device and the intended use for such tests. For this reason we would like to comment on the proposed standardized tests for evaluating PET machine performance that were presented in a recent article (*1*).

For truly standardized measurements of performance, definite instructions are required so that tests can and will be implemented by everyone in a predictable manner. We were puzzled by the number of tests in (*1*) that required the user to choose acquisition parameters (e.g., the transmission scan, blank scan, axial acceptance angle, slice thickness, energy window, coincidence time window, machine wobble, reconstruction filters) "as they would [be set] for a typical patient study." Given such loose prescriptions the typical PET practitioner might sense incongruity with the stated tasks "to establish a common methodologic language that clearly defines the experimental measurements which are to be performed" and "to provide reliable tests that can be used to evaluate different scanners, despite the differences that exist among them (*1*)."

One can only guess why this latitude in test definition was selected. One reason might have been that there are too many acquisition parameters, some of which may be viewed as study dependent, to definitively specify for all existing and future scanners. If this is true, the term standardization might be a misnomer and should be replaced with a term such as guidelines. We also should be prepared for the numerous qualifiers that will be associated with each test result. A second reason for selecting many acquisition parameters, as in a typical patient study, is that the tests will yield information that we can use to more directly understand how faithfully PET records in-vivo positron concentration data. At best, this is a difficult task, as the authors clearly note. Therefore, we feel it is essential to bury the mindset that we can, through simple, stylized, static phantom measurements, unambiguously establish the superiority (inferiority), suitability (unacceptability) and accuracy (inaccuracy) of PET machines in performing particular dynamic patient studies. Moreover, the best description of the performance of a PET machine will always be provided implicitly by patient *study specificity and study sensitivity* and by direct comparison of in-vivo PET data with trusted in-vivo *gold standards.*

Therefore, we feel the exclusive focus of standardized tests should be to reliably and efficiently characterize the intrinsic performance of a PET device and its fundamental correction algorithms. These tests would then serve as guidelines for machine acceptance testing, and as an integral part of an annual machine quality control program. A test design philosophy like "the overriding concern of the EEC group is that the measurements would approximate a clinical situation as closely as possible, so that the measurements can be used to predict and interpret patient studies" seems misguided to us.

With few exceptions, the tests outlined in (*1*), with more rigidly defined acquisition protocols, would be useful for elucidating tomograph specific parameters of performance. We have a few remaining concerns. Originally, when PET machines did not have scatter corrections, the scatter fraction measurement was useful for defining the magnitude of this error in PET. This physical artifact is now well appreciated. There seems little value in having a specific measurement of scatter fraction as part of standardized tests. Since scatter fraction is heavily object-dependent and the object in (*1*) has no physical significance, we are uncertain about how to use the measured number. In knowing that one machine has a scatter fraction of about 0.23, another of about 0.19 and yet another of about 0.53, for this stylized phantom, tells us nothing about their overall performance for patient studies nor will it help us with quality control. What is useful to know is how well the scatter correction is functioning. Even here the test for that is weak. Deletion of the scatter fraction measurement of course impacts the proposed sensitivity measurement. We would argue that the value of the sensitivity test is to: (a) verify the manufacturer's overall sensitivity specifications and (b) to assist in tracking machine stability with time. Furthermore, an estimate of scatter fraction, of about the same accuracy as the three line source test proposed in (*1*), could be obtained by reconstructing with and without scatter correction.

We were puzzled by the author's discussion of noise equivalent count rate. It was not suggested as a method of presenting the measured data, rather it was just mentioned as something one might do. Although this might be a useful calculation for measurements of a realistic phantom, we feel it is not relevant to the purpose of these tests.

We felt the discussion of axial profile variation and its implication for reduced partial volume error might be misleading in that loss of quantitation due to finite resolution in the transverse direction was not mentioned. That is, a machine with fine axial sampling does not necessarily image small structures more accurately than another with coarser axial sampling.

To partially address the need for reliable information regarding comparative performance of different PET machines under *realistic* imaging conditions, we would hope that the PET manufacturers who coauthored (*1*) would collectively design and build anatomically detailed brain, e.g., (*2*), and body phantoms. These phantoms, fillable with defined activity concentrations, then could be scanned with each new generation of machines on the market, using explicit acquisition protocols that would represent a spectrum of clinical PET studies. In this way more prototypic and usable information on image resolution, noise, contrast and loss of quantitation could be available for evaluation by consumers at a time when it would be most valuable. Obviously, potential purchasers of PET devices would rely less on phantom data as information on the clinical performance of a current generation machine became available.

It is admirable that the authors of (*1*) initiated the task of specifying standardized tests of PET devices. It certainly is a valuable goal. We hope everyone will recognize both the utility and the limitations of these tests and will support continued effort toward their final definition.

REFERENCES

1. Karp JS, Daube-Witherspoon ME, Hoffman EJ, et al. Performance standards in positron emission tomography. *J Nucl Med* 1991;32:2342–2350.
2. Chang W, Madsen MT, Wang L, Kirchner PT. A 3-D brain phantom for ECT applications. *J Nucl Med* 1989;30:851.

A.N. Bice
R.S. Miyaoka
University of Washington
Seattle, Washington

REPLY: We appreciate the comments by Drs. Bice and Miyaoka regarding our article on standardized tests of PET scanner performance. However, there seems to be some confusion about the proposed performance standards in PET. Their letter states, " . . . we feel the exclusive focus of standardized tests should be to

reliably and efficiently characterize the intrinsic performance of a PET device . . . ." We couldn't agree more. That is why we do not believe that the PET manufacturers should design and build anatomically detailed brain and body phantoms for standardized tests. This is incompatible with Bice and Miyaoka's and our stated goal to characterize the intrinsic performance of a PET device using basic and well-established parameters, such as spatial resolution, sensitivity, scatter and count rate capability.

Drs. Bice and Miyaoka are also puzzled by "the number of tests that required the user to choose acquisition parameters . . . as they would be set for a patient study." It seems obvious to us that a BGO system requires a wider energy window than a NaI (Tl) system, and a system with fixed septa cannot be tested with a large axial acceptance angle, to give two examples. No single set of parameters can be fairly applied to all PET scanner configurations. These parameters are optimized by the manufacturer, but they will depend on the particular scanner. It is important to keep the parameters fixed for all tests, but it is not possible to fix them for all scanners.

Another confusing suggestion is to eliminate the test of scatter fraction but retain the test of scatter correction. A system with 5% scatter is clearly preferable to one with 95% scatter, since scatter correction only subtracts the estimated scatter contribution but not the noise associated with the scatter. Also, knowledge of the scatter fraction allows one to calculate the true sensitivity and true count rate as a function of activity. While the phantom selected has no "physical significance," it is not so unrealistic as to preclude comparisons between scanners. The value measured for intrinsic scatter fraction may change with a more realistic phantom, but the relative values between scanners are unlikely to change.

We were somewhat dismayed at the reference to the measurement of the accuracy of scatter correction as "weak" without a suggestion as to how to make it better. As the proposed measurements come into routine use on a variety of scanners, especially those newer systems whose specifications are not yet known, specific ideas as to improvements to these measurements will be welcomed.

Finally, we disagree with Bice and Miyaoka that purchasers of PET devices will "rely less on phantom data as information on the clinical performance of a current generation machine" becomes available. Clinical PET studies will always be evolving, as will PET scanners, while the performance measurements were designed to serve as standards for a substantial period of time. Both the intrinsic performance and the clinical experience will be important considerations to potential purchasers of PET scanners.

Joel S. Karp
Hospital of The University of Pennsylvania
Philadelphia, Pennsylvania
Margaret E. Daube-Witherspoon
Edward J. Hoffman
Thomas K. Lewellen
Jonathan M. Links
Wai-Hoi Wong
Richard D. Hichwa
Michael E. Casey
James G. Colsher
Richard E. Hitchens
Gerd Muehllehner
Everett W. Stoub

Patterns of Dementia in Alzheimer's Disease

TO THE EDITOR: We have read with a great interest Holman et al.'s article in the February issue of the *Journal* (*1*). This paper points out the wide variety of patterns observed in dementia, particularly in Alzheimer's disease (AD).

They confirmed previously published results obtained with [^{123}I]IMP (*2*) or the ^{133}Xe noninhalation method (*3*). Unfortunately, the statistics were undoubtedly incorrect, and Holman and coworkers failed to calculate the predictive values (PV) of the HMPAO tomograms. By using the author's method and the same notation (Q_i: pattern $_i$; AD + or AD− for the presence or no presence of AD) we can emphasize that the meaning of P (Q_B/AD+) as it appears in the Results section is wrong. Indeed, P (Q_B/AD+) does not represent the probability for the patient to have AD if Q_B is present, but exactly the opposite: the probability to encounter the Q_B pattern if AD exists. This is the Bayesian notation corresponding to the sensitivity of the test. By using Holman's results, sensitivity is equal to 27% (14/52).

Moreover, Holman and coworkers said that the positive predictive value (PPV) for Q_B patterns is 82% (Table 1; summary). This is incorrect. Indeed, the PPV corresponds to P (AD+/Q_B). This value (the negative PV) can be calculated only if the sample represents the probability of the distribution in all populations. Clearly, it is not true here since the prevalence, p, of AD can be assumed to be equal to 5% (for individuals older than 65 yr, no comment is made) and in Holman's study p is nearly equal to 50% (52/113)!

PPV can be obtained using Bayes' theorem, which results in the following relationship:

$$PPV = p \times \text{sensitivity}/(p \times \text{sensitivity}) + (1 - p)(1 - Sp),$$

where Sp is the specificity: P (Q_B−/AD−).

Holman's data, (sensitivity = 27%; specificity = 95%) and assuming p = 5%, results in a PPV of only 21% and not 82%, the result obtained by Holman et al. With a similar calculation, the negative predictive value (NPV) is 50%. We agree with Holman that Q_B is one of the most probable patterns of AD (but only 14/52), but it is not pathognomonic. What is true for Q_B is even more true for other patterns. In a previous study using the cerebellum as reference (*4*), we showed that the best cutoff value to discriminate AD from normals was 0.8, with a sensitivity and specificity of 0.6 and 1, respectively. Thus, the NPV (P(AD−/Q_B−)) was equal to 100%. The main goal of Holman and coworkers' paper was to provide interesting raw data for several diseases according to their different patterns. This leads to the conclusion that HMPAO brain tomograms are of very low value in determining diagnostic causes of memory or cognitive complaints, or both. Holman et al. also provided for calculations of predictive values for each pattern, but a correct application of the Bayes' theorem was needed.

REFERENCES

1. Holman BL, Johnson KA, Gerada B, Carvalho PA, Satlin A. The scintigraphic appearance of Alzheimer's disease: a prospective study using technetium-99m-HMPAO SPECT. *J Nucl Med* 1992;33:181–185.
2. Derouesne C, Rancurel G, Leponcin Lafitte M, Rapin JR, Lassen NA. Variability of cerebral blood flow defects in Alzheimer's disease on I-123-iodo-isopropylamphetamine and single photon emission tomography. *Lancet* 1985;11:1282.
3. Celsis P, Agniel A, Puel M, Demonet JF, Rascol A, Marc-Vergnes JP.