GBEF was less than 35%, leaving only 13 subjects. The investigators did not claim to have established reference values in that publication.

One might expect the letter writers' method to have results similar to our 15-min infusion. Table 1 of our article provides the mean GBEF and SD for a 15-min infusion, imaged for and quantified at 30, 45, and 60 min. Our data strongly suggest that the authors' conclusion regarding reference values is not correct. From our data, the lower limits of normal for their method would more likely be less than 25% (fifth percentile). In regard to the statistical criticism, we clearly noted that the data were not Gaussian in distribution and that we therefore used percentiles rather than mean and SD to establish reference values.

The writers misinterpret the relationship between abdominal pain and sincalide infusion. No subject we studied had abdominal pain with a 30-min or 60-min infusion, although 2 subjects had symptoms with a 15-min infusion. Prior data have shown that the incidence of pain with a 3-min sincalide infusion approaches 50% (*2*). Thus, the incidence of abdominal pain in healthy subjects is seen only with short infusion times. In healthy controls and patients with gallbladder disease, Yap et al. found that none had abdominal pain with a 45-min infusion, including those with low GBEFs (*4*). This is another important reason to use the slow infusion method—the patients will appreciate it.

Referring clinicians have had 2 concerns about the use of sincalide cholescintigraphy to diagnose chronic acalculous gallbladder disease: the first is the lack of sufficient evidence-based data proving the diagnostic utility of the GBEF, and the second is the lack of standardization of the infusion methodology (*9,10*). We strongly recommend general acceptance of the 60-min infusion method as the standard and that this methodology and these reference values become the standard. We hope that a prospective multicenter investigation will be initiated using this methodology with the expectation that it will confirm the utility of sincalide cholescintigraphy for selecting candidates for cholecystectomy.

REFERENCES

1. Ziessman HA, Tulchinsky M, Lavely WC, et al. Sincalide-stimulated cholescintigraphy: a multicenter investigation to determine optimal infusion methodology and gallbladder ejection fraction normal values. *J Nucl Med.* 2010;51:277–281.
2. Ziessman HA, Fahey FH, Hixson DJ. Calculation of a gallbladder ejection fraction: advantage of continuous sincalide infusion over the 3-min infusion method. *J Nucl Med.* 1992;33:537–541.
3. Ziessman HA, Muenz LR, Agarwal AK, ZaZa A. Normal values of sincalide cholescintigraphy: comparison of two methods. *Radiology.* 2001;221:404–410.
4. Yap L, Wychereley AG, Morphett AD, et al. Acalculous biliary pain: cholecystectomy alleviates symptoms in patients with abnormal cholescintigraphy. *Gastroenterology.* 1991;101:786–793.
5. Krishnamurthy GT, Krishnamurthy S, Brown PH. Constancy and variability of gallbladder ejection fraction: impact on diagnosis and therapy. *J Nucl Med.* 2004;45:1872–1877.
6. Krishnamurthy GT, Brown PH. Comparison of fatty meal and intravenous cholecystokinin infusion for gallbladder ejection fraction. *J Nucl Med.* 2002;43:1603–1610.
7. Krishnamurthy GT, Krishnamurthy S. Diagnostic reliability of gallbladder ejection fraction. *Indian J Nucl Med.* 2002;17:13–17.
8. Krishnamurthy S, Ceruli-Switzer J, Chapman N, Krishnamurthy GT. Comparison of gallbladder ejection fraction obtained with regular CCK-8 and pharmacy compounded CCK-8. *J Nucl Med.* 2003;44:499–504.
9. DiBaise JK, Oleynikov D. Does gallbladder ejection fraction predict outcome after cholecystectomy for suspected chronic acalculous gallbladder dysfunction? A systematic review. *Am J Gastroenterol.* 2003;98:2605–2611.
10. Delgado-Aros S, Cremonini F, Bredenoord AJ, Camilleri M. Systematic review and meta-analysis: does gall-bladder ejection fraction on cholecystokinin cholescintigraphy predict outcome after cholecystectomy in suspected functional biliary pain? *Aliment Pharmacol Ther.* 2003;18:167–174.

Harvey A. Ziessman*
Mark Tulchinsky
Alan H. Maurer
**Johns Hopkins Medical Institute*
601 N. Caroline St., Suite 3231
Baltimore, MD 21278
E-mail: hziessm1@jhmi.edu

DOI: 10.2967/jnumed.110.076646

Validating PET Scanner Calibration for Multicenter Trials

TO THE EDITOR: With interest we read the recent publication of Scheuermann et al. (*1*), who reported on the experience of the American College of Radiology Imaging Network in qualifying PET scanners to participate in multicenter trials. The network does so by analyzing submitted PET scans of uniform cylinders (either solid ^{68}Ge or fillable with ^{18}F) to verify the accuracy of scanner calibration (in terms of standardized uptake values) and by qualitatively reviewing typical patient images. Because many of the sites tested have been unable to produce acceptable results on the first attempt, the authors concluded that a verification of the basic scanner calibration is extremely important before sites can be allowed to participate in multicenter trials.

From our experience (*2*), we fully support this final conclusion. In particular, we agree that testing with fillable phantoms provides an independent check of system calibration and is a useful metric in characterizing the operator's experience in measuring and recording the injected dose accurately. The problems encountered are likely to occur in clinical acquisitions, too. The authors claim that using an identical phantom for calibration or normalization and for standardized uptake value testing, that is, a ^{68}Ge cylinder, may propagate errors. This claim is reflected in our findings, also. In our opinion, using the same phantom for calibration and verification is in some way a circular argument and may even completely hide calibration errors.

In the qualification process for PET scanners used in German multicenter trials, a somewhat different approach is followed (*2*), emphasizing testing of all equipment involved in the final analysis chain. Basically, each scanner is calibrated in terms of activity concentration, which is rescaled to standardized uptake values by normalization to the ratio of injected activity to body volume (approximated by patient weight). Therefore, careful cross calibration between PET scanner and dose calibrator is essential (*2,3*). The verification chain therefore starts with the dose calibrator, whose accuracy is checked by certified ^{68}Ge sources. This test not only verified the instrument itself but also facilitated the identification of errors in the subsequent chain. The PET scanner calibration and processing is then tested through measurement of a cylindric phantom filled with a known activity concentration of ^{18}F solution, relying on the accuracy of the calibrator. Data were acquired to a high statistical quality to facilitate the detection of systematic errors during subsequent analysis of reconstructed images.

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.