In the beginning, the fraction of instruments failing on a first attempt was quite similar to the data reported, but there was an improvement for subsequent qualification processes associated with participation in further multicenter trials, an effect attributable to training and increasing experience.

REFERENCES

1. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel BA. Qualification of PET scanners for use in multicenter cancer clinical trials: the American College of Radiology Imaging Network experience. *J Nucl Med.* 2009;50:1187–1193.
2. Geworski L, Knoop BO, de Wit M, Ivancevic V, Bares R, Munz DL. Multicenter comparison of calibration and cross calibration of PET scanners. *J Nucl Med.* 2002;43:635–639.
3. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med.* 2009;50(suppl):11S–20S.

Lilli Geworski*
Bernd Knoop
**Hannover Medical School*
Carl Neuberg Strasse 1
Hannover 30625, Germany
E-mail: geworski.lilli@mh-hannover.de

DOI: 10.2967/jnumed.109.069989

SUVs: Always a Good Choice?

TO THE EDITOR: The excellent presentation by Scheuermann et al. (*1*) on how the PET community is addressing sometimes-impaired abilities to compare semiquantitative results among institutions was well worth reading. This article followed an earlier publication of many valuable recommendations (*2*) for multiinstitutional therapeutic response trials, including a preference for standardized uptake values (SUVs). But with candor, the current study reports specific problems with SUVs. The scanners of one manufacturer give 20% and 4% lower SUVs than the scanners of other manufacturers for a physiologic phantom and a physical phantom, respectively. The former, somewhat of a surrogate phantom, was a rather precise population average of normal-liver ^{18}F-FDG SUVs. More important, it is the absence of results from a quantitative measurement model of all factors controlling the magnitude of this phenomenon that can question confidence in SUVs.

Also disturbing are results from an earlier survey (*3*) of normal-liver ^{18}F-FDG SUV population averages: a rather wide range of values, from 1.5 to 3.6. This shows an error range far exceeding the somewhat low SE for this physiologic phantom: SE = (liver average SUV of 2.5) × (0.2)/(a significant number averaged)$^{1/2}$, where the same-scanner coefficient of variation for a normal-liver population is approximately 0.2.

Additionally, in the current study a significant number of participating institutions had difficulty in obtaining an SUV of 1.0 within a known physical phantom. This variation in accuracy occurred despite a necessarily biased sample of volunteering researchers making special efforts to qualify their PET quantitation methodology for clinical trials. It appears that the overall institution-dependent error magnitude would be a composite of these spurious errors (infrequent errors and perhaps of greater magnitude), systematic methodology errors, and instrument errors (probably of lesser magnitude).

It is good to see impressive results from the rigor of physical phantom use being supplemented with physiologic phantom data. I would like to call attention to a way to improve upon usage of liver averaging. A more robust reference having better statistics might be provided by the use of fully corrected population-averaged SUVs from a combination of several organs individually having low coefficients of variation—similar to an approach in a mouse study (*4*). An atlas compilation of SUV data of many organs shows several candidates whose coefficients of variation are about as low as that of the liver (*5*).

A step beyond impartial reporting of SUV measuring accuracy could be asking whether findings now suggest revisiting setting-specific decisions to choose the SUV over other markers—whether in the clinical trial setting or in the more commonly encountered single-patient clinical diagnostic setting. Are all systematic and random measurement uncertainties being adequately considered as judgments are made? Additionally, and more rigorously, should any preferred choice among competing markers be justified by studies (e.g., cost–benefit comparisons) for a particular setting? Further, are methodology subclasses of the SUV and other markers also explored as options?

There are various candidates for other markers that compete with the SUV. In a methodology hierarchy of increasing complexity and diagnostically enhancing information, some classes (with subclass examples) to consider include the following: single-scan tissue ratios (e.g., the ratio of a region to an organ average (*4*), to liver, to cerebellum, or to a contralateral side); single-scan SUVs (with or without varieties of corrections and transformations); dual-time scans (widely spaced in time or an extension of a whole-body scan, with or without patient-specific plasma tracer information for a 2-time-point Patlak plot); and dynamic scans (with a wide range of plasma tracer information options for Patlak or compartmental model analysis). The possible use of a transformed SUV mentioned in this list, such as ln(SUV), stems from statistical distribution considerations when the need for correctly quantified statistical significance plays a noticeable role in diagnostic decision making (*6*).

The message here is a suggestion to pause and reflect on whether to passively accept SUVs as presented from software, to aggressively improve on the SUV methodology used, or to expend the effort to evaluate and pursue other options. This last alternative is somewhat supported by a recommendation from a study that evaluated statistical considerations for SUV use in early clinical trials having few patients. This is to consider the advantages of a better-performing, accurate PET procedure that permits fewer patients than would the SUV for a given statistical performance, even if the methodology is complex (*7*).

Finally, returning to the endeavors of institutions to qualify their PET protocols, valuable and rarely tabulated information obtained from a large population of scanned patients has been acquired for this paper (*1*) regarding human errors and other methodology problems. If these problems involve larger-magnitude errors, even if less frequent and mostly controllable, they nevertheless can increase the probability that SUV measurements will be less accurate overall. If ideally limited by only modest random errors, the SUV might have acceptable potential in many settings. Additional specific recommendations from this work would be a beneficial resource for future PET procedural guidelines.

REFERENCES

1. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel BA. Qualification of PET scanners for use in multicenter cancer clinical trials: the American College of Radiology Imaging Network experience. *J Nucl Med.* 2009;50:1187–1193.

2. 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.
3. Park K, Ashlock R, Chang J, et al. High variation in standardized uptake values among PET systems from different manufacturers [abstract]. *J Nucl Med.* 2007;48(suppl):185P.
4. Yu C, Zhang X, Huang S. Establishment of a database of FDG SUV for major organ tissues in normal mice [abstract]. *J Nucl Med.* 2008;49(suppl):151P.
5. Wang Y, Chiu E, Rosenberg J, Gambhir SS. Standardized uptake value atlas: characterization of physiological 2-deoxy-2-[^{18}F]fluoro-d-glucose uptake in normal tissues. *Mol Imaging Biol.* 2007;9:83–90.
6. Thie JA, Hubner KF, Smith GT. The diagnostic utility of the lognormal behavior of PET standardized uptake values in tumors. *J Nucl Med.* 2000;41:1664–1672.
7. Doot R, Kurland B, Kinahan P, Mankoff D. Considerations for using PET as a response measure in multicenter clinical trials [abstract]. *J Nucl Med.* 2009; 50(suppl 2):140P.

Joseph A. Thie
University of Tennessee
12334 Bluff Shore Dr.
Knoxville, TN 37922
E-mail: jathie@utk.edu

DOI: 10.2967/jnumed.109.064022