation coefficients were given in Table 4. Among the measured SUVs, interobserver and intraobserver agreements were highest for SUV_{max} and lowest for SUV_{mean} in all of the classified groups. According to our findings, agreement in SUV_{mean} was slightly lower in larger tumor sizes and malignant groups compared to the smaller and benign ones. However, intraobserver and interobserver correlation for all measured PET/CT parameters were still very high in all classified groups. Figures 3 and 4 show the Bland-Altman plots for SUV_{max} and SUV_{mean} according to tumor size.

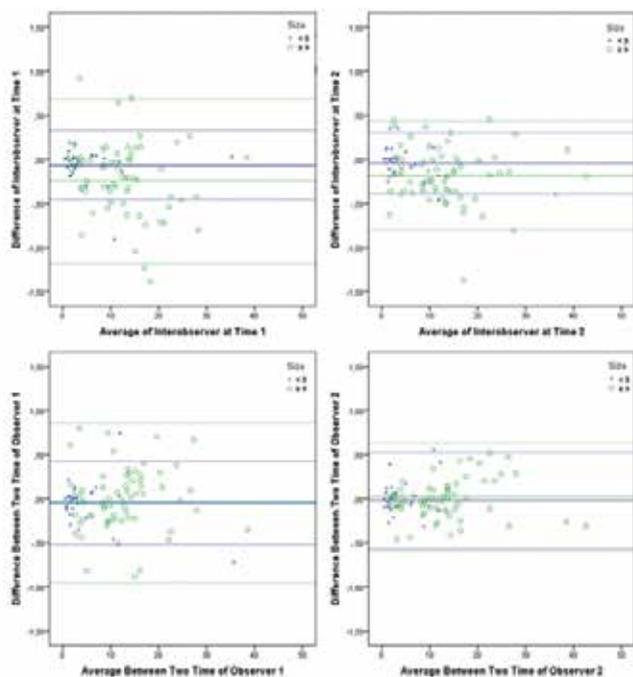


FIG. 3. Bland-Altman plots for SUV_{max} values in pulmonary lesions classified according to tumor size (<3 cm, $\geq 3\text{cm}$).

DISCUSSION

FDG PET/CT functional imaging is widely used in the assessment of pulmonary nodules and masses either to categorize a lesion as malignant or benign or to stage and monitor lung cancer (14). For the quantification of tumor glucose uptake, SUVs are commonly employed as a semi-quantitative index. For repeated tumors, measurements of SUV must be highly reproducible. This is particularly important for therapy assessment when small changes in tumor metabolism are being evaluated. Metabolic response on PET is manifested by decrease in the glycolytic activity of tumor (17,18). A lack of metabolic progression is associated with good improvement in outcome. Hence, the importance of reliable quantification of FDG uptake has been accentuated in many studies. Differences in measurements may lead to unnecessary operations or erroneous changes in therapy such as the discontinuation of prior curative therapies and the introduction of new chemotherapeutic drugs. Therefore, the interobserver and intraobserver variabilities of SUV analysis must be known. In the study by Marom et al. (3), to assess interobserver and intraobserver variabilities of SUV_{max} measurements, 5 observers determined the SUV_{max} in 20 patients with lung cancer. SUV_{max} was determined using 2 different methods: by manually shifting a fixed-size (1-cm) circular ROI around the lesion until SUV_{max} was detected and by creating a freehand drawing

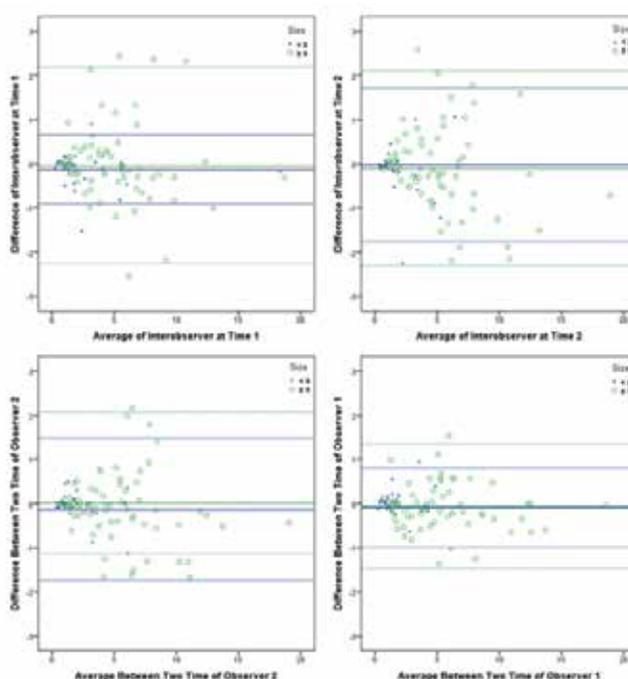


FIG. 4. Bland-Altman plots for SUV_{mean} values in pulmonary lesions classified according to tumor size (<3 cm, $\geq 3\text{cm}$).

placed over the lesion within the slice in which the lesion was visualized by the highest FDG concentration. The SUV_{max} measurements demonstrated good interobserver and intraobserver agreement with various statistical analyses (correlation, Bland-Altman, regression, and ANOVA). In the intraobserver and interobserver variability analysis, Pearson correlation coefficients were greater than 0.94 and 0.95, respectively. In that study, SUV measurements obtained from the study were also compared with the calculations reported in the initial clinical documents. In that case, agreement was poor. SUV_{max} measurements of the study for the same tumor showed a difference greater than 25% from the measurements of initial clinical documents in 45% of the tumors. In the study by Minn et al. (19), which evaluated 10 patients with lung cancer, and the study by Benz et al. (6), which evaluated 33 patients with high-grade sarcomas, there was 100% agreement in SUV_{max} measurements of the tumors determined by two observers.

The results of interobserver and intraobserver agreement in tumor size measurements in previous studies are discordant. In the literature, several studies have reported good interobserver and intraobserver agreement of tumor size measurements (20,21). However, some other studies have demonstrated substantial variation (22,23). Huang et al. (5) evaluated 43 pulmonary nodules on PET/CT and assessed interobserver variability of SUV_{max} , SUV_{mean} and tumor size measurements. They found no interobserver variation in SUV_{max} measure-

TABLE 4. Correlation Coefficients of the SUV_{max} , SUV_{mean} , $SUV_{nliv max}$ and UD_{CT} measurements according to tumor size and malignancy status

				Time 1		Time 2
				Observer 1	Observer 2	Observer 1
Tumor Size<3	Tumor Size	Time 1	Observer 1	1		
			Observer 2	0.912		
		Time 2	Observer 1	0.921	0.901	1
			Observer 2	0.911	0.946	0.937
	SUV_{max}	Time 1	Observer 1	1		
			Observer 2	1.00	1	
		Time 2	Observer 1	0.998	0.999	1
			Observer 2	1.00	0.999	1.0
	SUV_{mean}	Time 1	Observer 1	1		
			Observer 2	0.995	1	
		Time 2	Observer 1	0.993	0.986	1
			Observer 2	0.980	0.978	0.987
$SUV_{nliv max}$	Time 1	Observer 1	1			
		Observer 2	0.996	1		
	Time 2	Observer 1	0.997	0.998	1	
		Observer 2	0.996	0.998	0.999	
Tumor Size>3	Tumor Size	Time 1	Observer 1	1		
			Observer 2	0.981	1	
		Time 2	Observer 1	0.967	0.965	1
			Observer 2	0.965	0.978	0.988
	SUV_{max}	Time 1	Observer 1	1		
			Observer 2	0.999	1	
		Time 2	Observer 1	0.998	0.998	1
			Observer 2	0.998	0.999	0.999
	SUV_{mean}	Time 1	Observer 1	1		
			Observer 2	0.978	1	
		Time 2	Observer 1	0.946	0.945	1
			Observer 2	0.944	0.950	0.974
$SUV_{nliv max}$	Time 1	Observer 1	1			
		Observer 2	0.994	1		
	Time 2	Observer 1	0.992	0.995	1	
		Observer 2	0.993	0.996	0.993	
Benign	Tumor Size	Time 1	Observer 1	1		
			Observer 2	0.982	1	
		Time 2	Observer 1	0.970	0.980	1
			Observer 2	0.969	0.977	0.994
	SUV_{max}	Time 1	Observer 1	1		
			Observer 2	0.998	1	
		Time 2	Observer 1	0.998	0.998	1
			Observer 2	0.998	0.999	0.998
	SUV_{mean}	Time 1	Observer 1	1		
			Observer 2	0.97	1	
		Time 2	Observer 1	0.977	0.954	1
			Observer 2	0.949	0.925	0.974
$SUV_{nliv max}$	Time 1	Observer 1	1			
		Observer 2	0.997	1		
	Time 2	Observer 1	0.986	0.988	1	
		Observer 2	0.997	0.999	0.985	
Malignant	Tumor Size	Time 1	Observer 1	1		
			Observer 2	0.984	1	
		Time 2	Observer 1	0.976	0.970	1
			Observer 2	0.973	0.985	0.989
	SUV_{max}	Time 1	Observer 1	1		
			Observer 2	0.999	1	
		Time 2	Observer 1	0.999	0.998	1
			Observer 2	0.999	0.999	0.999
	UV_{mean}	Time 1	Observer 1	1		
			Observer 2	0.979	1	
		Time 2	Observer 1	0.948	0.948	1
			Observer 2	0.940	0.946	0.971
$SUV_{nliv max}$	Time 1	Observer 1	1			
		Observer 2	0.992	1		
	Time 2	Observer 1	0.991	0.995	1	
		Observer 2	0.992	0.995	0.994	

ments and low variability in SUVmean or size measurements. Jackson et al. (4) evaluated the interobserver agreement of the SUVmax, SUVnliv max and size measurements in primary tumors between observers with disparate practice in PET/CT. They found that in the setting of different reading experiences, FDG metabolic parameters have higher interobserver agreement than size measurements. They reported that size measurements were more manual and subjective than the SUV measurements as marginally altered angles in the measurements could produce greater variability.

Since there is no gold standard method in measuring SUV values, we used 3 commonly applied approaches (SUVmax, SUVmean and SUVnliv max) for SUV calculation in our study. Among the measured SUVs, intraobserver and interobserver agreements were highest for SUVmax and lowest for SUVmean in all subgroups, which were classified according to tumor size and malignancy status. Our findings showed that variability of SUVmean was slightly higher in larger tumor sizes and malignant groups compared to the smaller and benign ones. In general, the low activity in the background of the thorax might have facilitated tumor delineation in FDG PET/CT scans. However, in some of the necrotic regions of large tumors, tumor delineation was more difficult. In such cases, necrotic regions affected the SUVmean much more than SUVmax. The strong dependency of the SUVmean on the dimension and shape of the ROIs may explain the slightly higher variability of SUVmean in our study. However, intraobserver and interobserver correlation for all measured PET/CT parameters were still very high in all classified groups. This very good interobserver and intraobserver agreement for all parameters in our study may be related to the training and experience of the observers, as mentioned in previous studies (4,24). The other factor resulting in the lower variability in our study was probably the study environment. More attention of the readers to technique might affect the results. The present study had limitations. It was designed as a single center with a relatively small patient population (n=97). There were only 2 readers and the exams were limited to the lungs. The low level of background activity in the thorax might have facilitated tumor delineation in FDG PET/CT studies. The higher and more variable level of background activity in other parts of the body may lead to a higher variability of FDG PET/CT measurements. A larger, multi-institutional prospective studies are needed to expand our present findings.

This study shows that if PET/CT is evaluated by experienced observers with careful attention to technique, SUV and size measurements on FDG PET/CT are highly reproducible. However, despite the high interobserver and intraobserver reproducibility, the fact that some variability exists must be considered when evaluating tumor response. Consistency may be improved if the same observer performs serial measure-

ments for any one patient. If the difference in SUV and size measurements account for a modulation in treatment and there is discrepancy between the clinical and imaging findings, it would be helpful to repeat current and prior FDG uptake measurements by the same observer to improve reproducibility.

Ethics Committee Approval: Ethics committee approval was received for this study from the Ethics Committee of Çukurova University.

Informed Consent: Informed consent was obtained from the patients who participated in this study.

Peer-review: Externally peer-reviewed.

Author contributions: Concept - G.B.; Design - G.B.; Supervision - G.B., G.Ş.; Resource - G.B., M.G.; Materials - G.B., M.G.; Data Collection and/or Processing - G.B., M.G.; Analysis and/or Interpretation - G.B., M.G., G.Ş.; Literature Search - G.B.; Writing - G.B.; Critical Reviews - G.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. Dimitrakopoulou-Strauss A, Strauss L. Quantitative studies using positron emission tomography (PET) for the diagnosis and therapy planning of oncological patients. *Hell J Nucl Med* 2006;9:10-21.
2. Rousseau C, Devillers A, Sagan C, Ferrer L, Bridji B, Campion L, et al. Monitoring of early response to neoadjuvant chemotherapy in stage II and III breast cancer by [¹⁸F]fluorodeoxyglucose positron emission tomography. *J Clin Oncol* 2006;24:5366-72. [\[CrossRef\]](#)
3. Marom EM, Munden RF, Truong MT, Gladish GW, Podoloff DA, Mawlawi O, et al. Interobserver and intraobserver variability of standardized uptake value measurements in non-small-cell lung cancer. *J Thorac Imaging* 2006;21:205-12. [\[CrossRef\]](#)
4. Jackson T, Chung MK, Mercier G, Ozonoff A, Subramaniam RM. FDG PET/CT interobserver agreement in head and neck cancer: FDG and CT measurements of the primary tumor site. *Nucl Med Commun* 2012;33:305-12. [\[CrossRef\]](#)
5. Huang YE, Chen CF, Huang YJ, Konda SD, Appelbaum DE, Pu Y. Interobserver variability among measurements of the maximum and mean standardized uptake values on (18) F-FDG PET/CT and measurements of tumor size on diagnostic CT in patients with pulmonary tumors. *Acta Radiol* 2010;51:782-8. [\[CrossRef\]](#)
6. Benz MR, Evilevitch V, Allen-Auerbach MS, Eilber FC, Phelps ME, Czernin J, et al. Treatment monitoring by 18F-FDG PET/CT in patients with sarcomas: interobserver variability of quantitative parameters in treatment-induced changes in histopathologically responding and nonresponding tumors. *J Nucl Med* 2008; 49:1038-46. [\[CrossRef\]](#)

7. Ott K, Fink U, Becker K, Eilber FC, Phelps ME, Czernin J, et al. Prediction of response to preoperative chemotherapy in gastric carcinoma by metabolic imaging: results of a prospective trial. *J Clin Oncol* 2003;2:4604-10. [\[CrossRef\]](#)
8. Avril N, Sassen S, Schmalfeldt B, Naehrig J, Rutke S, Weber WA, et al. Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer. *J Clin Oncol* 2005;23:7445-53. [\[CrossRef\]](#)
9. Eschmann SM, Friedel G, Paulsen F, Reimold M, Hehr T, Budach W, et al. 18F-FDG PET for assessment of therapy response and preoperative re-evaluation after neoadjuvant radio-chemotherapy in stage III non-small cell lung cancer. *Eur J Nucl Med Mol Imaging* 2007;34:463-71. [\[CrossRef\]](#)
10. Hoekstra CJ, Stroobants SG, Smit EF, Vansteenkiste J, van Tinteren H, Postmus PE, et al. Prognostic relevance of response evaluation using [18F]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. *J Clin Oncol* 2005;23:8362-70. [\[CrossRef\]](#)
11. Weber WA, Ziegler SI, Thodtmann R, Thödtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG PET. *J Nucl Med* 1999;40:1771-7.
12. Scarfone C, Lavelly WC, Cmelak AJ, Delbeke D, Martin WH, Billheimer D, et al. Prospective feasibility trial of radiotherapy target definition for head and neck cancer using 3-dimensional PET and CT imaging. *J Nucl Med* 2004;45:543-52.
13. Weber WA, Wieder H. Monitoring chemotherapy and radiotherapy of solid tumors. *Eur J Nucl Med Mol Imaging* 2006;33:27-37. [\[CrossRef\]](#)
14. Mettler FA, Guiberteau MJ. F-18-FDG PET/CT neoplasm imaging. In: Mettler FA and Guiberteau MJ, editors. *Essentials of Nuclear Medicine Imaging*. Philadelphia: Elsevier Saunders 2012:361-96. [\[CrossRef\]](#)
15. Lin C, Itti E, Haioun C, Petegnief Y, Luciani A, Dupuis J, et al. Early 18F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis. *J Nucl Med* 2007;48:1626-32. [\[CrossRef\]](#)
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159-74. [\[CrossRef\]](#)
17. 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;50:122S-50S. [\[CrossRef\]](#)
18. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma AA, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer* 1999;35:1773-82. [\[CrossRef\]](#)
19. Minn H, Zasadny KR, Quint LE, Wahl RL. Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. *Radiology* 1995;196:167-73. [\[CrossRef\]](#)
20. Monsky WL, Raptopoulos V, Keogan MT, Doty D, Kamel I, Yam CS, et al. Reproducibility of linear tumor measurements using PACS: comparison of caliper method with edge-tracing method. *Eur Radiol* 2004;14:519-25. [\[CrossRef\]](#)
21. Wormanns D, Diederich S, Lentschig MG, Winter F, Heindel W. Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. *Eur Radiol* 2000;10:710-3. [\[CrossRef\]](#)
22. Revel MP, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology* 2004; 231:453-8. [\[CrossRef\]](#)
23. Mukherji SK, Toledano AY, Beldon C, Schmalfuss IM, Cooper JS, Sicks JD, et al. Interobserver reliability of computed tomography-derived primary tumor volume measurement in patients with supraglottic carcinoma. *Cancer* 2005;103:2616-22. [\[CrossRef\]](#)
24. Krabbe CA, Pruijm J, Scholtens AM, Roodenburg JL, Brouwers AH, Phan TT, et al. 18-F-FDG PET in squamous cell carcinoma of the oral cavity and oropharynx: a study on inter- and intraobserver agreement. *J Oral Maxillofac Surg* 2010;68:21-7. [\[CrossRef\]](#)