# Meta-Analysis of the Test-Retest Repeatability of [18F]-Fluorodeoxyglucose Standardized Uptake Values

Shankar, Lalitha K; Huang, Erich; Litière, Saskia; Hoekstra, Otto S; Schwartz, Larry; Collette, Sandra; Boellaard, Ronald; Bogaerts, Jan; Seymour, Lesley; de Vries, Elisabeth G E

*Published in:*
Clinical Cancer Research

*DOI:*
10.1158/1078-0432.CCR-21-3143

*Publication date:*
2023

*Citation for published version (APA):*
Shankar, L. K., Huang, E., Litière, S., Hoekstra, O. S., Schwartz, L., Collette, S., Boellaard, R., Bogaerts, J., Seymour, L., & de Vries, E. G. E. (2023). Meta-Analysis of the Test-Retest Repeatability of [18F]-Fluorodeoxyglucose Standardized Uptake Values: Implications for Assessment of Tumor Response. *Clinical Cancer Research*, *29*(1). https://doi.org/10.1158/1078-0432.CCR-21-3143

# Meta-Analysis of the Test–Retest Repeatability of [$^{18}$F]-Fluorodeoxyglucose Standardized Uptake Values: Implications for Assessment of Tumor Response

Lalitha K. Shankar[1], Erich Huang[1], Saskia Litiere[2], Otto S. Hoekstra[3], Larry Schwartz[4], Sandra Collette[5], Ronald Boellaard[3], Jan Bogaerts[2], Lesley Seymour[6], and Elisabeth G.E. deVries[7]

## ABSTRACT

**Purpose:** Currently, guidelines for PET with $^{18}$F-fluorodeoxyglucose (FDG-PET) interpretation for assessment of therapy response in oncology primarily involve visual evaluation of FDG-PET/CT scans. However, quantitative measurements of the metabolic activity in tumors may be even more useful in evaluating response to treatment. Guidelines based on such measurements, including the European Organization for Research and Treatment of Cancer Criteria and PET Response Criteria in Solid Tumors, have been proposed. However, more rigorous analysis of response criteria based on FDG-PET measurements is needed to adopt regular use in practice.

**Experimental Design:** Well-defined boundaries of repeatability and reproducibility of quantitative measurements to discriminate noise from true signal changes are a needed initial step. An extension of the meta-analysis from de Langen and colleagues (2012) of the test–retest repeatability of quantitative FDG-PET measurements, including mean, maximum, and peak standardized uptake values ($SUV_{max}$, $SUV_{mean}$, and $SUV_{peak}$, respectively), was performed. Data from 11 studies in the literature were used to estimate the relationship between the variance in test–retest measurements with uptake level and various study-level, patient-level, and lesion-level characteristics.

**Results:** Test–retest repeatability of percentage fluctuations for all three types of SUV measurement (max, mean, and peak) improved with higher FDG uptake levels. Repeatability in all three SUV measurements varied for different lesion locations. Worse repeatability in $SUV_{mean}$ was also associated with higher tumor volumes.

**Conclusions:** On the basis of these results, recommendations regarding SUV measurements for assessing minimal detectable changes based on repeatability and reproducibility are proposed. These should be applied to differentiate between response categories for a future set of FDG-PET–based criteria that assess clinically significant changes in tumor response.

## Introduction

PET with $^{18}$F-fluorodeoxyglucose (FDG-PET) plays a role in the staging and restaging of cancers and in the assessment of therapy response in lymphoma (1). The International Conference on Malignant Lymphomas criteria, based on FDG avidity of a lymphoma mass, were developed by panels of experts following international meetings and have been proven in subsequent studies involving large cohorts of patients to provide clinically meaningful information (2, 3). FDG-PET can support a diagnosis of progressive disease during systemic treatment by detecting new lesions (1). Other criteria such as International Working Group response criteria (4), Cheson criteria (5), and more recently, Lugano classification (6) and RECIL (7) have also been used in assessing response to therapy in lymphoma.

FDG-PET was incorporated into version 1.1 of the RECIST, although only as an adjunct to anatomical imaging for a more accurate determination of tumor progression (8). An expanded role for FDG-PET in RECIST-based response assessment is pending, whereas FDG-PET data obtained according to generally accepted standard procedures emerge from clinical trials. However, FDG-PET has not been widely accepted in assessing response to treatment in solid tumors in clinical trials, although qualitative assessment of FDG uptake changes guides clinical care. For clinicians and researchers to widely accept the use of FDG-PET in tumor response assessment, (i) results indicating adequate performance of FDG-PET in this role from clinical trial data obtained according to generally accepted standard procedures are necessary and (ii) should provide a better performance than measurement of tumor size on CT or MR.

Almost 20 years ago, the European Organization for Research and Treatment of Cancer (EORTC) PET Study Group published recommendations for response monitoring using quantitative PET data in an attempt to harmonize PET response reporting in clinical trials. This was based on the limited evidence available at that time (9). Between 2000 and 2010, it became apparent that guidelines were required to standardize the entire workflow from patient preparation and image acquisition, reconstruction, and analysis to calibration to ensure exchangeability of scan results between institutes, scanner vendors, and software platforms. Guidelines introduced during this period include the 2006 consensus recommendations from NCI (10), European Association of Nuclear Medicine (EANM) guidelines versions 1.0 and 2.0 (11, 12), the Netherlands Protocol for Standardization (13), and Perlman Uniform Protocols for Imaging in Clinical Trials (14). Other sets of response criteria were developed more recently in light of these guidelines, the most notable of which is the PET Response Criteria in Solid Tumors (PERCIST) introduced in 2009. Included in PERCIST are specific details of lesion selection and imaging analysis and proposing new methods to determine the PET "volume of interest" (VOI) as well as new definitions of levels of metabolic response (15).

[1]National Cancer Institute, Bethesda, Maryland. [2]European Organisation for Research and Treatment of Cancer Headquarters, Brussels, Belgium. [3]Vrije Universiteit Medical Center, Amsterdam, the Netherlands. [4]Columbia University Medical Center and New York Presbyterian Hospital, New York, New York. [5]Bristol Myers Squibb, Princeton, New Jersey. [6]Queen's University, Kingston, Ontario, Canada. [7]University of Groningen, Groningen, the Netherlands.

**Corresponding Author:** Lalitha K. Shankar, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892. E-mail: shankarl@mail.nih.gov

Clin Cancer Res 2023;29:143–53

**doi:** 10.1158/1078-0432.CCR-21-3143

## Translational Relevance

The role of FDG-PET in the assessment of treatment response in solid tumors has been hindered by a lack of prospective clinical trials sufficient in size to determine its utility. In addition, the performance characteristics of semiquantitative assessments of change in FDG uptake have been limited to small studies. This meta-analysis of test–retest data from the available published literature, including multi-center studies, provide a guide to develop response assessment criteria with FDG PET that can be assessed prospectively in clinical trials. Such studies can also assess whether the addition of FDG-PET response criteria can improve RECIST, thereby enhancing the assessment and selection of successful treatment strategies. The potential utility of increasing the performance of radiologic solid tumor response assessment with the added metabolic information from FDG-PET has been studied for decades. The potential benefits of this combined approach include improved efficiencies in the development and assessment of novel cancer therapies. In this analysis, the FDG-PET Working Group of the RECIST committee assessed the parameters of robustness (repeatability) of FDG-PET CT across available published studies. On the basis of this meta-analysis, we are able to define baseline characteristics of the tumor as well as the level of changes in SUV that correspond to significant changes in tumor metabolism. Using these parameters, clinical trials assessing tumor response with FDG-PET can assess the value of adding this metabolic information to morphologic assessments by CT or MR.

Even though prevailing guidelines for FDG-PET interpretation primarily involve visual evaluation of FDG-PET/CT scans, quantitative measurements of the metabolic activity in tumors can also be extracted and could potentially be even more useful in evaluating response to treatment. Evaluations of repeatability and reproducibility of quantitative measurements are essential in determining whether changes in the measurement values are likely due to actual changes in the underlying physiology rather than simply noise. For example, a 30% increase in standardized uptake value (SUV) is likely due to changes in metabolic activity if test–retest measurements deviate from each other by less than 10%. On the other hand, if such measurements regularly deviate by 50% or more, an increase of 30% may simply be attributable to natural measurement error. An earlier meta-analysis of test–retest repeatability studies of quantitative FDG-PET measurements, including maximum SUV ($SUV_{max}$) and mean SUV ($SUV_{mean}$), was performed using data from five studies consisting of 102 patients in total (16). On the basis of these data, the percentage of differences in repeat measurements of $SUV_{max}$ on the same patient using the same image acquisition and processing and SUV computation protocols in the absence of any intervention should be contained within the range (−28%–39%) with 95% probability (17). Hence, changes outside this range are indicative of likely underlying metabolic phenomena.

However, data on the current state of the art PET VOI methodologies were unavailable. Also, numerous issues remained unresolved, including how test–retest repeatability in these FDG-PET measurements varied as a function of characteristics such as tumor type and location and image acquisition and processing methodology and which FDG-PET measurements to use for evaluating response to treatment. The FDG-PET subcommittee of the RECIST group attempted to collect data from Phase II and III treatment trials involving FDG-PET in the assessment of solid tumors. Unfortunately, there is a paucity of data in this space. The subcommittee then collected data from more recent FDG-PET test–retest repeatability studies to produce a larger dataset with a broader range of lesion, patient, and image acquisition characteristics. Meta-analysis techniques as used earlier (16) were applied to this expanded dataset. The thresholds above which increases or decreases in FDG-PET measurements are likely attributable to actual changes in tumor metabolism instead of variabilities in physiological and image acquisition parameters were verified. Moreover, this meta-analysis will address issues not considered in previous analyses, such as how test–retest repeatability of various FDG-PET measurements differ as a function of image acquisition protocols and patient and lesion characteristics such as primary tumor type or lesion location. This analysis thus serves as an important intermediary step by addressing the imaging technique's reproducibility in establishing minimal detectable metabolic changes due to tumor response. These results should be applied to response-assessment criteria that are developed to evaluate meaningful clinical outcomes in the treatment of solid tumors.

## Materials and Methods

### Study search and inclusion/exclusion criteria

In addition to the studies included de Langen and colleagues (16), more recent studies on test–retest repeatability of FDG-PET measurements completed after the publication of that article were identified through PubMed searches. The additional included studies that satisfied all the following criteria: (i) Patients had one or more lesions of a solid tumor, (ii) patients underwent test and retest scans using the same scanner and same acquisition protocol within the space of 28 days; (iii) patients did not receive any treatment between scans; and (iv) quantitative FDG-PET parameters (e.g., SUV measurements), as opposed to qualitative assessment of FDG-PET scans were the focus.

### Data elements

The studies included in the meta-analysis contained test and retest measurements of one or more of the following FDG-PET parameters for each lesion for each patient: mean SUV ($SUV_{mean}$; the average uptake or SUV within a 3D VOI, typically defined by a 41% of SUVmax isocontour or a fixed SUV = 4 contour), maximum SUV ($SUV_{max}$; the maximum uptake or SUV of a single voxel seen within a 3D VOI) and peak SUV ($SUV_{peak}$; the average uptake or SUV based on a predefined fixed-sized VOI, usually a 3D spherical VOI of 1 mL volume and positioned in the tumor such to yield the highest $SUV_{peak}$ value; refs. 12, 15). Primary tumor type for each patient, site, and volume of each lesion were included in the data; summary statistics for these variables are provided in **Tables 1** and **3**.

The following image acquisition and processing information were available for each patient in each study: Time between test and retest scans, scanner used (e.g., PET alone or PET/CT and whether the scan was dynamic or static), thresholding technique to determine the VOI, technical scan details and reconstruction methods for SUV measurements, image-processing methodology for metabolic volume measurements, and total scan time. Summary statistics of these aspects are provided in **Table 2**.

### Data preprocessing

For quality control purposes, patients falling into any of the following categories were excluded from the analysis: (i) Those for which the time between injection and PET scan during the test PET scan differed from that during the retest scan by more than 15 minutes

**Table 1.** Summary characteristics of the studies included in this meta-analysis.

| Study (*denotes multicenter) | Included in meta-analysis from de Langen et al. | Scanner type and vendor used in number of patients | Number of patients | Tumor types and number of patients | Number of lesions | Acquisition protocol |
|---|---|---|---|---|---|---|
| Hoekstra et al (2002) | Yes | PET (Siemens): 10 (100%) | 10 | Lung: 10 (100%) | 27 | Dynamic |
| Weber et al (1999) | Yes | PET (Siemens): 16 (100%) | 16 | Lung: 8(50.0%) Colorectal: 3 (18.8%)<br>Esophageal: 1 (6.3%)<br>Lymphoma: 1 (6.3%)<br>Renal: 1 (6.3%)<br>Vulvar: 1 (6.3%)<br>Adenoid cystic: 1 (6.3%) | 50 | Dynamic |
| Nahmias and Wahl (2008) | Yes | PET/CT (Siemens): 21 (100%) | 21 | Lung: 8 (38.1%)<br>Esophageal: 3 (14.3%)<br>Breast: 4 (19.0%)<br>Head and neck: 1 (4.8%)<br>Lymphoma: 1 (4.8%)<br>Prostate: 1 (4.8%)<br>Melanoma: 1 (4.8%)<br>Thyroid: 1 (4.8%)<br>Renal: 1 (4.8%) | 21 | Static |
| Velasquez et al (2009) | Yes | PET (unknown vendor): 7<br>PET/CT (unknown vendor): 35 | 42 | Colorectal: 42 (100%) | 105 | Static (35)<br>Dynamic (7) |
| Minn et al (1995) | Yes | PET (unknown vendor): 10 | 10 | Lung: 10 (100%) | 10 | Dynamic |
| Hatt et al (2010) | No | PET/CT (Philips): 15 | 15 | Esophageal: 15 (100%) | 17 | Static |
| Heijmen et al (2012) | No | PET/CT (Siemens): 15 | 15 | Colorectal: 15 (100%) | 24 | Static |
| Kramer et al (2016) | No | PET/CT (Philips) 10 | 10 | Lung: 10 (100%) | 75 | Static |
| ACRIN 6678* | No | PET/CT (Philips): 8<br>PET/CT (GE): 5<br>PET/CT (Siemens): 19 | 32 | Lung: 32 (100%) | 113 | Static |
| MERCK MK-046–008* | No | PET/CT (unknown vendor): 39 | 39 | Lung: 39 (100%) | 116 | Unknown |
| Rockall et al (2014) * | No | PET/CT (Philips): 8<br>PET/CT (GE): 13 | 21 | Ovarian: 21 (100%) | 87 | Static |

(based on recommendations from EANM and NCI guidelines; ref. 10), and (ii) those for which the time between the test and retest scans was more than 10 days based on EANM guidelines. In addition, individual lesions falling into any of the following categories were also excluded for quality control purposes: (i) Those for which the mean of the test and retest measurements of $SUV_{max}$ was less than 4.0 (i.e., relatively low uptake of FDG), and (ii) those for which the mean of the test and retest measurements of $SUV_{max}$ fell below the first percentile and above the 99th percentile across all lesions, similar to de Langen and colleagues (16).

## Statistical analysis

For each FDG-PET parameter, the variance–mean relationship (i.e., the relationship between the variance of the test and retest measurements, namely the test–retest variance, and the natural logarithm of the mean of the test and retest measurements, namely the test–retest mean) was modeled through a generalized linear mixed-effects model (GLMM; ref. 18) with a logarithmic link function as done by de Langen and colleagues (16). The study cohort was considered a random effect in this model, whereas the test–retest mean's natural logarithm was considered a fixed effect. Lesions were used as the unit of analysis, and the analysis was performed for each FDG-PET parameter separately. Exact likelihoods were used to bypass potential problems associated with the small sizes of some of the studies. Details of the GLMM are provided in the Appendix; the SAS procedure GLIMMIX was used to compute estimates of the model parameters and the *P* values of tests of significance of the association between test–retest variance and test–retest mean.

Probabilities of changes in SUV measurements exceeding a specific threshold due solely to random fluctuations can be derived from the variance–mean relationship (see Appendix). However, these probabilities may vary as a function of the magnitude of the uptake, image acquisition, and processing protocols, and patient and lesion characteristics, as all these factors can affect the SUV measurement. Thus, a multivariate GLMM was used to model the variance of the test and retest measurements as a function of one or more baseline imaging and processing protocol aspects, patient, and lesion characteristics, plus the logarithm of the mean of the test and retest measurements. As was done in the analysis of the variance–mean relationship described above, a logarithmic link function was used for this analysis, study cohorts were treated as random effects, and individual lesions were used as the unit of analysis. Forward stepwise selection served to identify which baseline characteristics, on top of the test–retest mean, to include in the model. Details of this GLMM are provided in the Appendix; the SAS procedure GLIMMIX was used to compute estimates of the model parameters and the *P* values of tests of significance of the association between the test–retest variance and the test–retest mean or baseline characteristic.

## Data availability

The data used in this study were shared by the investigators conducting the original studies. Data requests should be addressed to the original study sponsors.

**Table 2.** Additional summary statistics of image acquisition and processing parameters used in each of the studies.

| Study | Number of patients | Thresholding technique | SUV normalization technique | Injected dose (median and range) | Injection time per scan and number of patients | Median and range of days between test and retest scans |
|---|---|---|---|---|---|---|
| Hoekstra et al. (2002) | 10 | 50% of maximum voxel | Unknown | 370 (370–370) | 60 min: 10 (100%) | 1 (1–1) |
| Weber et al. (1999) | 16 | 50% of maximum voxel | Body weight | 370 (370–370) | 70 min: 16 (100%) | 3 (1–10) |
| Nahmias and Wahl (2008) | 21 | 50% of maximum voxel and manual delineation | Unknown | | 90 min: 21 (100%) | 2 (1–5) |
| Velasquez et al. (2009) | 42 | 70% of maximum and 50% of maximum voxel | Bodyweight and lean body mass | | 60 min: 42 (100%) | 4 (1–7) |
| Minn et al (1995) | 10 | 4×4 voxels around | Lean body mass | | 60 min: 10 (100%) | 2 (1–7) |
| Hatt et al (2010) | 15 | 50% of maximum voxel | Bodyweight | 6.0 (6.0–6.0) | 60 min: 15 (100%) | 4.1 (2–7) |
| Heijmen et al. (2012) | 15 | 41% of maximum voxel | Bodyweight | 264 (175–405) | 56 to 65 min: 10 (66.7%)<br>66 to 75 min: 5 (33.3%) | 5.5 (2–7) |
| Kramer et al. (2016) | 10 | 50% of maximum voxel | Lean body mass | | 60 min: 10 (100%); analyzed at both 60 min and 90 min. | 1 (1–2) |
| ACRIN 6678 | 32 | Cylindrical 1.5 cm$^3$ VOI around maximum voxel | Bodyweight | 368 (215–580) | Less than 60 min: 6 (18.8%)<br>60 min: 18 (56.3%)<br>70 min: 5 (15.6%)<br>80 min: 2/(6.3%) | 5 (1–7) |
| MERCK MK-046–008 | 39 | Cylindrical 1.5 cm$^3$ VOI around maximum voxel | Bodyweight | | Less than 60 min: 2 (5.1%)<br>60 min: 35 (89.7%)<br>70 min: 2 (5.1%) | 2 (1–6) |
| Rockall et al. (2014) | 21 | Manual delineation | Bodyweight | | 60 min: 21 (100%) | 3 (2–8) |

Note: Within each study, the investigator applied the same thresholding and SUV normalization technique to each patient. Exact timing was given for all studies but Heijmen et al, ACRIN 6678, MERCK MK-046–008, and Rockall et al.

## Results

### Summary statistics

Applying the search and inclusion/exclusion criteria yielded 11 studies ranging in size between 10 and 45 patients and 10 and 116 lesions (19–29), four of which (19–22) were multicenter studies. Three of these studies (23–25) used PET only, one (19) used both PET and PET/CT, and the others used PET/CT. Studies that used PET only used a dynamic scan acquisition, whereas those that used PET/CT used a static scan acquisition; one used both static and dynamic acquisitions (19). The 11 studies contained a total of 242 patients and 645 lesions; complete summary statistics of the studies are provided in **Table 1**.

$SUV_{max}$ measurements were available for 226 patients (595 lesions) across 10 studies (all but Weber and colleagues; ref. 24), whereas $SUV_{mean}$ measurements were available for 168 patients (416 lesions) across nine of these studies (all but ACRIN 6678 and MERCK MK-0646–008; refs. 20, 21) and $SUV_{peak}$ measurements were available for 102 patients (328 lesions) across four studies (20, 21, 26, 27). Application of the quality control criteria described previously in §2.3 resulted in the removal of 12 lesions from the ACRIN 6678 and MERCK MK-0646–08 trials (20, 21) and removal of four lesions in Heijmen and colleagues (26) in which the time difference between injection and scan during the test scan compared with during the retest scan exceeded 15 minutes. 53 lesions across eight of the studies (19–23, 26–28) were removed because of $SUV_{max}$ test–retest means not exceeding 4.0. Twelve lesions for which the $SUV_{max}$ test–retest means were less than the first percentile or greater than the 99th percentile were also removed. Overall, 88% of the lesions were evaluable except for the Merck study (77%).

After the data were screened using these criteria, $SUV_{max}$ measurements were available for 522 lesions across 222 patients. The test–retest means ranged from 4.0 to 27.1 (median 8.6), differences in test and retest measurements ranged from −7.4 to 7.7 (median 0.0), and magnitudes of the percentage of fluctuation (measurement divided by the test–retest mean) ranged from 0.0% to 90.8% (median 8.3%). $SUV_{mean}$ measurements were available for 384 lesions across 167 patients. Here the test–retest means ranged from 1.1 to 17.4 (median 5.6), differences in test and retest measurements ranged from −2.6 to 4.6 (median 0.0), and magnitudes of percent fluctuations ranged from 0.0% to 53.2% (median 7.0%). $SUV_{peak}$ measurements were available for 99 patients (274 lesions); test–retest means ranged from 2.3 to 20.8 (median 6.1), differences in test and retest measurements ranged from −4.9 to 5.4 (median 0.0), and magnitudes of percent fluctuations ranged from 0.0% to 85.0% (median 8.8%). Summary statistics of the SUV measurements are provided in **Tables 4** and **5**.

After the data were screened using the quality criteria, 117 patients (50.6%) had primary lung tumors, 60 (26.0%) had primary colorectal tumors, 21 (9.1%) had primary ovarian tumors, and 19 (8.2%) had primary esophageal tumors. Other tumor types included breast, head and neck, prostate, and renal. One hundred eight lesions (18.8%) were

**Table 3.** Summary statistics for lesion-level characteristics (localization and lesion size).

| Study | Number of evaluable lesions | Median and range of number of lesions per patient | Localizations of the number of lesions | Mean and range of lesion volumes (cm$^3$) | Number of lesions with volume $\geq$4.2 cm$^3$ |
|---|---|---|---|---|---|
| Hoekstra et al. (2002) | 26 | 2 (1–7) | Lung/mediastinum: 26 (100%) | 7.1 (0.8–111) | 15(57.7%) |
| Weber et al. (1999) | 50 | 2.5 (1–8) | Lung/mediastinum: 34/ (68.0%)<br>Lymph node: 8/ (16.0%)<br>Liver: 8 (16.0%) | 5.2 (0.6–86.9) | 27(54.0%) |
| Nahmias and Wahl (2008) | 21 | 1 (1–1) | Lung/mediastinum: 14 (66.7%)<br>Liver: 1 (4.8%)<br>Bone: 2 (9.5%)<br>Esophagus: 2 (9.5%)<br>Other: 2 (9.5%) | 5.0 (1.0–80.0) | 10 (47.6%) |
| Velasquez et al. (2009) | 93 | 3 (1–4) | Information missing for all lesions | 7.1 (0.4–493) | 54 (58.1%) |
| Minn et al. (1995) | 10 | 1 (1–1) | Lung/mediastinum: 10 (100%) | 42.7 (18.6–231) | 10 (100%) |
| Hatt et al. (2010) | 15 | 1 (1–1) | Esophagus: 15 (100%) | Measurements missing for all lesions | |
| Heijmen et al. (2012) | 20 | 1 (1–3) | Liver: 20 (100%) | Measurements missing for all lesions | |
| Kramer et al. (2016) | 66 | 6.5 (1–15) | Lung/mediastinum: 11 (16.7%)<br>Lymph node: 44 (66.7%)<br>Bone: 6 (9.1%)<br>Other: 5 (7.6%) | 23.3 (1.2–309) | 30(45.4%) |
| ACRIN 6678 | 98 | 3 (1–7) | Lung/mediastinum: 10 (10.2%)<br>Lymph node: 28 (28.6%)<br>Liver: 29 (29.6%)<br>Bone: 16 (16.3%)<br>Other: 15 (15.3%) | 22.5 (0.0–964)<br>Measurements missing for 14 lesions | 63 (64.3%) |
| MERCK MK-046-008 | 90 | 2 (1–6) | Lung/mediastinum: 14 (15.6%)<br>Lymph node: 31 (34.4%)<br>Liver: 31 (34.4%)<br>Bone: 10 (11.1%)Other: 4 (4.4%) | 8.9 (0.3–428) | 54 (60.0%) |
| Rockall et al. (2014) | 83 | 5 (2–5) | Information missing for all lesions | 11.5 (0.8–514) | 63 (75.9%) |

localized in the lung or mediastinum, and 89 (15.5%) were in the liver. Other lesion locations included lymph nodes (67 lesions; 11.7%), bone (28 lesions; 4.9%), and esophagus (17 lesions; 3.0%). The localization of 244 lesions (42.5%) was unknown.

The volumes of these lesions ranged from 0.6 to 963.7 mL, with a mean of 9.2 mL. 163 lesions (28.4%) were smaller than 4.2 mL, 296 (51.6%) were larger, and for 115 (20.0%) unknown. Various thresholding methods (e.g., a fixed percent of maximum voxel, a fixed area around the region of interest, manual delineation) and SUV normalization techniques (body weight or lean body mass) were used.

The number of days between test and retest scans varied from 1 to 10 days, and the amount of time for the injection for each scan ranged from less than one hour to greater than 90 minutes. Summary statistics of these variables are provided in **Tables 1** through **3**.

### Repeatability of the SUV measurements

Bland–Altman plots are given in **Fig. 1** and scatter plots of the percentage of fluctuation a function of test–retest mean are presented in **Fig. 2**. Regression coefficients associated with the test–retest means in the variance–mean relationships were positive for all three SUV measurements in both the univariate [1.73; 95% confidence interval (CI), 1.34–2.12 for SUV$_{max}$, 1.53; 95% CI, 1.13–1.93 for SUV$_{mean}$, and 1.89; 95% CI, 1.41–2.36 for SUV$_{peak}$; see **Table 6**] and multivariate analyses (1.83; 95% CI, 1.45–2.21 for SUV$_{max}$, 1.66,}95% CI, 1.26–2.85) for SUV$_{mean}$, and 1.88, 95% CI, 1.44–2.32 for SUV$_{peak}$; see **Table 8**). Variance–mean relationships of the percentage of fluctuations of the SUV measurements were derived directly from the variance–mean relationships for the SUV measurements themselves (see Appendix). Higher uptake values were associated with worse test–retest repeatability (i.e., higher test–retest variance) of the SUV measurements themselves, as indicated by the positive regression coefficient associated with the test–retest means. But higher uptake values were associated with better test–retest repeatability of percent fluctuations for all three SUV measurements, as indicated by the negative regression coefficients in both the univariate (−0.27}for SUV$_{max}$, −0.47 for SUV$_{mean}$, and −0.11 for SUV$_{peak}$) and multivariate analyses (−0.17}for SUV$_{max}$, −0.34 for SUV$_{mean}$, and −0.12 for SUV$_{peak}$).

According to the multivariate analyses, lesions in the lungs and lymph nodes and single-center studies had the better test–retest repeatability. Lesions in the bone, liver, and those measured in multicenter studies, had worse test–retest repeatability, but the amounts varied on the basis of SUV max, peak or mean. Worse test–retest repeatability of SUV$_{max}$ measurements was also associated with whether the lesion was from a study involving more than two centers (0.711; 95% CI, 0.326–1.10) and bone lesions (0.975; 95% CI, 0.332–1.62). Worse test–retest repeatability of SUV$_{mean}$ measurements was associated with whether the lesion was associated with a study involving more than two centers (1.15; 95% CI, 0.734–1.36) and whether volume measurements were missing (2.86; 95% CI, 1.81–3.39), but better test–retest repeatability was associated with liver lesions (−1.92; 95% CI, −2.81 *to* −1.47). Worse test–retest repeatability of SUV$_{peak}$ was associated with whether the lesion was associated with a study involving more than two centers (0.957; 95%

**Table 4.** Summary statistics of FDG-PET uptake measurements for the individual studies included in the meta-analysis.

| Study | Proportion of lesions evaluable for SUV | Median and range of test–retest differences | | | Median and range of test–retest means among individual patients | | | Median and range of relative test–retest differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ |
| Hoekstra et al. (2002) | 26/27 (96.3%) | 0.0 (−2.2–3.3) | −0.2 (−1.8–1.2) | | 8.4 (4.7–18.5) | 5.6 (3.3–11.3) | | 7.0 (0.5–36.9) | 7.0 (0.8–27.1) | |
| Weber et al. (1999) | 50/50 (100%) | | 0.0 (−0.7–0.9) | | | 4.5 (1.3–10.6) | | | 6.8 (0.4–23.0) | |
| Nahmias and Wahl (2008) | 21/21 (100%) | −0.1 (−1.8–3.4) | 0.0 (−0.4–0.5) | | 8.9 (4.0–23.9) | 5.1 (1.6–17.4) | | 6.1 (0.2–22.7) | 4.3 (0.2–17.7) | |
| Velasquez et al. (2009) | 93/105 (88.6%) | −0.2 (−3.9–4.7) | −0.2 (−2.6–4.4) | | 8.9 (4.2–20.5) | 7.0 (3.4–16.5) | | 11.4 (0.4–42.7) | 10.5 (0.3–41.2) | |
| Minn et al. (1995) | 10/10 (100%) | 0.5 (−1.3–2.8) | 0.4 (−0.9–2.3) | | 9.2 (4.7–19.5) | 8.1 (3.9–16.4) | | 8.8 (2.2–22.1) | 8.0 (1.1–22.8) | |
| Hatt et al. (2010) | 15/17 (88.2%) | −0.6 (−2.4–3.2) | 0.5 (−1.1–4.6) | | 9.1 (5.2–19.0) | 3.7 (1.1–11.5) | | 9.5 (0.9–30.2) | 25.3 (7.0–40.7) | |
| Heijmen et al. (2012) | 20/24 (83.3%) | 0.0 (−3.6–4.5) | 0.1 (−1.5–2.3) | 0.1 (−1.6–2.8) | 8.3 (5.5–17.0) | 5.6 (3.9–10.5) | 6.3 (4.0–12.3) | 12.9 (0.5–35.6) | 10.7 (0.9–30.9) | 13.3 (1.6–38.4) |
| Kramer et al. (2016) | 66/75 (88.0%) | −0.2 (−1.9–1.6) | 0.0 (−1.4–0.7) | −0.1 (−1.7–1.4) | 7.6 (4.1–27.1) | 4.8 (2.6–11.5) | 5.5 (2.7–20.8) | 6.6 (0.3–24.0) | 4.4 (0.1–22.8) | 5.5 (0.1–26.9) |
| ACRIN 6678 | 98/113 (86.7%) | 0.3 (−7.4–6.6) | | 0.1 (−4.6–5.4) | 10.1 (4.6–22.9) | | 7.8 (3.1–18.8) | 10.2 (0.2–48.3) | | 10.1 (0.4–48.7) |
| MERCK MK-046-008 | 90/116 (77.6%) | 0.0 (−5.5–7.7) | | 0.0 (−4.9–5.2) | 7.6 (4.2–17.8) | | 5.5 (3.1–18.8) | 9.2 (0.1–90.8) | | 9.8 (0.0–85.0) |
| Rockall et al. (2014) | 83/87 (95.4%) | 0.0 (−2.8–3.9) | −0.1 (−1.7–2.9) | | 8.7 (4.5–18.5) | 5.8 (2.5–14.6) | | 6.2 (0.0–44.2) | 5.6 (0.0–53.2) | |

Note: Blank cells indicate that the corresponding uptake measurement was not acquired for that study.

**Table 5.** Summary statistics of the test–retest differences in FDG-PET uptake measurements for the individual studies included in the meta-analysis.

| Study | Proportion of lesions evaluable for SUV | Median and range of test–retest differences | | | Median and range of test–retest means among individual patients | | | Median and range of relative test–retest differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ |
| Hoekstra et al. (2002) | 26/27 (96.3%) | 0.0 (−2.2–3.3) | −0.2 (−1.8–1.2) | | 8.4 (4.7–18.5) | 5.6 (3.3–11.3) | | 7.0 (0.5–36.9) | 7.0 (0.8–27.1) | |
| Weber et al. (1999) | 50/50 (100%) | | 0.0 (−0.7–0.9) | | | 4.5 (1.3–10.6) | | | 6.8 (0.4–23.0) | |
| Nahmias and Wahl (2008) | 21/21 (100%) | −0.1 (−1.8–3.4) | 0.0 (−0.4–0.5) | | 8.9 (4.0–23.9) | 5.1 (1.6–17.4) | | 6.1 (0.2–22.7) | 4.3 (0.2–17.7) | |
| Velasquez et al. (2009) | 93/105 (88.6%) | −0.2 (−3.9–4.7) | −0.2 (−2.6–4.4) | | 8.9 (4.2–20.5) | 7.0 (3.4–16.5) | | 11.4 (0.4–42.7) | 10.5 (0.3–41.2) | |
| Minn et al. (1995) | 10/10 (100%) | 0.5 (−1.3–2.8) | 0.4 (−0.9–2.3) | | 9.2 (4.7–19.5) | 8.1 (3.9–16.4) | | 8.8 (2.2–22.1) | 8.0 (1.1–22.8) | |
| Hatt et al. (2010) | 15/17 (88.2%) | −0.6 (−2.4–3.2) | 0.5 (−1.1–4.6) | | 9.1 (5.2–19.0) | 3.7 (1.1–11.5) | | 9.5 (0.9–30.2) | 25.3 (7.0–40.7) | |
| Heijmen et al. (2012) | 20/24 (83.3%) | 0.0 (−3.6–4.5) | 0.1 (−1.5–2.3) | 0.1 (−1.6–2.8) | 8.3 (5.5–17.0) | 5.6 (3.9–10.5) | 6.3 (4.0–12.3) | 12.9 (0.5–35.6) | 10.7 (0.9–30.9) | 13.3 (1.6–38.4) |
| Kramer et al. (2016) | 66/75 (88.0%) | −0.2 (−1.9–1.6) | 0.0 (−1.4–0.7) | −0.1 (−1.7–1.4) | 7.6 (4.1–27.1) | 4.8 (2.6–11.5) | 5.5 (2.7–20.8) | 6.6 (0.3–24.0) | 4.4 (0.1–22.8) | 5.5 (0.1–26.9) |
| ACRIN 6678 | 98/113 (86.7%) | 0.3 (−7.4–6.6) | | 0.1 (−4.6–5.4) | 10.1 (4.6–22.9) | | 7.8 (3.1–18.8) | 10.2 (0.2–48.3) | | 10.1 (0.4–48.7) |
| MERCK MK-046-008 | 90/116 (77.6%) | 0.0 (−5.5–7.7) | | 0.0 (−4.9–5.2) | 7.6 (4.2–17.8) | | 5.5 (3.1–18.8) | 9.2 (0.1–90.8) | | 9.8 (0.0–85.0) |
| Rockall et al. (2014) | 83/87 (95.4%) | 0.0 (−2.8–3.9) | −0.1 (−1.7–2.9) | | 8.7 (4.5–18.5) | 5.8 (2.5–14.6) | | 6.2 (0.0–44.2) | 5.6 (0.0–53.2) | |

Note: Blank cells indicate that the corresponding uptake measurement was not acquired for that study.

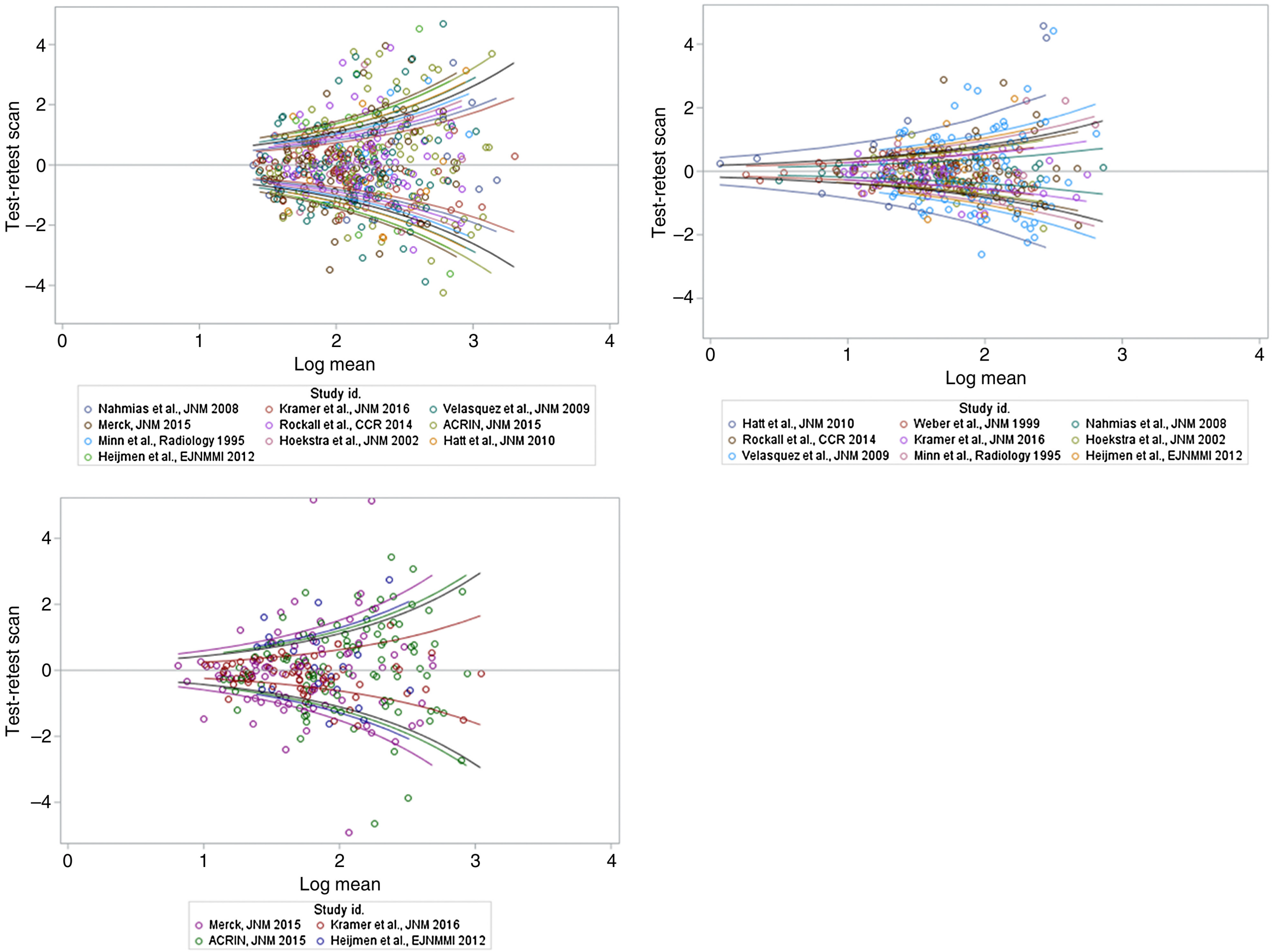

**Figure 1.**

Bland–Altman plots for the SUV parameters, with 95% normal limits by cohort and for all studies included in the meta-analysis. Top left plot, Maximum SUV (578 lesions). Top right plot, Mean SUV (404 lesions). Bottom left plot, Peak SUV (316 lesions).

CI, 0.546–1.37) and bone (1.033; 95% CI, 0.300–1.77) and liver lesions (0.735; 95% CI, 0.123–1.35). These regression coefficients and confidence intervals are the same for the test–retest variance of the percentage of fluctuations due to its relationship between the test and retest variance of the measurements themselves (see Appendix).

These variance–mean relationships were used to estimate the probabilities of repeat measurements differing by more than a prescribed threshold in the absence of any intervention or underlying metabolic changes. For $SUV_{max} \geq 3.5$, differences in such repeat measurements will exceed 30% (threshold chosen on the basis of PERCIST) and 1.05 U (i.e., $0.3 \times 3.5$) with probability no greater than 0.067. Similarly, for $SUV_{mean} \geq 3.5$ and $SUV_{peak} \geq 3.5$, differences in such repeat measurements will exceed 30% and 1.05 U with probability no greater than 0.023 and 0.056, respectively. Meanwhile, for $SUV_{max} \geq 4.0$, $SUV_{mean} \geq 4.0$, and $SUV_{peak} \geq 4.0$, differences in such repeat measurements will exceed 30% and 1.2 units with probabilities no greater than 0.063, 0.019, and 0.054, respectively. These probabilities are summarized in **Table 7**. On the basis of these results, the minimum demonstrable change in tumor metabolism can be reliably seen in target lesions (minimum size of 1 $cm^3$) with a baseline of $SUV_{peak} \geq 4.0$ with a change of 30% (absolute change of 1.2). Requiring a baseline FDG uptake to have baseline of $SUV_{peak} \geq 4.0$ will limit the assessment of lesions with low metabolic uptake but is needed for reliable assessment of changes in tumor metabolism. Prior studies have demonstrated the difficulty in assessing response in lesions that start with a low metabolism (30). This has been partially addressed by performing a dynamic assessment of FDG uptake kinetics in Doot and colleagues (31) but is problematic for routine FDG-PET CT acquisitions.

On the basis of (A.4), a change of 30% in $SUV_{max}$ corresponds to a test–retest $SUV_{max}$ mean of 5.9, whereas a change of 25% corresponds to a test–retest mean of 22.6. Because of the monotonically decreasing relationship between change and test–retest mean, test–retest means above 5.9 will correspond to changes less than 30%; therefore, provided the test–retest mean is at least 5.9, a 30% change (1.8 $SUV_{max}$ units) or more will be strong evidence of actual underlying metabolic phenomena. Meanwhile, on the basis of these variance–mean relationships, a change of 30% in $SUV_{mean}$ corresponds to a test–retest $SUV_{mean}$ mean of 1.9 and a change of 25% corresponds to a test–retest mean of 4.1, whereas a change of 30% in $SUV_{peak}$ corresponds to

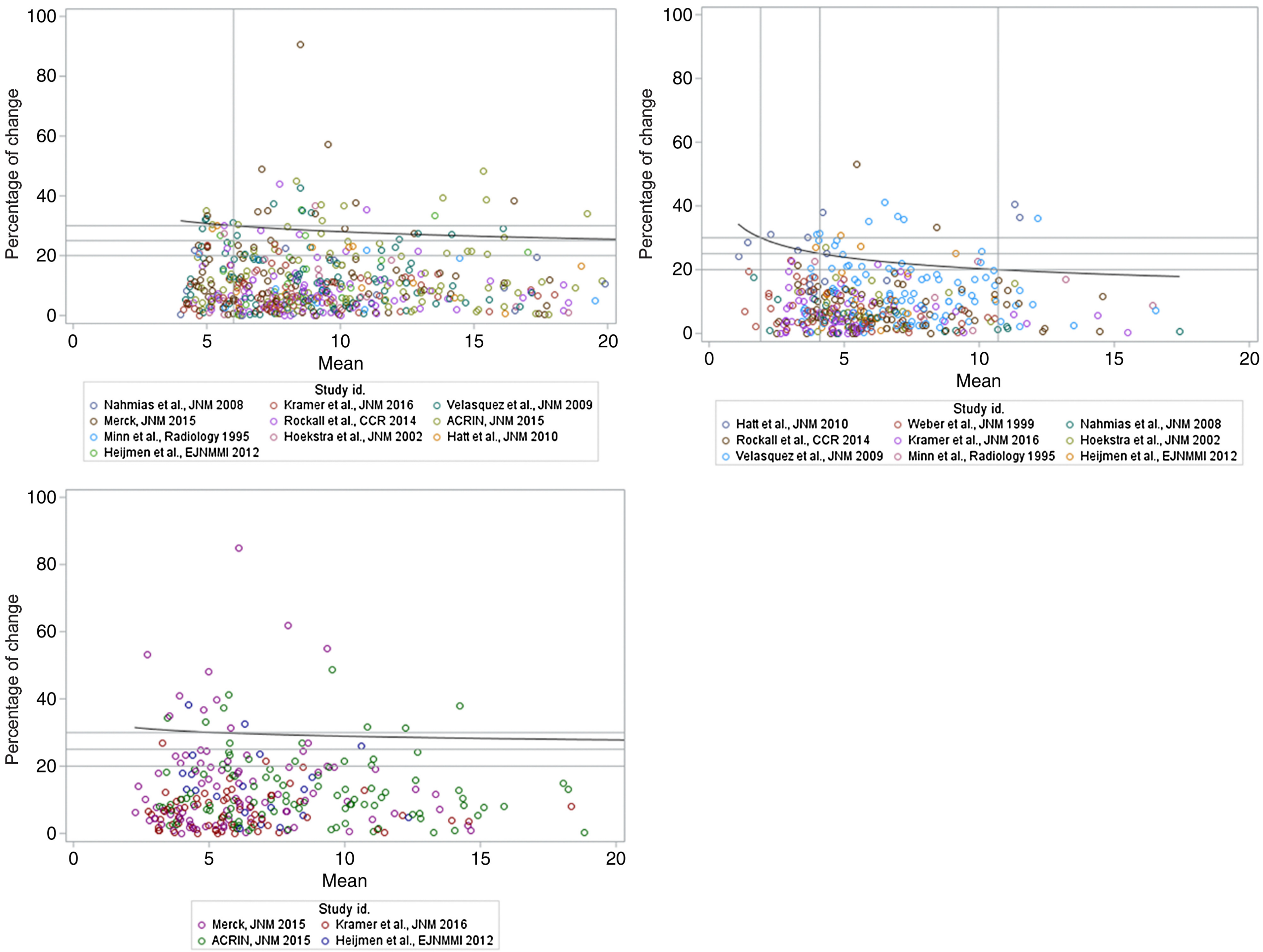

**Figure 2.**

Plot of the percentage of change versus test–retest mean for the different SUV parameters. Top left plot, Maximum SUV (578 lesions). Top right plot, Mean SUV (404 lesions). Bottom left plot, Peak SUV (316 lesions).

a test–retest $SUV_{peak}$ mean of 5.5 and a change of 25% corresponds to a test–retest mean of 150.6; similar reasoning produces the thresholds for declaring underlying change as described previously in §3.2.

## Discussion

An analysis of the test–retest repeatability of FDG $SUV_{max}$, $SUV_{mean}$, and $SUV_{peak}$ can provide insight into what types of changes are likely to constitute actual underlying metabolic phenomena versus measurement noise. Many test–retest repeatability studies involve sample sizes too small for any conclusive findings. Hence, data from multiple such studies in the literature were combined. However, because of the heterogeneity in baseline patient characteristics, the acquisition protocol and image reconstruction and processing, and the SUV measurement computation from study to study, a simple meta-analysis could have led to misleading results. Thus, meta-regressions to examine test–retest variance in these SUV measurements as a function of these factors were performed.

These meta-regressions provide insight into several important considerations in the establishment of a set of FDG-PET–based response criteria. These analyses can be used to select which of the three SUV measurements should be used in such criteria. These results can be used to estimate true and false-positive rates associated with prespecified definitions of progressive disease, partial response, complete response, and stable disease in identifying presence or absence of actual underlying metabolic changes. The variance–mean relationship and the multivariate meta-regression results can be used to assess whether these response category definitions are broadly applicable for a wide range of uptake values, lesion locations, and lesion sizes.

**Table 6.** Parameter estimates and corresponding *P* values for the mean–variance relationship.

| | Parameter estimates with 95% confidence intervals | | |
|---|---|---|---|
| | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ |
| Model intercept | −3.97 (−4.99 to −2.94) | −4.14 (−5.16 to −3.13) | −4.26 (−6.03 to −2.49) |
| Log test–retest mean regression coefficient | 1.73 (1.34–2.12) | 1.53 (1.13–1.93) | 1.89 (1.41–2.36) |

**Table 7.** Estimated probabilities of repeat measurements exceeding the indicated thresholds in the absence of any intervention or underlying metabolic change.

| Thresholds | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ |
|---|---|---|---|
| **≥30% change, SUV  ≥3.5 (absolute change of 1.05)** | 0.067 | 0.023 | 0.056 |
| **≥30% change, SUV ≥4.0 (absolute change of 1.2)** | 0.063 | 0.019 | 0.054 |
| **≥25% change, SUV ≥3.5 (absolute change of 0.88)** | 0.128 | 0.059 | 0.111 |
| **≥25% change, SUV ≥4.0 (absolute change of 1.0)** | 0.121 | 0.051 | 0.108 |

Note: If these thresholds are used to differentiate partial response versus stable disease versus progressive disease, these probabilities are false-positive rates (i.e., probabilities that stable disease is declared when there is actual underlying increase or decrease in metabolic activity).

It was shown that the PERCIST definitions of progressive disease versus stable disease versus partial response, namely a 30% increase or decrease in $SUV_{peak}$ for $SUV_{peak} \geq 3.5$, were associated with a false-positive rate (i.e., probability of incorrectly indicating disease progression or treatment response when no changes in metabolism actually occurred) no greater than 0.056. Analogous definitions using $SUV_{max}$ and $SUV_{mean}$ were associated with false-positive rates no greater than 0.067 and 0.023, respectively. Although low false-positive rates are desirable, values that are too low (e.g., markedly lower than 0.05) typically mean excessive conservatism in declaring progressive disease or partial response, namely an increased tendency to indicate stable disease when underlying changes in metabolism did indeed occur. Such conservatism can be seen in definitions based on $SUV_{mean}$, suggesting that a lower threshold (e.g., 25%) may be more appropriate here.

In order for response category definitions to be broadly applicable in a wide range of scenarios, the test–retest repeatability should ideally be associated with very few FDG uptake variables, image acquisition, and processing aspects, or study- and patient-level baseline characteristics. The test–retest repeatability for all three SUV measurements was associated with the lesion's location and that of $SUV_{mean}$ was associated with lesion volume. This could entail the need for different response category definitions for bone or liver lesions, and, in the case of $SUV_{mean}$, larger-sized lesions that would be impractical. Test–retest repeatability of the percentage of fluctuations improved with increasing uptake levels for all three SUV measurements. However, the repeatability of $SUV_{peak}$ was more robust to uptake values than the repeatability of the other two SUV metrics were; in the multivariate regression analyses, the coefficient associated with the percentage of fluctuations in $SUV_{peak}$ was closer to zero than those associated with the percentage of fluctuations in $SUV_{max}$ and $SUV_{mean}$.

Thus, some preliminary recommendations regarding thresholds defining the different response categories can be made on the basis of the results of this meta-analysis, including whether those from PERCIST (progressive disease and partial response corresponding to an increase in $SUV_{peak}$ by 30% and a decrease in $SUV_{peak}$ by 30% or more, respectively, with $SUV_{peak} \geq 3.5$) may be adopted. However, instead of the requirement of absolute change of 0.8 $SUV_{peak}$ units for progressive disease and partial response prescribed in PERCIST, a more stringent requirement of 1.1 U is recommended (a 30% or greater change in $SUV_{peak}$ for $SUV_{peak} \geq 4.0$ translates to an absolute change of 1.1 or greater). Response criteria based on $SUV_{peak}$ may need to differ slightly for bone and liver lesions. RECIST considers bone lesions as unevaluable and as such not followed for size changes. EORTC criteria propose a 25% change threshold (9). However, application of this lower threshold to $SUV_{max}$ or $SUV_{peak}$ will result in an inflated false-positive rate (0.128 and 0.111 respectively).

These analyses suffered from some limitations involving the availability of test–retest repeatability and image acquisition and processing and study- and patient-level characteristic data for FDG

**Table 8.** Model intercept and regression coefficient estimates from multivariate models of the test–retest variance of maximum, mean, and peak SUV given the baseline characteristics, with *P* values.

| Baseline characteristic and characteristic levels | | Regression coefficient estimates and 95% confidence intervals | | |
|---|---|---|---|---|
| | | $SUV_{max}$ | $SUV_{mean}$ | $SUV_{peak}$ |
| Model intercept | | −4.69 (−5.55 to − 3.84) | −4.76 (−5.41 to − 4.43) | −5.17 (−6.16 to − 4.19) |
| Log test–retest mean | | 1.83 (1.45–2.21) | 1.66 (1.26–2.85) | 1.88 (1.44–2.32) |
| Two or more centers | No | Baseline | Baseline | Baseline |
| | Yes | 0.711 (0.326–1.10) | 1.15 (0.734–1.36) | 0.957 (0.546–1.37) |
| Lesion localization | Lung | Baseline | Baseline | Baseline |
| | Bone | 0.975 (0.332–1.62) | 0.384 (−0.650–0.911) | 1.033 (0.300–1.77) |
| | Esophagus | 0.125 (−0.826–1.08) | −0.972 (−2.14 to − 0.379) | |
| | Liver | 0.177 (−0.353 to 0.706) | −1.92 (−2.81 to − 1.47) | 0.735 (0.123–1.35) |
| | Lymph node | −0.144 (−0.599–0.312) | −0.330 (−0.803 to − 0.09) | 0.157 (−0.746–0.431) |
| | Other | −0.126 (−0.813–0.562) | −0.228 (−1.32–0.328) | −0.388 (−1.18–0.403) |
| | Unknown | 0.006 (−0.518–0.529) | | |
| Lesion volume greater than 4.2 mL | No | Baseline | Baseline | |
| | Yes | 0.015 (−0.306–0.336) | 0.092 (−0.475–0.104) | |
| | Missing values | 0.535 (−0.128–1.20) | 2.86 (1.81–3.39) | |

Note: Baseline characteristics omitted from this table were the ones not included in any of the multivariate models. Blank cells indicate that the quantity was not estimable, and the variable was thus automatically excluded from the model.

SUV measurements. Measurements of the different FDG-PET parameters could not be harmonized because of lack of access to the scans themselves. Certain baseline characteristics (e.g., scanner vendor, acquisition type, and SUV normalization technique) were unavailable for some patients, and the reason for the data gaps is unclear; therefore, the analyses may be restricted to the specific set of scenarios. Not all studies obtained data on all three SUV measurements, and this often prevented the assessment of the association of the test–retest repeatability with some baseline characteristics; for example, no lesion volumes were measured in any studies in which SUV$_{\text{peak}}$ data were obtained. In addition to these analyses, there are other important considerations in the development of FDG-PET–based response criteria.

Data were collected using different PET and PET-CT generation systems. The repeatability of these systems may differ in small degrees and will continue to do so as technology advances. However, repeatability is obtained by scanning the same patient on the same system, using the same reconstruction method and setting and analyzed using the same software and methods, thus mitigating the effect of many of the SUV uncertainties on repeatability (ref. 16; e.g., the effect of uptake time variations, biological factors, and so on are the same regardless of system used). In fact, in our analyses, we could not identify a scanner or site effect. The main factors associated with repeatability were the quantitative metrics (maximum, peak, or mean) and whether the data came from a single or multicenter study.

The use of SUV$_{\text{peak}}$ as a quantitative metric for assessing metabolic changes should be considered. SUV$_{\text{peak}}$ is less sensitive to image noise and therefore less affected by differences in administered activities and scanner sensitivities (32, 33) than SUV$_{\text{max}}$. Moreover, SUV$_{\text{peak}}$ varies less with image reconstruction methodologies and settings (27, 34, 35), and we therefore expect (as was also found) that scanner or site effects can be neglected. Although SUV$_{\text{mean}}$ is also calculated by averaging over an extended tumor area and may be less sensitive to noise as well, we observed that SUV$_{\text{mean}}$ repeatability was worse compared with that of SUV$_{\text{peak}}$. SUV$_{\text{peak}}$ and SUV$_{\text{max}}$ are not affected by the tumor segmentation used nor by segmentation uncertainties seen with semi-automated segmentations (27, 36), whereas this is not the case for SUV$_{\text{mean}}$. For larger tumors, typically showing increased levels of uptake heterogeneity, segmentation variability is likely to increase, resulting in an even more reduced repeatability of SUV$_{\text{mean}}$ as compared with SUV$_{\text{peak}}$. Consequently, for multicenter studies, the use of SUV$_{\text{peak}}$ is strongly recommended. One potential reason for why there is larger variability in SUV$_{\text{mean}}$ than SUV$_{\text{peak}}$ and SUV$_{\text{max}}$, is the variability in segmentation performance. Test–retest of PET-based volumes is around 35% (27). Consequently, SUV$_{\text{mean}}$ is affected by segmentation uncertainties, whereas SUV$_{\text{max}}$ and SUV$_{\text{peak}}$ are not. Technical and biological uncertainties are the same for all three metrics, but SUV$_{\text{mean}}$ has an additional factor related to segmentation uncertainties. Because SUV$_{\text{max}}$ is more sensitive to image noise than SUV$_{\text{peak}}$, it makes sense that SUV $_{\text{peak}}$ is demonstrating the best test retest repeatability. In addition, SUV$_{\text{peak}}$ is also less sensitive to small variations in image quality (different scanners, different reconstruction etc.), so overall, SUV$_{\text{peak}}$ would be the most precise and most reproducible SUV-metric.

Although a lesion-level meta-regression was needed to identify thresholds above which changes can be attributable to actual underlying phenomena, patient-level analyses will be needed in the future to further solidify how to use this information in response assessment in practice. The number of target lesions to follow (e.g., a maximum of five lesions similar to RECIST 1.1) and which lesions to consider the target lesions (e.g., those with the highest baseline uptakes) need to be specified. A minimum lesion size (e.g., 1.0 cm$^3$ as prescribed by PERCIST) should also be included because of the difficulty of reliably measuring very small lesions. A pilot study by Kramer and colleagues (27) considered the use of one, five (as done in PERCIST), or all lesions when assessing repeatability. This study showed that using multiple lesions to measure metabolic changes improves repeatability. However, in such a scenario, all target lesions are assumed to have more or less the same metabolic responses. A procedure on how to handle mixed response, namely heterogeneous responses across lesions in which uptake increases in some lesions but decrease in others, will need to be developed. For example, partial response may require that a decrease in SUV$_{\text{peak}}$ by a prespecified amount is observed in all target lesions, and progressive disease may require that an increase in SUV$_{\text{peak}}$ by a prespecified amount is observed in at least one target lesion. Also, the majority of the patients included in the test–retest studies and hence in this meta-analysis had either lung or colorectal cancers. Whether other histologies have lesions affected by similar factors (i.e., lesion size, level of metabolic activity, and location) will need to be assessed in future studies.

Further data collection and analyses are necessary to develop a consensus on how to handle these considerations. This meta-analysis provides the parameters for minimal detectable metabolic change in a tumor that should be applied to response criteria that are developed to assess clinically significant outcomes. Importantly, a lack of decrease in SUV$_{\text{peak}}$ by 30% in a lesion with a baseline of $\text{SUV}_{\text{peak}} \geq 4.0$ signifies a lack of demonstrable improvement/response on a therapeutic regimen. This is significant information for patient management in both clinical trials as well as clinical care. Important aspects of any response criteria assessed in future prospective data analyses should include whether a set of response criteria is associated with more definitive clinical outcomes and whether the use of FDG-PET/CT will improve or significantly change response assessment performed by RECIST. These data could be acquired as part of clinical trials and registries to obtain the volume of information needed for these assessments.

## Authors' Disclosures

L.H. Schwartz reports other support from Merck, personal fees from Eisai and Regeneron, and grants from Janssen and Boehringer outside the submitted work. S. Collette reports personal fees from Bristol Myers Squibb outside the submitted work. L. Seymour reports personal fees from Boehringer Ingelheim, grants and other support from AstraZeneca, grants from Bayer, non-financial support from Pfizer, as well as grants from Merck and Roche, and other support from Keosys outside the submitted work. E.G.E. deVries reports other support from NSABP, Daiichi Sankyo, and Crescendo Biologics, and grants from Amgen, Genentech, Roche, CytomX, G1 Therapeutics, Bayer, Synthon, Servier, Regeneron, and Crescendo outside the submitted work; and reports Non-financial interests, including non-remunerated activities and public positions: Member ESMO-MCBS, Chair ESMO Cancer Medicines Working Group, Co-chair RECIST committee, and member of expert panel for selection of essential medicine list, WHO. No disclosures were reported by the other authors.

## Authors' Contributions

**L.K. Shankar:** Conceptualization, formal analysis, supervision, methodology, writing–original draft, project administration, writing–review and editing. **E. Huang:** Data curation, formal analysis, methodology, writing–original draft, writing–review and editing. **S. Litiere:** Data curation, software, formal analysis, methodology, writing–review and editing. **O.S. Hoekstra:** Resources, supervision, writing–original draft, project administration, writing–review and editing. **L. Schwartz:** Data curation, investigation, writing–original draft, writing–review and editing. **S. Collette:** Data curation, formal analysis, methodology, writing–original draft. **R. Boellaard:** Formal analysis, investigation, writing–original draft, writing–review and editing. **J. Bogaerts:** Resources, supervision, methodology, writing–review and editing. **L. Seymour:** Conceptualization, supervision, investigation,

methodology, writing–original draft. **E.G.E. deVries:** Conceptualization, supervision, writing–original draft, project administration, writing–review and editing.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

## Note

Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Received September 9, 2021; revised January 28, 2022; accepted October 24, 2022; published first October 27, 2022.

## References

1. Cheson BD. Role of functional imaging in the management of lymphoma. J Clin Oncol 2011;29:1844–54.
2. Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutchings M, Müeller SP, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. J Clin Oncol 2014;32:3048–58.
3. Weiler-Segie M, Bushelev O, Epelbaum R, Dann EJ, Haim N, Avivi I, et al. 18F-FDG avidity in lymphoma addressed: a study of 766 patients. J Nucl Med 2010;51:25–30.
4. Cheson BD, Horning SJ, Coiffier B, Shipp MA, Fisher RI, Connors JM, et al. Report of an International Workshop to standardize response criteria for non-Hodgkin's lymphomas. J Clin Oncol 1999;17:1244–53.
5. Cheson BD, Pfistner B, Juweid ME. Revised response criteria for malignant lymphoma. J Clin Oncol 2007;25:579–86.
6. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol 2014;32:3059–68.
7. Younes A, Hilden P, Coiffier B, Hagenbeek A, Salles G, Wilson W, et al. International working group consensus response evaluation criteria in lymphoma (RECIL 2017). Ann Oncol 2017;28:1436–47.
8. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumors: revised RECIST guideline (Version 1.1). Eur J Cancer 2009;45:228–47.
9. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma AA, et al, for the European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. Measurement of clinical and subclinical tumour response using [$^{18}$F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. Euro J Cancer 1999;35:1773–82.
10. Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, et al. Consensus recommendations for the use of $^{18}$F-FDG–PET as an indicator of therapeutic response in patients in national cancer institute trials. J Nucl Med 2006;47:1059–66.
11. Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG. FDG-PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. Eur J Nucl Med Mol Imaging 2010;37:181–200.
12. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG-PET/CT: EANM procedure guidelines for tumor imaging: version 2.0. Eur J Nucl Med Mol Imaging 2015;42:328–54.
13. Boellaard R, Oyen WJG, Hoekstra CJ, Hoekstra OS, Visser EP, Willemsen AT, et al. The Netherlands protocol for standardization of FDG whole body PET studies in multi-center trials. Eur J Nucl Med Mol Imaging 2008;35:2320–33.
14. The American College of Radiology. Uniform protocols for imaging in clinical trials. Philadelphia, PA: American College of Radiology; 2006.
15. 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.
16. de Langen AJ, Vincent A, Velasquez LM, van Tinteren H, Boellaard R, Shankar LK, et al. Repeatability of $^{18}$F-FDG uptake measurements in tumors: a meta-analysis. J Nucl Med 2012;53:701–8.
17. Kinahan PE, Perlman ES, Sunderland JJ, Subramaniam R, Wollenweber SD, Turkington TG, et al. The QIBA profile for FDG-PET/CT as an imaging biomarker measuring response to cancer therapy. Radiology 2020;294:647–57.
18. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. JASA 1993;889–25.
19. Velasquez LM, Boellaard R, Kollia G, Hayes W, Hoekstra OS, Lammertsma AA, et al. Repeatability of $^{18}$F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. J Nucl Med 2009;50:1646–54.
20. Weber WA, Gatsonis CA, Mozley PD, Hanna LG, Shields AF, Aberle DR, et al. Repeatability of $^{18}$F-FDG PET/CT in advanced non–small cell lung cancer: prospective assessment in two multicenter trials. J Nucl Med 2015;56: 1137–43.
21. Phase I imaging study evaluating dalotuzumab (MK0646) in combination with erlotinib for patients with non-small cell lung cancer (MK-0646–008). Available from: https://clinicaltrials.gov/ct2/show/NCT00729742 (2016; accessed January 13, 2021).
22. Rockall AG, Avril A, Lam R, Iannone R, Mozley PD, Parkinson C, et al. Repeatability of quantitative FDG-PET/CT and contrast-enhanced CT in recurrent ovarian carcinoma: test–retest measurements for tumor FDG uptake, diameter, and volume. Clin Cancer Res 2014;20;2751–60.
23. Hoekstra CJ, Hoekstra OS, Stroobants SG, Vansteenkiste J, Nuyts J, Smit EF, et al. Methods to monitor response to chemotherapy in non–small cell lung cancer with $^{18}$F-FDG PET. J Nucl Med 2002;43:1304–9.
24. Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, Schwaiger M. Reproducibility of metabolic measurements in malignant tumors using FDG-PET. J Nucl Med 1999;40:1771–7.
25. 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.
26. Heijmen L, de Geus-Oei L, de Wilt J, Visvikis D, Hatt M, Visser EP, et al. Reproducibility of functional volume and activity concentration in $^{18}$F-FDG PET/CT of liver metastases on colorectal cancer. Eur J Nucl Med Mol Imaging 2012;39:1858–67.
27. Kramer GM, Frings V, Hoetjes N, Hoekstra OS, Smit EF, de Langen AJ, et al. Repeatability of quantitative whole-body $^{18}$F-FDG PET/CT uptake measures as function of uptake interval and lesion selection in non–small cell lung cancer patients. J Nucl Med 2016;57:1343–9.
28. Hatt M, Cheze-Le Rest C, Aboagye EO, Kenny LM, Rosso L, Turkheimer FE, et al. Reproducibility of $^{18}$F-FDG and 3'-deoxy-3'-$^{18}$F-fluorothymidine PET tumor volume measurements. J Nucl Med 2010;51:1368–76.
29. Nahmias C, Wahl LM. Reproducibility of standardized uptake value measurements determined by $^{18}$F-FDG–PET in malignant tumors. J Nucl Med 2008;49: 1804–8.
30. McDermott GM, Welch A, Staff RT, Gilbert FJ, Schweiger L, Semple SIK, et al. Monitoring primary breast cancer throughout chemotherapy using FDG-PET. Breast Cancer Res Treat 2007;102:75–84.
31. Doot RK, Dunnwald LK, Schubert EK, Muzi M, Peterson LM, Kinahan PE, et al. Dynamic and static approaches to quantifying 18F-FDG uptake for measuring cancer response to therapy, including the effect of granulocyte CSF. J Nucl Med 2007;48:920–5.
32. Boellard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. J Nucl Med 2004;45:1519–27.
33. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. J Nucl Med 2012;53: 1041–7.
34. Mansor S, Pfaehle E, Heijtel D, Lodge MA, Boellard R, Yaqub M. Impact of PET/CT system, reconstuction, protocol, data analysis method, and repositioning on PET/CT precision: an experimental evaluation using an oncology and brain phantom. Med Phys 2017;44:6413–24.
35. Kaalep A, Burggraaff CN, Pieplenbosch S, Verwer EE, Sera T, et al. Quantitative implications of the updated EARL 2019 PET-CT performance standards. EJNNMI Phys 2019;6:28.
36. Kolinger GD, Vállez García D, Kramer GM, Frings V, Smit EF, de Langen AJ, et al. Repeatability of [18F] FDG-PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. EJNNMI Res 2019;9:14.