

Quantitation of cancer treatment response by 2-[¹⁸F]FDG PET/CT: multi-center assessment of measurement variability using AUTO-PERCIST™

JooHyun O

Catholic University of Korea

Su Jin Lim

Johns Hopkins University

Hao Wang

Johns Hopkins University

Jeffrey P. Leal

Johns Hopkins University

Hui-Kuo G. Shu

Emory University

Richard L. Wahl (✉ rwahl@wustl.edu)

Washington University <https://orcid.org/0000-0002-7306-2590>

Original research

Keywords: [¹⁸F]FDG PET/CT, response assessment, quantification

Posted Date: January 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-93282/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at EJNMMI Research on February 12th, 2021.
See the published version at <https://doi.org/10.1186/s13550-021-00754-1>.

Abstract

Background: The aim of this study was to assess the reader variability in quantitatively assessing pre- and post-treatment 2-deoxy-2-[¹⁸F]fluoro-D-glucose positron emission tomography/computed tomography ([¹⁸F]FDG PET/CT) scans in a defined set of images of cancer patients using the same semi-automated analytical software (Auto-PERCIST™), which identifies tumor peak standard uptake value corrected for lean body mass (SUL_{peak}) to determine [¹⁸F]FDG PET quantitative parameters.

Methods: Paired pre- and post-treatment [¹⁸F]FDG PET/CT images from 30 oncologic patients and Auto-PERCIST™ semi-automated software were distributed to 13 readers across US and international sites. One reader was aware of the relevant medical history of the patients (read_{reference}), whereas the 12 other readers were blinded to history but had access to the correlative images. Auto-PERCIST™ was set up to first automatically identify the liver and compute the threshold for tumor measurability (1.5 x liver mean) + (2 x liver standard deviation [SD]), and then detect all sites with SUL_{peak} greater than the threshold. Next, the readers selected sites they believed to represent tumor lesions. The main performance metric assessed was the percent change in the SUL_{peak} (%ΔSUL_{peak}) of the hottest tumor identified on the baseline and follow up images.

Results: The intra-class correlation coefficient (ICC) for the %ΔSUL_{peak} of the hottest tumor was 0.87 (95%CI: [0.78, 0.92]) when all reads were included (n=297). Including only the measurements that selected the same target tumor as the read_{reference} (n=224), the ICC for %ΔSUL_{peak} was 1.00 (95%CI: [1.00, 1.00]). The Krippendorff alpha coefficient for response (complete or partial metabolic response, versus stable or progressive metabolic disease on PET Response Criteria in Solid Tumors 1.0) was 0.91 for all reads (n=380), and 1.00 including for reads with the same target tumor selection (n=270).

Conclusion: Quantitative tumor [¹⁸F]FDG SUL_{peak} changes measured across multiple global sites and readers utilizing Auto-PERCIST™ show very high correlation. Harmonization of methods to single software, Auto-PERCIST™, resulted in virtually identical extraction of quantitative tumor response data from [¹⁸F]FDG PET images when the readers select the same target tumor.

Introduction

2-deoxy-2-[¹⁸F]fluoro-D-glucose positron emission tomography/computed tomography ([¹⁸F]FDG PET/CT) is increasingly applied in monitoring treatment response in patients with cancer. While PET is intrinsically a quantitative imaging technique, many PET assessments of cancer response are qualitative, as for example in lymphoma where quantitative PET data are converted into a 5 point qualitative scale which is practical and highly useful (1,2). Quantitative PET assessments of response have been deployed in many research imaging studies, especially in examining early treatment response related changes in metabolism including breast cancer where these changes can predict much later pathological outcomes (3,4). The PET Response Criteria in Solid Tumors 1.0 (PERCIST 1.0) were

proposed in 2009 as a method to standardize the assessment of tumor response on [^{18}F]FDG PET and emphasized use of the peak standard uptake value corrected for lean body mass (SUL_{peak}) in contrast to the maximum standardized uptake value (SUV_{max}) (5,6). While the SUV_{max} is reasonably easy to determine with many forms of software though the SUL_{peak} is more challenging to measure (7).

Thus despite its attractiveness, quantitative PET utilizing PERCIST is not routinely performed for assessing response to therapy in patients with cancer in the clinic or many clinical trial settings, contrary to the routinely utilized Response Evaluation Criteria in Solid Tumors for assessment of anatomical imaging. One way to expand the use of quantitative [^{18}F]FDG PET/CT in clinical trials and clinical practice is to reduce reader variability of SUV measurements and make the measurements rapid and automated. In a previous multi-center, multi-reader study we conducted, multiple sites assessed the same paired pre- and post-treatment [^{18}F]FDG PET/CT images in cancer patients. The intra-class correlation coefficient (ICC) of percent change in SUV_{max} was 0.89 (95% confidence interval (CI): [0.81, 0.94]) across multiple performance sites using a variety of analytical software tools. The ICC for the SUL_{peak} was lower at 0.70 (95% CI: [0.54, 0.80]). SUL_{peak} is, in principle, the more statistically sound of the PET parameters and it is the suggested metric in PERCIST (7). However, if there is considerable variability among sites in how SUL_{peak} is generated and measured, then the PERCIST metric potentially may introduce variability into assessments of treatment response, as opposed to reducing variability (8).

The aim of the present study was to determine if the utilization of Auto-PERCIST™, a semi-automated software system for the quantitative assessment [^{18}F]FDG PET images, could lower the reader variability in quantitatively assessing pre- and post-treatment [^{18}F]FDG PET/CT studies for response in a multi-center, multi-reader, multi-national study assessing identical images.

Materials And Methods

Pre- and post-treatment [^{18}F]FDG PET/CT images of 30 oncologic patients selected from a group of tumor types having representative patterns of FDG-avidity contained a mix of single and multiple tumors on the pretreatment scan (1 tumor, n=6; >1 but < 10 tumors, n=19; \geq 10 tumors, n=5), and a mix of the four major response categories using PERCIST (complete metabolic response, n=6; partial metabolic response, n=11; stable metabolic disease, n=4; and progressive metabolic disease, n=9).

Sites both with National Cancer Institute Quantitative Imaging Network affiliation and without which did not participate in the previous study with the same data set were recruited by email and conference calls. The dataset was based on a previous study of reader variability (9).

Thirty anonymized cases of pre- and post-treatment [^{18}F]FDG PET/CT studies (total 60 studies) were distributed along with directions for installing and utilizing the Auto-PERCIST™ software. Approval from the Johns Hopkins Institutional Review Board was obtained, and the need for patient informed consent was waived for this study of anonymized image data.

Measurement

Individual measurements from coupled pre- and post-treatment [¹⁸F]FDG PET/CT images from one patient were counted as a *read*. The coupled pre- and post-treatment measurements for all 30 cases from a single reader were counted as a *set of reads*. One reader from the central site (reader 1) had full knowledge of the primary tumors, treatment histories and subsequent follow-up results, but all other readers had no knowledge of the patients' medical histories as the reader is often intentionally blinded in the setting of multicenter trials. For statistical purpose, the measurements by reader 1 were considered as the reference standard for comparison ($read_{reference}$).

Each reader determined which tumor to measure. The Auto-PERCIST™ loads the PET images and automatically obtains liver measurements from a 3 cm diameter sphere in the right side of the liver to compute the threshold for lesion detection. The default setting is 1.5 x liver mean + 2 standard deviations (SD) at baseline to ensure the decline in [¹⁸F]FDG uptake is less likely due to chance and to minimize overestimation of response or progression. For follow up images, the default setting is lower at 1.0 x liver mean + 2SD, to allow detection of lesions with lower SUL_{peak}. If a lesion was perceptible visually but not detected using the default threshold settings, the reader had the choice to manually lower the threshold for detection. The Auto-PERCIST™ would detect all sites with SUL_{peak} higher than the threshold (Figure 1). It was up to the readers to determine whether the detected sites were true tumor lesions or not. The reader could also separate a detected focus of [¹⁸F]FDG uptake into separate smaller lesions when needed – to exclude adjacent physiologic [¹⁸F]FDG uptake or break down a large conglomeration of tumors into smaller separate lesions. The reader could also add smaller [¹⁸F]FDG uptake lesions to make them a single lesion if the reader decided the separate [¹⁸F]FDG uptakes were parts of a larger single lesion. The readers were instructed to select up to 5 of the hottest tumors for cases with multiple lesions. The readers could view the PET/CT images on any reading software they preferred, but the measurements came only from the Auto-PERCIST™. The measurements from Auto-PERCIST™ included SUL_{peak}, maximum and mean SUL, number of counts, geometric mean, exposure, kurtosis, skewness, and metabolic volume. After the readers selected and quantified the lesions, the measurements were saved as text files and sent for central compilation and analysis to the Image Response Assessment Core at Johns Hopkins University.

Statistical analysis

The primary study metric was the percentage change in SUL_{peak} ($\% \Delta SUL_{peak}$) from baseline to follow-up. Percentage change was defined as $[(\text{follow-up measurement} - \text{baseline measurement}) / (\text{baseline measurement})] \times 100$. For assessment of up to 5 lesions, the percentage change was computed from the sum of the lesions. Treating both case and site as random-effects, a linear random-effects model was fit via the restricted maximum likelihood estimation method, which estimated variance components of the

random-effects in the model. As a measure of inter-rater agreement, the intra-class correlation coefficient (ICC) was computed using the variance components of the random-effects. The ICC was computed as $[\text{inter-subject variance} / (\text{inter-subject variance} + \text{intra-subject variance} + \text{residual variance})]$. The bias-corrected and accelerated bootstrap method was implemented with 1,000 bootstrap replicates to construct the 95% confidence interval of the computed ICC. The sampling unit was a *read*.

To assess agreement between the reference reader ($\text{read}_{\text{reference}}$) and another reader, the ICC was computed for each pair of the reference reader and 12 other readers. The mean of these ICCs and its range (minimum, maximum) were reported.

Krippendorff alpha reliability coefficient was computed as a measure of agreement between multiple readers for response outcome, which was classified into four ordered major response categories using PERCIST 1.0 as: complete metabolic response (CMR), partial metabolic response (PMR), stable metabolic disease (SMD), and progressive metabolic disease (PMD). The measurements were classified: PMD for SUL_{peak} increase $\geq 30\%$ (and 0.8 units) or new lesions; SMD for SUL_{peak} increase or decrease $< 30\%$ (or 0.8 units); PMR for SUL_{peak} decrease $\geq 30\%$ (and 0.8 units); and CMR for no perceptible tumor lesion. Additionally, Krippendorff coefficient was computed with the response categories being dichotomized into two levels: clinical benefit (CMR/PMR/SMD) and no benefit (PMD); or response (CMR/PMR) and no-response (SMD/PMD). Krippendorff suggests 0.8 as a threshold for satisfactory reliability, but if tentative conclusions are acceptable, 0.667 is the lowest conceivable threshold (10).

Results

All reads

Reads were received from 13 different sites from January to September of 2018. A single reader (nuclear medicine physician/radiologist/radiological scientist) at each site measured all 30 cases. Measurements were treated as missing when a reader did not submit data.

Among a total of 390 possible *reads* by 13 readers, 347 baseline *reads* and 329 follow-up *reads* were reported, of which 297 *reads* were complete baseline and follow-up pairs. Such *reads* were used to compute the ICC with all readers and agreement with $\text{read}_{\text{reference}}$ for the baseline, follow-up and percentage change in SUL_{peak} , respectively. The ICC for $\% \Delta \text{SUL}_{\text{peak}}$ was 0.87 (95% CI: [0.78, 0.92]), and agreement with $\text{read}_{\text{reference}}$ was 0.88 (Range: [0.61, 1.00]). The ICC and agreement with $\text{read}_{\text{reference}}$ of other metrics are in Table 1. The overall within-subject coefficient of variance (COV; overall SD / average of the case means) for $\% \Delta \text{SUL}_{\text{peak}}$ change was computed as 2.293. The Bland-Altman plot of the $\% \Delta \text{SUL}_{\text{peak}}$ is shown in Figure 2.

TABLE 1. ICC and agreement for single or up-to-5 SUL_{peak} selections

	ICC (95% CI)			Agreement with read _{reference} ICC (Min, Max)		
	Baseline	Follow-up	Percentage Change	Baseline	Follow-up	Percentage Change
1 SUL _{peak} (All reads)	0.90 (0.82, 0.94)	0.75 (0.64, 0.82)	0.87 (0.78, 0.92)	0.95 (0.72, 1.00)	0.78 (0.37, 1.00)	0.88 (0.61, 1.00)
1 SUL _{peak} (Reads with same tumor selected)	1.00 (0.49, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	0.95 (0.43, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
Sum of up to 5 SUL _{peak} (All reads)	0.93 (0.85, 0.96)	0.85 (0.67, 0.92)	0.77 (0.61, 0.85)	0.95 (0.84, 0.99)	0.89 (0.46, 0.99)	0.80 (0.39, 0.95)
Sum of up to 5 SUL _{peak} (Reads with same tumors selected)	0.98 (0.94, 0.98)	0.98 (0.94, 0.98)	0.96 (0.92, 0.98)	0.97 (0.95, 0.99)	0.98 (0.97, 1.00)	0.93 (0.87, 1.00)

ICC - Intra-class correlation coefficient

CI - Confidence interval

SUL_{peak} - Peak standardized uptake value corrected for lean body mass

Agreement with read_{reference} - ICC between reference reader and each of the other 12 readers

Reads with same target tumor

Among the 360 possible *reads* from the 12 readers, the readers selected a different lesion compared to the read_{reference} in 46 *reads* at baseline, 43 *reads* at follow-up, and 29 *reads* at both baseline and follow-up. The 241 *reads* agreeing on target selection with the read_{reference} were used to compute the ICCs with all readers and agreement with read_{reference}. The ICC for %ΔSUL_{peak} among all readers was 1.00 (95% CI: [1.00, 1.00]), and agreement with read_{reference} was 1.00 (Range: [1.00, 1.00]). The ICC and agreement with read_{reference} of other metrics are in Table 1. The overall within-subject COV for %ΔSUL_{peak} change was computed as 0.007. The Bland-Altman plot is shown in Figure 3.

Sum of up to 5 SUL_{peak}

In addition to the SUL_{peak} measurement of a single lesion, the sum of SUL_{peak} measurements of up to 5 of the selected lesions was used to compute the ICC and agreement with read_{reference} for all reads and reads with the same target lesion (Table 1). Even when the same lesions were selected, the ICCs and agreement with read_{reference} were not a perfect 1.00 due to (a) differences in the manual thresholds used

for lesion detection and (b) utilization of the 'erosion option' for breaking up [¹⁸F]FDG uptake volumes by the individual readers.

TABLE 2. Inter-rater reliability of readers on response assessment

	Krippendorff alpha coefficient of 13 readers on response	
	All reads (n=380*)	Reads with same target (n=270*)
Response vs. no response (CMR/PMR vs. SMD/PMD)	0.91	1.00
Clinical benefit vs. no benefit (CMR/PMR/SMD vs. PMD)	0.81	1.00
Response categories (ordinal scale) CMR vs. PMR vs. SMD vs. PMD	0.86	1.00

*Reads with missing response were excluded (10 for all reads, and 1 for reads with same target)

Inter-rater reliability of readers on responses

Among the 390 *reads* for all reads, 380 *reads* reported response categories. Among the 271 *reads* agreeing on target selection with the read_{reference}, 270 reported response categories. The Krippendorff alpha coefficient of 13 readers for binary response measure (response (CMR/PMR) versus no-response (SMD/PMD)) was 0.91 for all *reads*, and 1.00 for only the *reads* with the same target lesion selection. When assessing clinical benefit (SMD/PMR/CMR representing clinical benefit versus PMD representing no benefit), the Krippendorff alpha coefficient was 0.81 for all *reads* and 1.00 for only the *reads* with the same target selection. With the four response categories treated in an ordinal scale, the Krippendorff alpha coefficient was 0.86 for all *reads* and 1.00 for only the *reads* with the same target selection (Table 2).

Discussion

Variability in measurements across readers and sites is an often cited hurdle to broader utilization of quantitative [¹⁸F]FDG PET/CT for response assessment of cancer treatment (11). Test-retest studies have demonstrated high repeatability of [¹⁸F]FDG and other radiopharmaceutical PET parameters (12-15). The variance of SUVs could be greater in clinical practice compared to ideal study setting (16). In the clinical setting, measurement of SUV_{max} was demonstrated to have high agreement in our previous paper, while the statistically more robust SUL_{peak} showed suboptimal agreement (9). We wanted to know if using

uniform software could eliminate the variability associated with the computation differences for SUL_{peak} across multiple vendors and software.

The localization of the liver, SUL measurements from the liver, computation of a threshold for lesion detection, and identification of candidate lesions were all performed automatically on Auto-PERCIST™. Following detection of all sites with SUL_{peak} higher than the set threshold, various $[^{18}F]FDG$ uptake intensity or pattern measurements and textural features for each of the detected sites were also performed automatically. When the readers chose the same single target tumor, the measurements were identical, as could be expected. For up to 5 hottest lesions measurements, the agreement was near perfect. However, agreement was not a perfect 1.00 even when the readers chose the same tumors because the readers had the option to break down a single volume of $[^{18}F]FDG$ uptake to separate parts, or add up two or more $[^{18}F]FDG$ uptake sites to a single volume as they determined appropriate. Some readers chose to break down a lesion detected on Auto-PERCIST™ to avoid including physiologic $[^{18}F]FDG$ uptake, or to separate a conglomeration of multiple tumors lesions. And some readers intentionally chose a detection threshold lower than the default software setting to include lesions with relatively low $[^{18}F]FDG$ uptake for assessment on the follow-up PET images. The agreement was lower for follow-up images for the all-reads assessment. The readers disagreed more often on what was tumor and what was physiologic or inflammatory response on the follow-up images.

A previous paper that showed excellent correlation between two different vendor software for SUL_{peak} had the tumor sites predefined by the readers to exclude interpretive error (13). Determining which $[^{18}F]FDG$ uptake site is true tumor remained a challenge even for experienced readers. In the outlier case in Figure 2 showing an average difference greater than 100%, some readers considered an intense $[^{18}F]FDG$ uptake in the colon on the follow-up image to be new tumor lesion, while the $read_{reference}$ considered it physiologic in nature. Of the 360 non-reference baseline reads (including missing measurements) in this study, only 241 reads (67%) chose the same lesion and went on to make the same measurements as the $read_{reference}$ at both baseline and follow-up. Among the 30 cases, the target lesion (hottest tumor) on the post-therapy scan was different from the target lesion noted on the pre-therapy scan in 11 cases. For example in one case, the target lesion was in a mediastinal node on pre-therapy scan, and then a lung lesion became the hottest tumor in the post-therapy scan. In 3 cases, nodes in different stations were the target lesions at different time points. Among patients with multiple bone or lung metastases, different lesions in the same organ could be observed becoming the target tumors at different time points. As seen in inter-observer agreement studies of $[^{18}F]FDG$ PET/CT performed in patients with lymphoma after therapy, even experienced readers do not always agree on what is tumor $[^{18}F]FDG$ uptake and what is physiologic $[^{18}F]FDG$ uptake (17,18). Rather than relying solely on the reading experience of the local site, discussions and consensus meetings and better training methods are necessary to implement $[^{18}F]FDG$ PET/CT to its full potential. It almost certainly is the case that the availability of more relevant patient history would result in better accuracy and consistency in tumor detection.

While PERCIST 1.0 is quantitative, the category of CMR is dependent on the reader's judgement, and software quantification alone could not determine the response to be CMR. There were 6 cases considered to have reached CMR by the read_{reference}. The twelve other readers categorized the case correctly as CMR in 44 *reads* out of 72 (12 readers x 6 cases), PMR was designated in 21 *reads*, SMD in 5 *reads* and PMD in 1 *read*, with 1 missing *read*. Thus, in addition to selection of different target tumor from the read_{reference}, the reader's decision between CMR and PMR leaves room for variability in response categorization, even if quantitation produces identical results. Detailed definition or consensus on findings compatible with the CMR category, or addition of quantitative threshold to clarify the CMR category is necessary for use in trials and in the clinical setting. A lesion could be considered present and thus not CMR even with very low SUL_{peak} for example in the lungs, or a lesion could be considered resolved and thus CMR even with relatively high SUL_{peak} for example in tonsils. The threshold computed from liver measurements (liver SUL_{mean} + 2SD) was viewed by the readers as too high a cutoff for CMR in this study as could be inferred by how the readers manually lowered the threshold on the follow-up images.

Revealing a potential limitation in the software, and of the PERCIST criteria, there was a small tumor with clearly perceptible [¹⁸F]FDG uptake visually, which was not detectable by Auto-PERCIST™ due to the volume below the PERCIST definition of SUL_{peak} sphere of 1 cubic centimeter (Figure 4). More mundane limitation of applying PERCIST includes the need to measure the patient's height. That many of the referring physicians and radiologist are not familiar with the SUL_{peak} parameters is another limitation to overcome. When there are multiple lesions showing intense [¹⁸F]FDG uptake, the lesion with the worst response may not be the target lesion, and PERCIST needs to specify how to address such poorly behaving lesions for categorizing the overall response.

Auto-PERCIST™ has the ability to automatically detect potentially new lesions for co-registered studies based on the location of the classified lesions. Auto-PERCIST™ also computed additional PET parameters representing tumor features, such as metabolic tumor volume, geometric mean, exposure, kurtosis and skewness, which have been reported as prognostic markers and diagnostic tools (19-22). Discordance among readers was minimal for the additional PET parameters, and the cause for any variance arose when the reader manually changed the tumor boundary. Even with the addition of several PET parameters, the measurement took seconds to at the longest and a few minutes for cases with many lesions. In addition to reducing variability in measurement, the software reduced the measurement time radically. Auto-PERCIST™ may become adjunct reading software the way myocardial perfusion and metabolism studies utilize cardiac image analysis software. Auto-PERCIST is available to academic researchers who register their interest with the Johns Hopkins Technology Transfer office.

Conclusion

Harmonization of methods to single software Auto-PERCIST™ resulted in virtually identical extraction of quantitative data including the SUL_{peak} when the readers selected the same target tumor, and should

promote greater use of [¹⁸F]FDG PET/CT for response assessment in cancer treatment. Nonetheless, the findings show caution remains in order as lesion selection still results on qualitative assessments of whether a lesion is tumor or physiological uptake.

Abbreviations

[¹⁸F]FDG: 2-deoxy-2-[¹⁸F]fluoro-D-glucose; PET/CT: positron-emission tomography/computed tomography; SULpeak : peak standard uptake value corrected for lean body mass; %ΔSULpeak: percent change in the SULpeak; PERCIST 1.0: PET Response Criteria in Solid Tumors 1.0; SUVmax: maximum standardized uptake value; CI: confidence interval; SD: standard deviation; ICC: intraclass correlation coefficient; CMR: complete metabolic response; PMR: partial metabolic response; SMD: stable metabolic disease; PMD: progressive metabolic disease; COV: coefficient of variance

Declarations

Ethics approval and consent to participate

This retrospective study was approved by our Institutional Review Board. Informed consent was waived.

Consent for publication

Not applicable

Availability of data and material

The datasets used in this study are available from the corresponding author on reasonable request.

Competing interests

No relevant conflicts of interest were identified except two of the authors, JL and RW, are co-inventors on a patent underlying the Auto-PERCISTä software.

Funding

Supported in part by grants awarded by the Radiological Society of North America, the Quantitative Imaging Biomarkers Alliance, National Cancer Institute (5U01CA140204-04), National Institutes of Health (NCI CCSG P30CA006973 and U01CA140204), and National Research Foundation of Korea (NRF-2019R1G1A009158).

Authors' contributions

JHO, SJL, HW, JPL, HGS and RLW participated in the study design, collecting data and data analysis. JHO, SJL, HW, JPL prepared the manuscript and contributed to data analysis and interpretation. HGS and

RLW supervised the project and reviewed manuscript. All authors read and approved the final manuscript.

Acknowledgements

The QIN Readers were Seong Young Kwon from Chonnam National University; Hui-Kuo Shu from Emory University; Masatoyo Nakajo from Kagoshima University; Evelyn de Jong from Maastricht University; Tadashi Watabe from Osaka University; Jin Chul Paeng and Seo Young Kang from Seoul National University; Woo Hee Choi, Eun Ji Han and Hyelim Park from The Catholic University of Korea; Ella Jones and Youngho Seo from University of California, San Francisco; John Buatti from University of Iowa; James Mounts and Matthew Oborski from University Pittsburgh Medical Center; and Joyce Mhlanga from Washington University in St. Louis.

References

1. Barrington SF, Mikhaeel NG, Kostakoglu L, 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;32:3048-3058.
2. Gallamini A, Barrington SF, Biggi A, et al. The predictive role of interim positron emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the Deauville five-point scale. *Haematologica*. 2014;99:1107-1113.
3. Wahl RL, Zasadny K, Helvie M, Hutchins GD, Weber B, Cody R. Metabolic monitoring of breast cancer chemohormonotherapy using positron emission tomography: initial evaluation. *J Clin Oncol*. 1993;11:2101-2111.
4. Connolly RM, Leal JP, Solnes L, et al. TBCRC026: Phase II Trial Correlating Standardized Uptake Value With Pathologic Complete Response to Pertuzumab and Trastuzumab in Breast Cancer. *J Clin Oncol*. 2019;37:714-722.
5. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50 Suppl 1:122S-150S.
6. O JH, Lodge MA, Wahl RL. Practical PERCIST: A Simplified Guide to PET Response Criteria in Solid Tumors 1.0. *Radiology*. 2016;280:576-584.
7. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med*. 2012;53:1041-1047.
8. Quak E, Le Roux PY, Lasnon C, et al. Does PET SUV Harmonization Affect PERCIST Response Classification? *J Nucl Med*. 2016;57:1699-1706.

9. O JH, Jacene H, Luber B, et al. Quantitation of Cancer Treatment Response by (18)F-FDG PET/CT: Multicenter Assessment of Measurement Variability. *J Nucl Med.* 2017;58:1429-1434.
10. Krippendorff K. Measuring the reliability of qualitative text analysis data. *Quality & Quantity.* 2004;38:787-800.
11. Ding Q, Cheng X, Yang L, et al. PET/CT evaluation of response to chemotherapy in non-small cell lung cancer: PET response criteria in solid tumors (PERCIST) versus response evaluation criteria in solid tumors (RECIST). *J Thorac Dis.* 2014;6:677-683.
12. Lodge MA. Repeatability of SUV in Oncologic (18)F-FDG PET. *J Nucl Med.* 2017;58:523-532.
13. Mhlanga JC, Chirindel A, Lodge MA, Wahl RL, Subramaniam RM. Quantitative PET/CT in clinical practice: assessing the agreement of PET tumor indices using different clinical reading platforms. *Nucl Med Commun.* 2018;39:154-160.
14. Fox JJ, Autran-Blanc E, Morris MJ, et al. Practical approach for comparative analysis of multilesion molecular imaging using a semiautomated program for PET/CT. *J Nucl Med.* 2011;52:1727-1732.
15. Fraum TJ, Fowler KJ, Crandall JP, et al. Measurement Repeatability of (18)F-FDG PET/CT Versus (18)F-FDG PET/MRI in Solid Tumors of the Pelvis. *J Nucl Med.* 2019;60:1080-1086.
16. Kumar V, Nath K, Berman CG, et al. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. *Clin Nucl Med.* 2013;38:175-182.
17. Burggraaff CN, Cornelisse AC, Hoekstra OS, et al. Interobserver Agreement of Interim and End-of-Treatment (18)F-FDG PET/CT in Diffuse Large B-Cell Lymphoma: Impact on Clinical Practice and Trials. *J Nucl Med.* 2018;59:1831-1836.
18. Han EJ, O JH, Yoon H, et al. FDG PET/CT response in diffuse large B-cell lymphoma: Reader variability and association with clinical outcome. *Medicine (Baltimore).* 2016;95:e4983.
19. Pinker K, Riedl C, Weber WA. Evaluating tumor response with FDG PET: updates on PERCIST, comparison with EORTC criteria and clues to future developments. *Eur J Nucl Med Mol Imaging.* 2017;44:55-66.
20. Tixier F, Hatt M, Valla C, et al. Visual versus quantitative assessment of intratumor 18F-FDG PET uptake heterogeneity: prognostic value in non-small cell lung cancer. *J Nucl Med.* 2014;55:1235-1241.
21. Huang W, Fan M, Liu B, et al. Value of metabolic tumor volume on repeated 18F-FDG PET/CT for early prediction of survival in locally advanced non-small cell lung cancer treated with concurrent chemoradiotherapy. *J Nucl Med.* 2014;55:1584-1590.

Figures

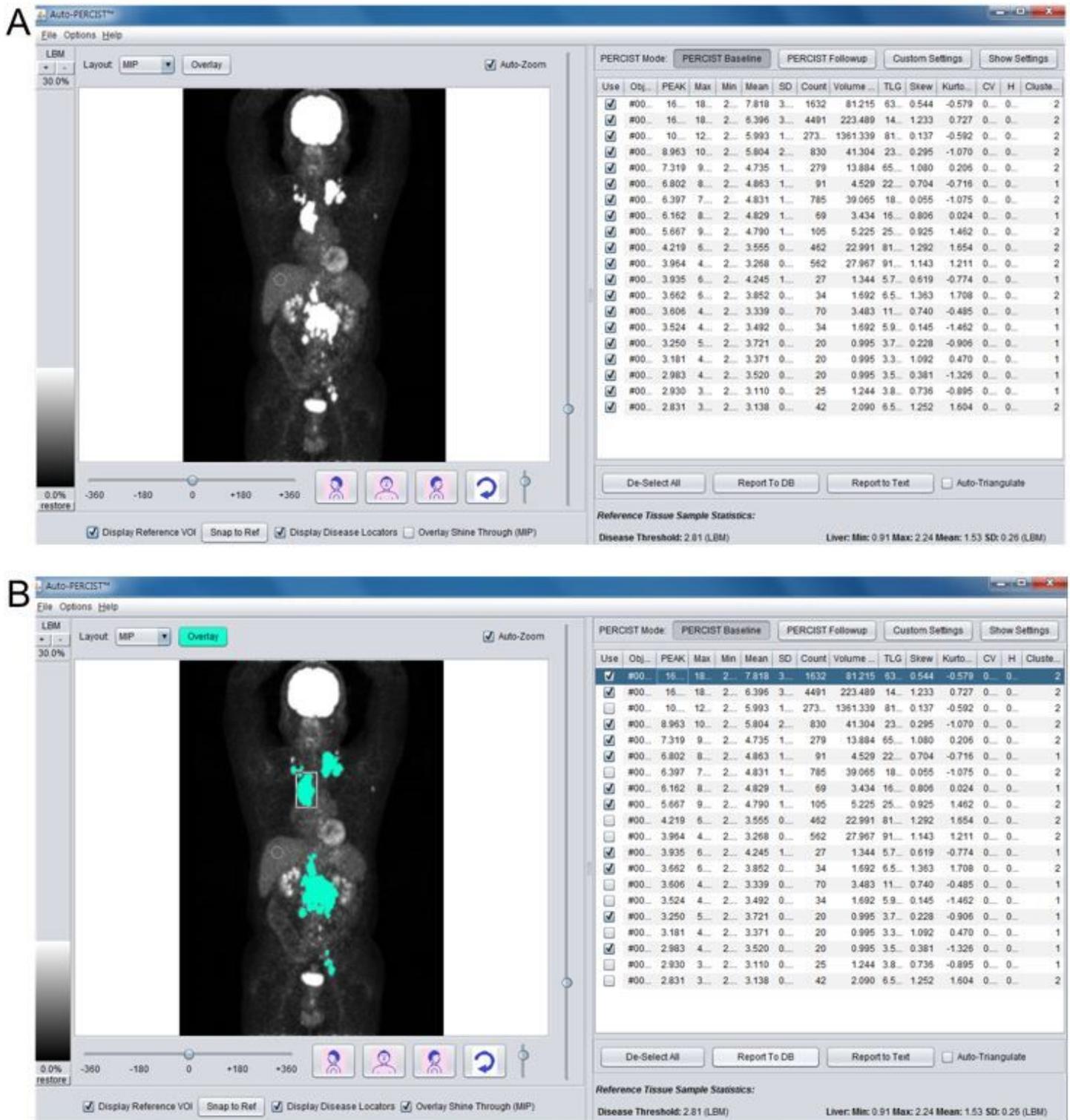


Figure 1

Screen captures from Auto-PERCISTTM. (A) The software detected all sites with SULpeak higher than the computed or manually set threshold. (B) The reader then selected the true tumor lesions (shaded in green), excluding physiologic [18F]FDG activity.

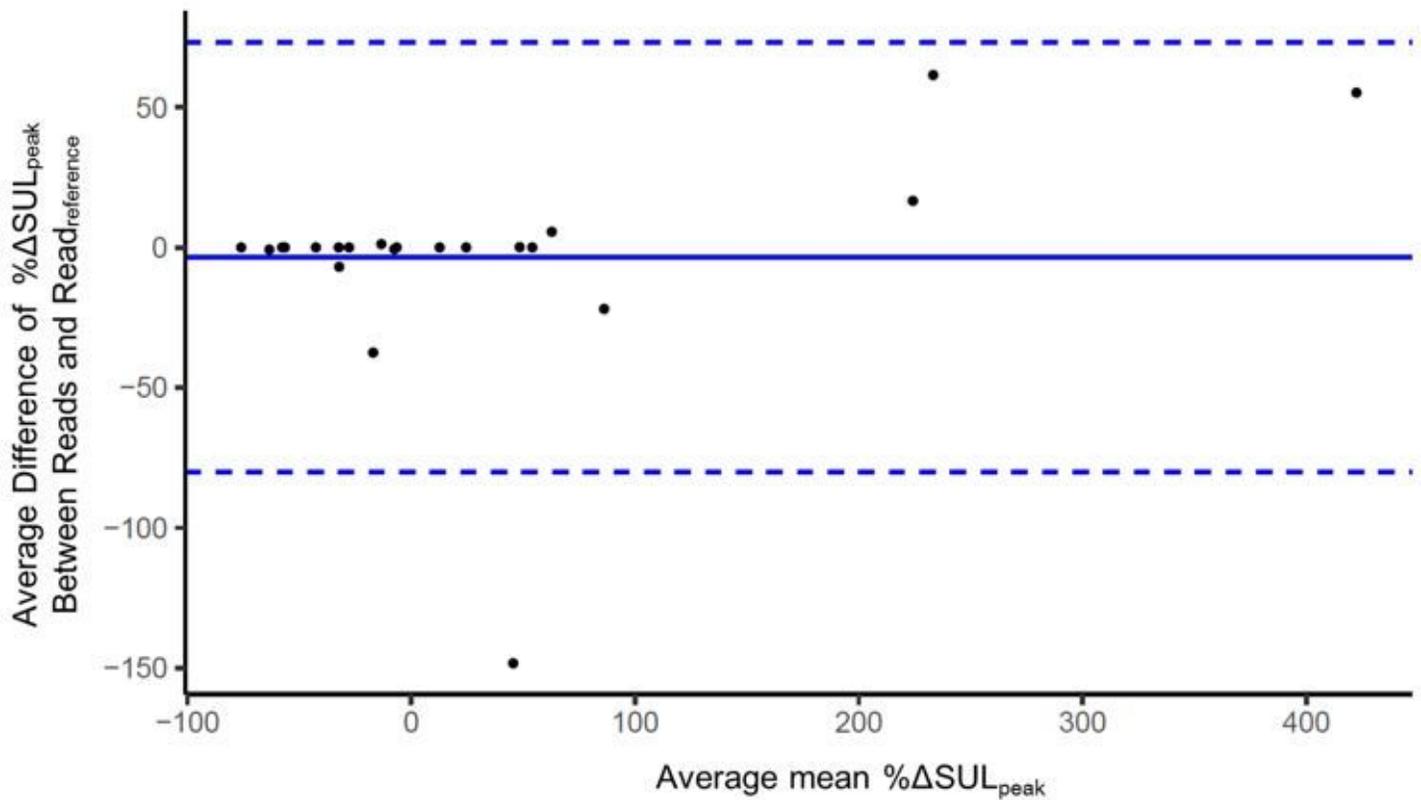


Figure 2

Bland-Altman plot of the percentage change of tumor [18F]FDG uptake from baseline to follow-up. The plot is for the percentage changes of SULpeak for all reads. Each dot represents a case (30 cases in total). The x-axis represents the average mean percentage change measurement by all readers. The y-axis represents the average difference between the 12 readers and the reference reader (readreference). The solid line represents the average bias, and the dashed lines represent the corresponding bias \pm 2 standard deviations (SD).

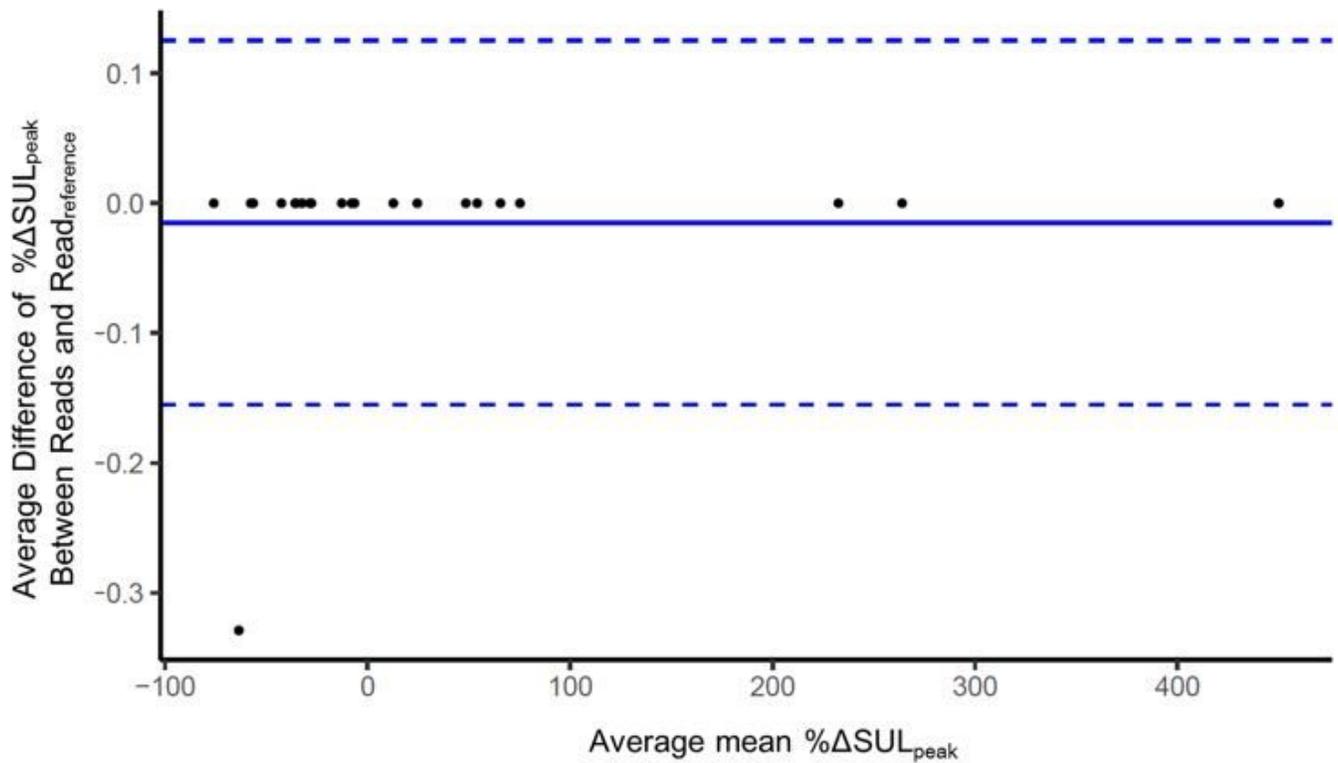


Figure 3

Bland-Altman plot of the percentage change of tumor [18F]FDG uptake from baseline to follow-up. The plot is for the percentage changes of SUL_{peak} (%ΔSUL_{peak}) for only the reads with same lesion selected as the readreference. Each dot represents a case (30 cases in total). The x-axis represents the average mean %ΔSUL_{peak} measurement by all readers. The y-axis represents the average difference between the 12 readers and the reference reader (readreference) and the y-axis unit is one tenth of one percent. The solid line represents the average bias, and the dashed lines represent the corresponding bias +/- 2 standard deviations (SD).

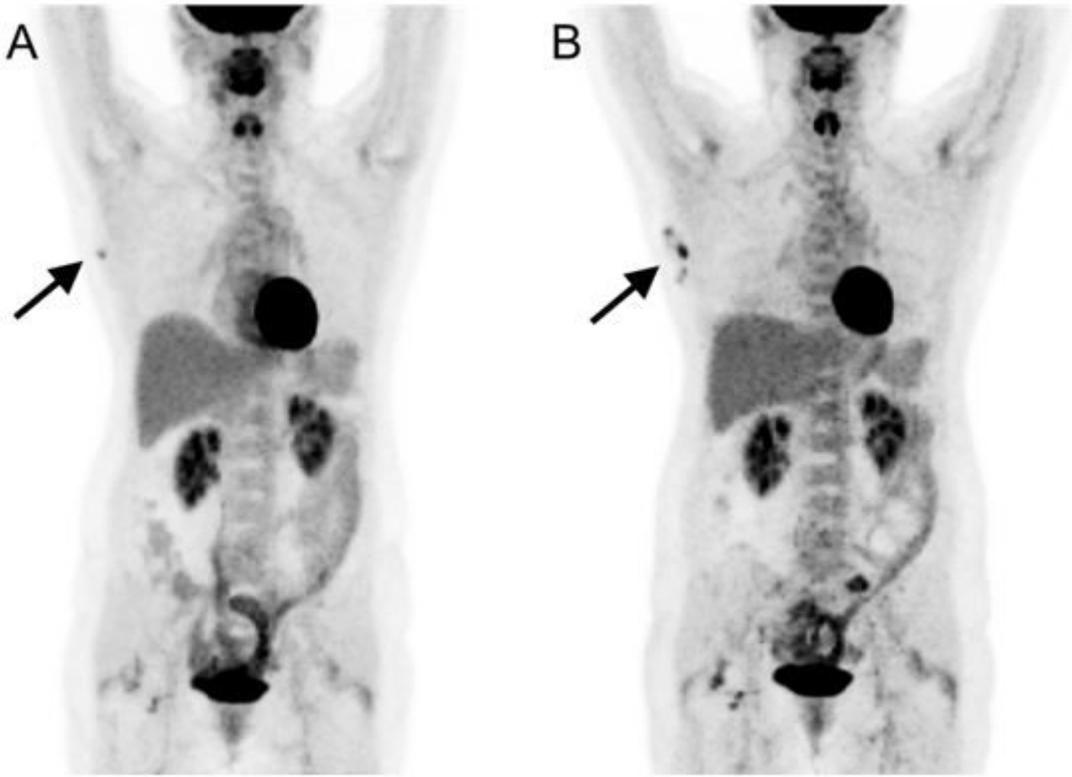


Figure 4

(A) PET maximum intensity projection (MIP) image of patient with right axillary node metastasis at baseline with SULpeak of 1.62 and tumor volume of less than 1.00 cc. Though visually perceptible, Auto-PERCISTTM failed to detect the lesion due to small size. (B) On the follow-up MIP image, the number of metastatic nodes and the [18F]FDG uptake intensity are increased to SULpeak of 2.84, allowing detection by Auto-PERCISTTM.