Journal of Nuclear Medicine, published on December 8, 2011 as doi:10.2967/jnumed.111.094763

Mutatis Mutandis: Harmonize the Standard!

Quantitative imaging is taking a more prominent role in the clinical arena as well as in research (*1,2*). There is increased interest both from the radiology and nuclear medicine fields and from pharmaceutical companies to use quantitative reads from clinical images, such as for diagnosis/stratification or treatment response assessment (*3–5*). In fact, PET/CT examinations are already used as imaging biomarkers. Staging based on PET/CT examination may be used to assess patient eligibility for trial participation. Quantitative reads—such as for tumor size or ^{18}F-FDG or ^{18}F-FLT uptake—may be used as a tool for assessing prognostic factors and treatment response as well (*6*).

See page ■■■

Any quantitative read from any type of image data is affected by the quality of the acquired images and the methods used to extract quantitative measures from them. For example, 1-dimensional, 2-dimensional, or volumetric tumor size measurements from diagnostic CT images will differ numerically by definition. Consequently, when tumor size is used as a prognostic factor or when changes in tumor size are used to assess treatment response, the thresholds applied (e.g., response classification) need to be adapted or recalibrated for the way quantitative reads are generated.

Whole-body ^{18}F-FDG and ^{18}F-FLT PET/CT studies are usually quantified using standardized uptake values (SUVs). SUVs may differ in the normalization factor being used (e.g., body weight or lean body mass) or in whether corrections for blood glucose are applied (*7*). In all cases, SUVs are derived from a PET study by first placing a 2-dimensional region of interest (ROI) or 3-dimensional volume of interest (VOI) in or around the tumor. The most commonly derived SUVs are maximum, peak, and mean SUV (SUV_{max}, SUV_{peak}, and SUV_{mean}, respectively).

SUV_{mean} is the average value within a manually or automatically defined VOI. At present, however, there is not a widely available and accepted automated VOI method allowing SUVs to be obtained in a consistent manner across sites (*8*). As a result, SUV_{max} has become the standard in many quantitative PET studies. SUV_{max} is defined as the SUV derived from a single voxel showing the highest uptake across the tumor. SUV_{max} exhibits several benefits. For example, it can be fairly easily derived from a VOI encompassing the tumor and may therefore be nearly free of observer variability. Second, it may represent the metabolically most active part of the tumor, which could be considered the best index for use as a prognostic or predictive factor. Finally, it may suffer least from partial-volume effects or patient motion, although this depends strongly on the tumor shape and tracer uptake heterogeneity. Mainly the ease of obtaining SUV_{max}, as well as its being representative of the most active part of the tumor, has made it one of the most commonly used parameters. Yet, drawbacks of using SUVmax arise from the single-voxel VOI definition. This can result in noise-induced upward bias (*9*) and voxel sizes and thus VOI sizes are different between various PET/CT systems having different default image reconstruction methods and settings. These drawbacks of SUV_{max} have resulted in the use of SUV_{peak}, as was also suggested in the PET Response Criteria in Solid Tumors (PERCIST) (*10*). SUV_{peak} is defined as a circular ROI or spheric VOI having a fixed, predefined diameter. These ROIs and VOIs can be centered on SUV_{max} or localized such that SUV_{peak} is highest across all locations within the tumor. SUV_{peak} may be less sensitive to noise because SUV_{peak} is based on the average value within an ROI or VOI. Moreover, SUVs are derived from an ROI or VOI having the same size regardless of the PET/CT system and reconstruction settings being used. Finally, recently it was shown that SUV_{peak} may be less sensitive to variability in image characteristics (resolution) than is SUV_{max}, likely because of the inherent smoothing that occurs when a fixed-size VOI is used (*11*). A drawback is the nonstandardized use of various peak ROI/VOI definitions or the lack of proper peak ROI/VOI implementations in image-viewing and analysis software.

The paper of Vanderhoek et al. (*12*) in this issue of *The Journal of Nuclear Medicine* explores the impact of SUV_{max} and SUV_{peak} definition on the quantification of ^{18}F-FLT uptake and its change during treatment. Vanderhoek et al. showed that ^{18}F-FLT SUVs derived with various ROI or VOI definitions (SUV_{max} and several SUV_{peaks}) can be substantially different from one another. This observation by itself is not surprising, as changing the size of the peak VOI will change the average SUV by definition. However, the authors showed that by varying the peak VOI diameter across the range seen in the liter-

Received Nov. 7, 2011; revision accepted Nov. 10, 2011.

For correspondence contact: Ronald Boellaard, Department of Nuclear Medicine and PET Research, VU University Medical Centre, De Boelelaan 1117, P.O. Box 7057, Amsterdam 1007 MB, The Netherlands.
E-mail: r.boellaard@vumc.nl

Published online ■■■■.

DOI: 10.2967/jnumed.111.094763

ature, SUV varied from 49% above to 46% below the mean SUV averaged across all SUV_{peaks}. More importantly, it was found that the definition of SUV_{peak} also affected observed treatment responses derived using relative (percentage) SUV changes, even when the same ROI or VOI method was applied to all PET scans of the same subject. That observation is of particular interest because it is generally assumed that the procedures followed to acquire PET images and the methods applied to obtain SUV have a minor effect on observed relative SUV changes (*13,14*). That assumption turned out to be correct, on average; that is, the average tumor response seen across all lesions and subjects was not significantly affected by changing the SUV_{peak} definition.

The observations in the paper of Vanderhoek et al. may have several implications. It is not always realized that absolute, or baseline, SUVs are being used in treatment response studies. For example, in trials, SUVs may be used for assessing subject eligibility or for stratification. Recently, PERCIST (*10*) suggested using the 5 hottest lesions per scan as the set of target lesions for response assessment. SUV in these target lesions should also exceed a minimal uptake threshold defined, for example, as 1.5 × liver SUV + 2 × liver SUV noise (as determined by SUV SD within a VOI). This minimal uptake threshold corresponds in practice to an SUV of about 3.5–4.0. Moreover, apart from a percentage change in SUV, a minimal change of 0.8 SUV (normalized using lean body mass) is considered. These absolute thresholds or minimal uptake changes are sensible and valid only for specific definitions of SUVs. Although PERCIST applies to ^{18}F-FDG PET, similar criteria—minimal thresholds for target lesion selection and a mixture of relative and absolute changes for response classification—are likely needed for ^{18}F-FLT PET as well. The huge variability from −46% to +49% in SUV_{peak} by changing its definition, as observed by Vanderhoek et al., illustrates that we can no longer afford to just derive SUV from PET scans in a nonharmonized manner. This will obscure the definition of any clinically useful threshold or response classification from literature or data that multicenter studies may provide.

When only relative or percentage SUV changes are considered as an index for treatment response or drug efficacy between populations of subjects, the specific definition of SUV_{peak} seems less relevant. Vanderhoek et al. showed that population differences in responses or response classifications using different SUV_{peak} definitions were not statistically different. This observation is consistent with observations of Yap et al. (*15*). In fact, that recent study demonstrated an extremely high correlation and correspondence between percentage SUV changes derived using SUV_{max} and SUV_{peak}. Although the study of Yap et al. concerned only ^{18}F-FDG, their observations seem to be valid for ^{18}F-FLT as well because a high correlation between various SUV responses was also seen in the study of Vanderhoek et al. (*12*). Yet, Vanderhoek et al. showed that in individual cases, ^{18}F-FLT SUV responses (relative changes) could differ substantially depending on the VOI definition being used. This finding may have implications in, for example, a crossover trial design in which subjects may cross over to the best-of-standard-care arm, such as might happen when a patient treated within an experimental arm experiences metabolic progression. Second, thresholds and criteria that have and will be derived from data collected during trials are and will be used clinically. In that case, translation of trial-based information and criteria into the clinic would require the use of the same SUV and VOI definitions. Moreover, we need to derive the SUV definition that provides the most robust and best predictive information that can be used in practice for our patients, not only at a group level but also in individual cases. Studies that directly compare use of these standardized SUV methods in combination with proposed criteria for response assessment (*10,16*) are needed to address these issues.

Evolution comes from diversity, and only those species that are best fitted to their surroundings will survive. PET/CT vendors and institutions operate in a competitive market and consequently need to continuously enhance and change their procedures and the performance of their systems. Striving to enhance image quality, image accuracy, and precision of PET quantification is a shared goal, not only for PET/CT vendors to survive but also for academia and practitioners to obtain the best results for their studies or best care for their patients. However, how do we assess accuracy and precision, and how do we collect evidence to demonstrate that improvements in accuracy and precision result in clinical benefit (evidence-based medicine)? One of the requirements to solve this question is the use of standards. As an example, we can take the evolution of the standard meter. Use of thumbs, feet, knots, etc., made it impossible to assess the size of any object, although at that time also there likely was a lot of debate about which method was best. The nonstandardized way of measuring length hampered trade and the gain of scientific knowledge, as data collected in one institution or country could not be used elsewhere. This had already been realized more than 200 years ago, in 1795, when the first definition of the standard meter was proposed. New, more accurate and precise standard meters were developed as technology moved forward, with the latest standard defined in 1983 (*17*). From the standard meter example, we learn 2 things. First, it is absolutely essential to have a standard for use as a reference in order to further develop the field; that is, only by performing measurements in a consistent and calibrated manner across the field can we move forward. It allows comparison of other measures and methods against the common standard and direct linkage of results obtained in different studies. Although we still use various units for length (meters, inches, miles), they can be directly converted into one another, as they are all calibrated to the same standard. Likewise, we can use different PET/CT systems,

provided harmonization is in place. Second, as technology improves as a result of competitive markets, standards will be enhanced as well, just as the definition of the meter evolved as well. In this way, we can benefit from improvements in technology while at the same time having standards to collect clinical evidence from and for our studies. Change or possible future improvements in technology therefore cannot be used as an excuse not to harmonize or standardize our PET/CT procedures and methodology but are required to demonstrate the clinical benefits of these improvements. Mutatis mutandis, the evolution of technology should not be held back, but standards will become more accurate and precise over time as well. The challenge here is to evolve standards in a harmonized manner along with technologic developments.

The need for harmonized PET/CT study procedures and scanner performance not only is realized by individual investigators but also is acknowledged by various scientific societies (*13*). Clinical validation of quantitative ^{18}F-FDG PET/CT by use of standards is at this moment even more important than achieving the best possible image quality in individual cases (patients, scanners, institutions). Harmonization and standards are required for making PET/CT a validated quantitative imaging biomarker tool. The paper of Vanderhoek et al. (*12*) demonstrates that harmonization of our data analysis procedures in order to harmonize results obtained with our PET/CT studies is needed as well.

ACKNOWLEDGMENT

No potential conflict of interest relevant to this article was reported.

Ronald Boellaard
Department of Nuclear Medicine and PET Research
VU University Medical Centre
Amsterdam, The Netherlands

REFERENCES

1. Hunter AJ. The Innovative Medicines Initiative: a pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov Today.* 2008;13:371–373.
2. Woosley RL, Myers RT, Goodsaid F. The Critical Path Institute's approach to precompetitive sharing and advancing regulatory science. *Clin Pharmacol Ther.* 2010;87:530–533.
3. Fletcher JW, Djulbegovic B, Soares HP, et al. Recommendations on the use of ^{18}F-FDG PET in oncology. *J Nucl Med.* 2008;49:480–508.
4. de Geus-Oei LF, van der Heijden HF, Corstens FH, Oyen WJ. Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer: a systematic review. *Cancer.* 2007;110:1654–1664.
5. Hoekstra CJ, Stroobants SG, Hoekstra OS, et al. The value of [^{18}F]fluoro-2-deoxy-D-glucose positron emission tomography in the selection of patients with stage IIIA-N2 non-small cell lung cancer for combined modality treatment. *Lung Cancer.* 2003;39:151–157.
6. Weber WA. Use of PET for monitoring cancer therapy and for predicting outcome. *J Nucl Med.* 2005;46:983–995.
7. Stahl A, Ott K, Schwaiger M, Weber WA. Comparison of different SUV-based methods for monitoring cytotoxic therapy with FDG PET. *Eur J Nucl Med Mol Imaging.* 2004;31:1471–1478.
8. Cheebsumon P, Yaqub M, van Velden FH, Hoekstra OS, Lammertsma AA, Boellaard R. Impact of [^{18}F]FDG PET imaging parameters on automatic tumour delineation: need for improved tumour delineation methodology. *Eur J Nucl Med Mol Imaging.* 2011.
9. Boellaard 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;45:1519–1527.
10. 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.
11. Makris NE, Huisman MC, Lammertsma AA, Boellaard R. Comparison of scanner validation programs used in oncology FDG PET/CT trials: impact of image reconstruction settings and VOI definition method [abstract]. *Eur J Nucl Med Mol Imaging.* 2011;38(suppl 2):S172.
12. Vanderhoek M, Perlman SB, Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response [abstract]. *J Nucl Med.* 2012;53:■■–■■.
13. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med.* 2009; 50(suppl 1):11S–20S.
14. Shankar LK, Hoffman JM, Bacharach S, 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;47:1059–1066.
15. Yap J, Locascio T, Tanaka Y, Syrkin L, Van Den Abbeele A. Impact of variations in SUV methods for assessing cancer response using FDG-PET [abstract]. *J Nucl Med.* 2011;52(suppl 1):392P.
16. Young H, Baum R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [^{18}F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer.* 1999;35:1773–1782.
17. Metre. Wikipedia: The Free Encyclopedia. Available at: http://en.wikipedia.org/wiki/Meter. Accessed November 22, 2011.