Evaluating Image Reconstruction Methods for Tumor Detection in 3-Dimensional Whole-Body PET Oncology Imaging

Carole Lartizien, PhD[1,2]; Paul E. Kinahan, PhD[1,3]; Richard Swensson[†], PhD[1]; Claude Comtat, PhD[2]; Michael Lin, MS[1]; Victor Villemagne, MD[4]; and Régine Trébossen, PhD[2]

[1]Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania; [2]Service Hospitalier Frédéric Joliot, Commissariat à l'Energie Atomique, Orsay, France; [3]Department of Radiology, University of Washington, Seattle, Washington; and [4]Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania

We compare 3 image reconstruction algorithms for use in 3-dimensional (3D) whole-body PET oncology imaging. We have previously shown that combining Fourier rebinning (FORE) with 2-dimensional (2D) statistical image reconstruction via the ordered-subsets expectation-maximization (OSEM) and attenuation-weighted OSEM (AWOSEM) algorithms demonstrates improvements in image signal-to-noise ratios compared with the commonly used analytic 3D reprojection (3DRP) or FORE+FBP (2D filtered backprojection) reconstruction methods. To assess the impact of these reconstruction methods on detecting and localizing small lesions, we performed a human observer study comparing the different reconstruction methods. The observer study used the same volumetric visualization software tool that is used in clinical practice, instead of a planar viewing mode as is generally used with the standard receiver operating characteristic (ROC) methodology. This change in the human evaluation strategy disallowed the use of a ROC analysis, so instead we compared the fraction of actual targets found and reported (fraction-found) and also investigated the use of an alternative free-response operating characteristic (AFROC) analysis. **Methods:** We used a non-Monte Carlo technique to generate 50 statistically accurate realizations of 3D whole-body PET data based on an extended mathematic cardiac torso (MCAT) phantom and with noise levels typical of clinical scans performed on a PET scanner. To each realization, we added 7 randomly located 1-cm-diameter lesions (targets) whose contrasts were varied to sample the range of detectability. These targets were inserted in 3 organs of interest: lungs, liver, and soft tissues. The images were reconstructed with 3 reconstruction strategies (FORE+OSEM, FORE+AWOSEM, and FORE+FBP). Five human observers reported (localized and rated) 7 targets within each volume image. An observer's performance accuracy with each algorithm was measured, as a function of the lesion contrast and organ type, by the fraction of those targets reported and by the area below the AFROC curve. This AFROC curve plots the fraction of reported targets at each rating threshold against the fraction of cases with (≥1) similarly rated false reports. **Results:** Images reconstructed with FORE+AWOSEM yielded the best overall target detection as compared with FORE+FBP and FORE+OSEM, although these differences in detectability were region specific. The FORE+FBP and FORE+AWOSEM algorithms had similar performances for liver targets. The FORE+OSEM algorithm performed significantly worse at target detection, especially in the liver. We speculate that this is the result of using an incorrect statistical model for OSEM and that the incorporation of attenuation weighting in AWOSEM largely compensates for this model inaccuracy. These results were consistent for both the fraction of actual targets found and the AFROC analysis. **Conclusion:** We demonstrated the efficacy of performing observer detection studies using the same visualization tools as those used in clinical PET oncology imaging. These studies demonstrated that the FORE+AWOSEM algorithm led to the best overall detection and localization performance for 1-cm-diameter targets compared with the FORE+OSEM and FORE+FBP algorithms.

Key Words: PET; observer performance; detection and localization; alternate free-response operating characteristic analysis; image reconstruction algorithms

J Nucl Med 2003; 44:276–290

We present a comparison of 3 image reconstruction algorithms for use in 3-dimensional (3D) whole-body PET oncology imaging. PET scanning with the labeled glucose analog ^{18}F-FDG is increasingly being used in whole-body oncology imaging to stage cancer and metastatic diseases in all regions of the body (*1*,*2*). Whole-body PET scanning, however, is typically constrained to short imaging times at each bed position to maintain a total scan duration that is acceptable to patients. The 5- to 10-min typical acquisition time for both transmission and emission data for each bed position in whole-body imaging results in images with, in general, poor signal-to-noise levels.

Noise reduction in whole-body emission images has been addressed by the use of statistical image reconstruction techniques with standard 2-dimensional (2D) acquisition mode data. Improvements in image signal-to-noise ratio

SOCIETY OF NUCLEAR MEDICINE

Received Oct. 2, 2001; revision accepted May 21, 2002.

For correspondence or reprints contact: Paul E. Kinahan, PhD, University of Washington Medical Center, 1959 N.E. Pacific St., Box 356004, Room NW040, Seattle, WA 98195-6004.

[†]Deceased.

E-mail: kinahan@u.washington.edu

(SNR) have been demonstrated with maximum likelihood (ML) algorithms compared with the standard analytic 2D filtered-backprojection (FBP) algorithm (*3,4*). The ML algorithms incorporate a model of Poisson photon counting statistics, albeit at the expense of computation time compared with FBP. The introduction of accelerated iterative algorithms (*5*) has significantly reduced computation time. The ordered-subsets expectation-maximization (OSEM) (*6*) algorithm in particular is now being routinely used for 2D whole-body image reconstruction in some centers (*7*), although convergence in the case of noisy data has not been proven.

An alternative approach to reducing statistical noise in whole-body imaging is by acquiring data in 3D mode (*8,9*). These data are then typically reconstructed with the analytic 3D reprojection (3DRP) algorithm (*10*). A synergistic combination is the use of iterative statistical reconstruction methods with 3D acquisition mode data. Fully 3D statistical reconstruction methods, however, are computationally intensive (*11–14*). As a faster alternative, hybrid methods have been proposed that combine 3D imaging with accelerated 2D statistical image reconstruction by rebinning the 3D data into a "stack" of 2D sinograms (*15,16*). For whole-body imaging, the Fourier rebinning (FORE) technique (*17*) compresses 3D datasets into 2D datasets with sufficient accuracy (*18*). The combination of FORE+OSEM, however, does not account for the effect of necessary data correction procedures on the statistical distribution of the rebinned data, even though the multiplicative correction terms for the effects of attenuation can be as high as 100 in some cases. A refinement of the hybrid approach was the incorporation of attenuation weighting (AW) in the statistical model of the data acquisition, resulting in the FORE+AWOSEM (where AWOSEM is attenuation-weighted OSEM) algorithm (*16*), which has shown improvements in contrast-to-noise trade-offs compared with both the FORE+OSEM and the 3DRP algorithms (*18*). The 2D AWOSEM algorithm incorporates the attenuation factors in the system matrix used to forward project the estimated sinogram based on the current estimate of the image. The estimated sinogram is then compared with the measured (attenuated) sinogram, thus approximately preserving the Poisson statistics assumed by the EM algorithm. This approach was first proposed for the EM algorithm by Hebert and Leahy (*19*) and can be extended to other data multiplicative correction terms (*20*). The 3D FORE+AWOSEM algorithm is somewhat more complex, as the 3D sinogram data must first be corrected for all physical effects, including attenuation, before the FORE step. To properly model the statistics, 2D attenuation factors are applied to attenuate the 2D sinograms, which are then reconstructed with AWOSEM. The attenuated 2D sinograms are Poisson-like, in that the variance is proportional to the mean. We also note that the proportionality constant is typically $\ll$1, which is the desired goal. Under these conditions we have shown that the behavior of the EM (and OSEM) algorithm is unchanged (*18*). The combinations of FORE+OSEM and FORE+AWOSEM may lead to apparent improvements in image SNR in clinically feasible reconstruction times, but the resulting effect on diagnostic utility is not clear. The purpose of this work is to evaluate the impact of 3 different reconstruction strategies now available in clinical practice (FORE+FBP, FORE+OSEM, and FORE+AWOSEM) on human observers' ability to detect foci of elevated FDG uptake, a characteristic of several malignancies. Figure 1 shows illustrations of these 3 reconstruction strategies for the same patient study.

Human observer detection capability is typically assessed by psychophysical studies that use the receiver operating characteristic (ROC) methodology (*21*). The analysis of ROC curves has been applied to conventional radiologic imaging, using displays of either single, transverse image planes or a central target image supported by adjacent image planes on either side to provide anatomic reference points (*22,23*). Similar methodologies have been adopted for ROC studies in PET and SPECT (*24–27*). This 2D or planar mode of display, however, substantially simplifies the volumetric display procedures routinely used for clinical interpretations of PET and SPECT images, which simultaneously display the 3 main planar sections (transverse, coronal, and sagittal) through the contiguous image volume. Another important limitation of the standard ROC analysis is that it uses only a single detection rating on each image or case and cannot consider observer's multiple reports of several possible targets.

For these reasons, we developed a different procedure for evaluating human observer performance with PET images, incorporating the same software tools for enhanced volume visualization used in PET oncology imaging. We used simulated data to overcome the practical impediment of acquiring large numbers of experimental datasets and to provide a gold standard for the location and contrast of the inserted targets. To replicate clinical conditions and to shorten the total reading times to feasible levels, we used multiple targets per volume and did not include any volumes without targets. These conditions are not consistent with ROC analysis, but they are compatible with an analysis of the observer's alternative free-response operating characteristic (AFROC) curve (*28,29*), which considers multiple reports and requires accurate localizations of the targets on the images. In addition to the AFROC approach, we also used a nonparametric analysis of measuring the fraction of actual targets that were found and reported ('fraction-found'), which provides a simple and robust measure of observer performance accuracy.

MATERIALS AND METHODS

For the studies reported here we used simulated 3D whole-body data with randomly located lesions (targets) of varying contrast levels. The simulation parameters were based on clinical relevance and a series of calibration studies. The data were reconstructed using the different image reconstruction methods, and human

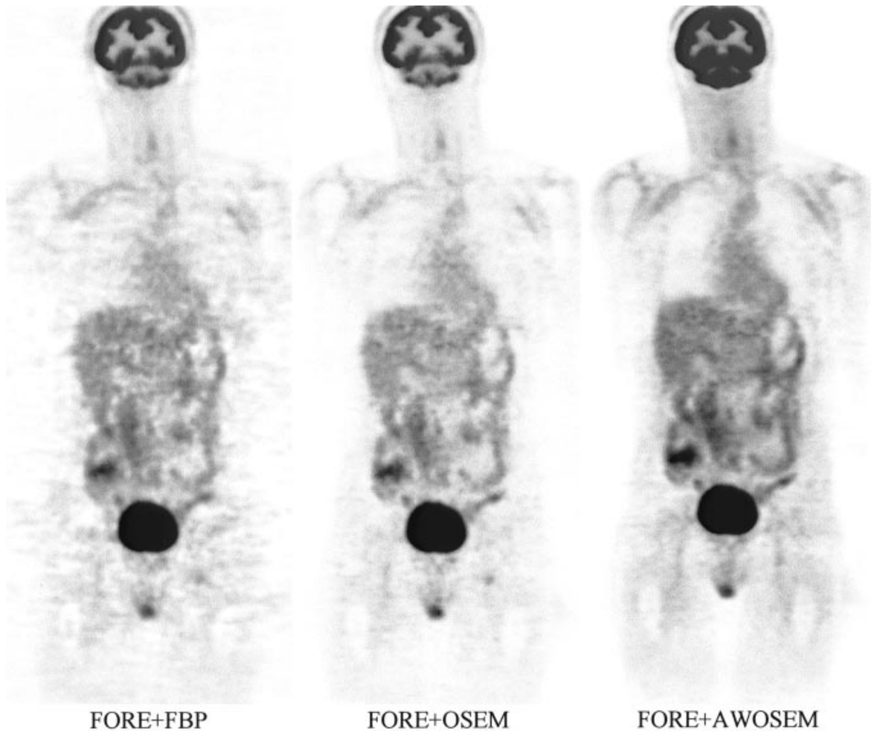


FIGURE 1. Illustration of the 3 reconstruction strategies. These images are same coronal views of same patient study reconstructed with FORE+FBP, FORE+OSEM, and FORE+AWOSEM algorithms.

observers attempted to localize the lesions using a modified version of the display software used for routine clinical studies. The results of the localization studies were analyzed with both a simple fraction-found approach and an AFROC analysis.

Data Simulations

We used the analytic (i.e., non-Monte Carlo) simulation method (ASIM) (*30*) to allow for the generation of multiple noisy realizations of 3D whole-body sinogram datasets. The principle of ASIM is to first analytically calculate noiseless transmission and attenuated emission sinograms on the basis of geometric specifications of the emission and attenuation objects (i.e., the whole-body phantom), the geometry of the scanner, the position of the object in the field of view, and the number of acquisitions (bed positions) for whole-body acquisitions. Sinograms of noiseless random and scattered coincidences are approximated by assuming that the total activity in each transaxial slice of the object is concentrated along the axis of the scanner. These values are scaled by measured scanner profiles. Second, predetermined levels of Poisson noise are optionally added to each of the 4 sinograms on the basis of user-specified count levels. These sinograms are then combined appropriately to duplicate measured raw data sinograms with accurate noise properties. Finally, the raw data are corrected for attenuation, scatter, and randoms by the same techniques as those used in practice, assuming that the correction terms, although they may be noisy, are accurate.

The whole-body simulator accounts for effects that are important in whole-body PET imaging, including attenuation, random and scattered coincidences arising from the activity inside and outside the field of view, detector efficiencies, activity decay between bed positions, and noise arising from the transmission scan (*30*). In addition, resolution effects and detector efficiencies can be included. The different sources of noise can be turned on or off as desired. The advantage of this approach is that multiple independent realizations of 3D whole-body sinograms can be rapidly generated, unlike Monte Carlo methods, which track individual photons. The disadvantage of this approach is that scatter and random scaling profiles must be measured for each tomograph and that there is no energy information (if needed). The simulator was validated by comparing the means and variances measured from multiple realizations of simulated and measured studies based on identical phantoms, with several activity levels inside and outside the scanner field of view. The simulator was also shown to predict plausible results for a more realistic whole-body geometry using a simple model of human FDG distribution when compared with clinical scans.

We generated data that reproduced the FDG distribution in the torso, with a geometry based on the 3D mathematic cardiac torso (MCAT) phantom (*31*) and with the addition of head, arms, and bladder objects to simulate a true imaging protocol. For the simulated scanner, we modeled the geometry and the characteristics of the ECAT EXACT HR+ scanner (CTI/Siemens, Knoxville, TN) (*32*). The tomograph consists of 32 rings of detectors acquiring data over 15.5 cm in the axial direction.

Simulated emission data were generated in the 3D mode of acquisition with a maximum ring difference of 22, which is equivalent to a maximum acceptance angle of 12.5°. The lines of response were mashed axially into groups of 4 or 5 lines of response, equivalent to a span factor of 9 (*33*). These acquisition parameters replicated those used in clinical practice.

The noise level in the phantom data was set to an average from clinical whole-body scans corresponding to an injection rule of approximately 3.7 MBq/kg and a 5-min acquisition per bed position starting 60 min after injection. For an acquisition centered on the liver, the average numbers of true and random coincidences were set to 22 and 35 million counts, respectively. Because of the wide variability in how transmission scans are acquired and processed, noiseless attenuation factors were used to avoid including the noise in the transmission scan as a free parameter of the study.

Image Reconstruction

The 3D emission datasets were rebinned using the same implementation of FORE into sets of contiguous transaxial 2D sinograms. Images were reconstructed using the FORE+FBP, FORE+OSEM, and FORE+AWOSEM algorithms. The OSEM and AWOSEM implementations used were those described by Comtat et al. (*16*), whereas the FBP implementation used was the standard software available on the ECAT EXACT HR+ scanner. The OSEM and AWOSEM algorithms used 16 subsets and 4 iterations.

To control the contrast-to-noise trade-off, the images reconstructed by FBP varied the cutoff of a Hann apodizing window, whereas the OSEM and AWOSEM algorithms varied the full width at half maximum (FWHM) of the kernel of a postreconstruction 3D gaussian smoothing filter. This form of regularization for the FORE+OSEM and FORE+AWOSEM algorithms is problematic as the objective function (the log likelihood in this case) is not maximized, and so the solution unfortunately depends on the starting point. Further, the effectively nonconvex objective function may not even have a unique global maximum. Algorithms with convex objective functions that are iterated to convergence may have improved detection task performance. The purpose of this study, however, was to compare algorithms used in clinical practice for whole-body oncology image reconstruction.

The smoothing parameters for the 3 reconstruction strategies were selected by maximizing the contrast-to-noise ratio (CNR), closely related to a nonprewhitening matched filter (*34*), for 1-cm-diameter spheric targets placed inside a simulated elliptic cylinder with dimensions similar to the MCAT torso (200 × 128 mm). Five targets were added at different radial positions within the cylinder with a contrast of 8:1 defined as the target-to-background activity concentration ratio, that is, (concentration in the target/concentration in the background). The attenuation coefficient for all objects was equivalent to water and the sinograms had a noise level similar to that used for the final observer study. Twenty-five noisy realizations of this configuration were generated and reconstructed by each of the 3 algorithms. The FWHM of the gaussian smoothing filter used with OSEM and AWOSEM was varied from 4 to 14 mm. For the FBP images, the cutoff frequency of the Hann-windowed ramp filter was varied from 0.05 to 1.0 of the Nyquist rate, corresponding to a smoothing with a gaussian filter of 4.2- to 21.5-mm FWHM.

For each smoothing value or cutoff frequency the image CNR for each target was calculated as:

$$CNR = \frac{\langle (T - B)/B \rangle}{\sqrt{\sigma^2(T) + \sigma^2(B)}}, \qquad \text{Eq. 1}$$

where T and B are the activity concentrations, measured using 8-voxel volumetric regions of interest placed over the targets and background regions in the reconstructed image volumes, $\langle\ \rangle$ represents the ensemble average, and $\sigma^2(T)$ and $\sigma^2(B)$ are the variances of these activities estimated across the 25 realizations.

A transverse section of a typical reconstructed image of the phantom used for the free parameter optimization study is shown in Figure 2. The average CNRs in the FORE+FBP, FORE+OSEM, and FORE+AWOSEM reconstructed images are plotted in Figure 3 as a function of the free parameter.

SOCIETY OF NUCLEAR MEDICINE

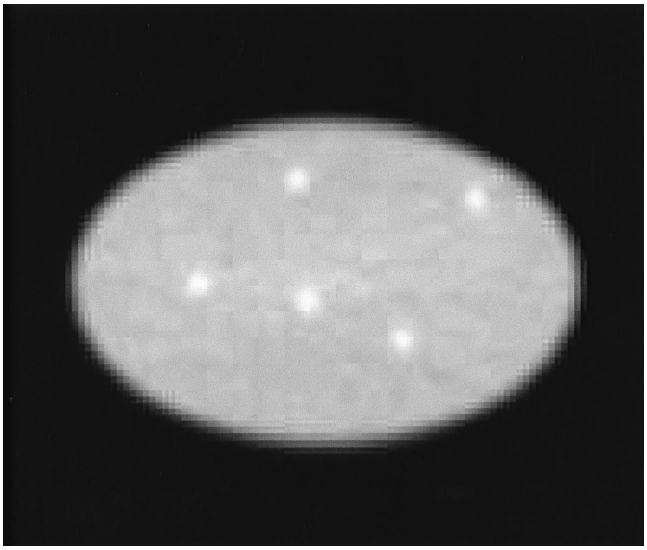


FIGURE 2. Transaxial section through reconstructed volumetric image of phantom used for optimization study illustrates locations of 5 targets, each 1 cm in diameter.

The CNR for the FBP algorithm index was maximum for a cutoff frequency of 0.4 of the Nyquist rate (corresponding to a smoothing with a gaussian filter of 10.6-mm FWHM). The CNR for the OSEM and AWOSEM algorithms both peaked with a postreconstruction gaussian smoothing filter of 12-mm FWHM. These values were used for the observer study described below.

Volumetric Observer Tool

For the purpose of evaluating observer performances in clinical practice, we have developed an “observer tool” as part of an enhanced volume visualization software tool used in clinical PET oncology imaging. The volumetric observer scoring procedure was incorporated into our dual/fusion clinical software visualization tool, which is based on the Clinical Application Programming Package (CTI, Knoxville, TN), which in turn is a superset of the Interactive Data Language IDL (Research Systems Inc., Boulder, CO). The dual/fusion display (Fig. 4) allows the standard procedure of interpreting whole-body PET scans by searching a contiguous volume in 3 orthogonal directions. It also supports the simultaneous display of emission and transmission image volumes with linked cross-hairs or image fusion. The dual/fusion display can load sequential emission scans for longitudinal studies and supports simple region-of-interest functions. When the observer tool is used with this dual/fusion display, all reports of target locations and confidence ratings are recorded and linked to their specified 3D positions within the emission-image volume. An observer can review and edit all of this information and can choose to display or hide the superimposed image markers for locations of the reported targets.

Calibration Study

Two important considerations for the observer study are contrast levels for the added targets and distance threshold used to determine when a target has been correctly reported. The target contrast levels must be selected so that the targets are sufficiently detectable by human observers, but not too obvious.

When observers report a target by clicking on its location with the mouse, it is unlikely that they will point exactly at the center

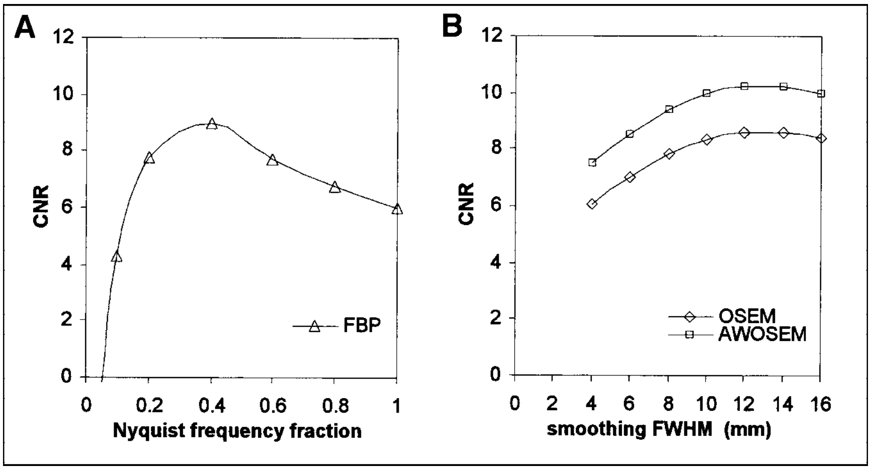
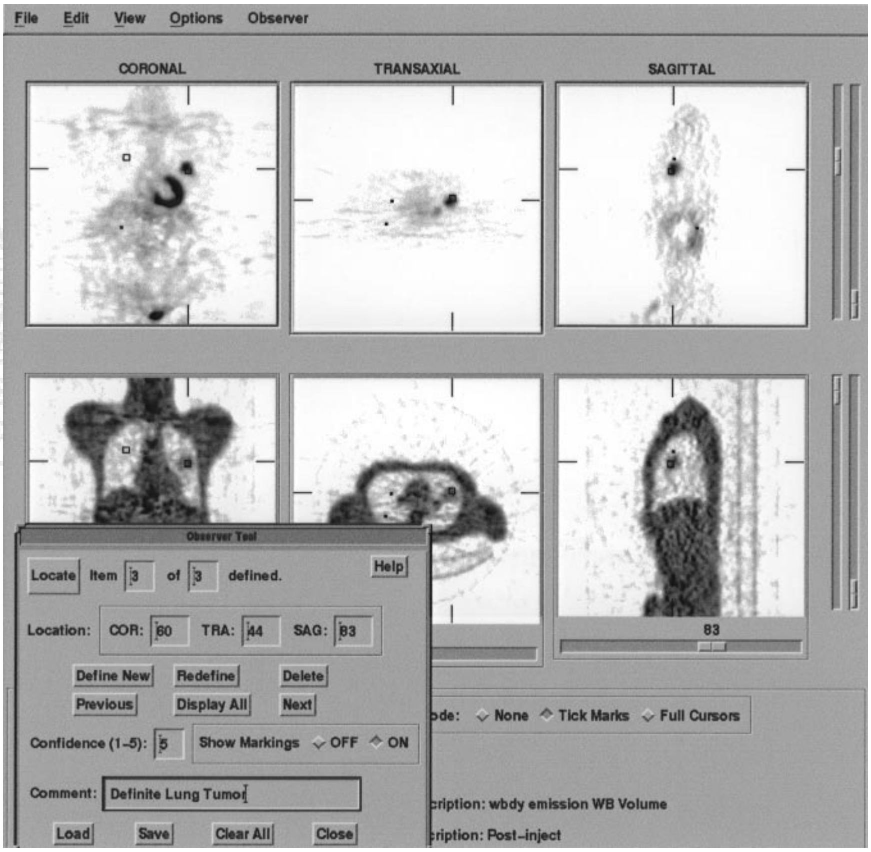


FIGURE 3. Measured CNR in image volumes according to Equation 1, averaged over 5 targets shown in Figure 2. (A) FBP images: CNR vs. cutoff frequency of Hann-windowed ramp filter. (B) OSEM and AWOSEM images: CNR vs. FWHM of post-reconstruction gaussian smoothing filter.

of the target. The distance between the real and the reported location may depend on the observer's visual acuteness and manual dexterity, on the image display and resolution, and on the lesion contrast. This will be a 3D Euclidean distance, because the observer can localize the target by clicking on any 1 of the 3 views (coronal, transverse, or sagittal). The distance threshold needs to be large enough to allow for observer localization error, but small enough to minimize the chance identification of a true target.

We performed a preliminary calibration observer study to determine an appropriate value for the threshold distance and to calibrate sets of target contrast levels that would cover the entire range of detectability within each organ of interest. Twenty-five 3D emission scans were simulated for the calibration study, each corresponding to a 2-bed position whole-body scan extending 28 cm from the upper part of the lungs to the bottom of the liver. Five targets of 1-cm diameter were randomly distributed in the lungs, the liver, and the surrounding soft tissues. All targets in a given volume had the same contrast defined as the target-to-organ activity concentration ratio. Five different contrasts of 12:1, 10:1, 8:1, 6:1, and 4:1 were simulated. Data were reconstructed with the FORE+OSEM algorithm using 16 subsets, 2 iterations, and a 7-mm FWHM gaussian smoothing. Three observers participated in the study, each reading 5 images per contrast level (25 volumetric images).



Item 3 of 3 defined.

Location: COR: 80 TRA: 44 SAG: 83

Confidence (1–5): 5

Show Markings ○ OFF ● ON

Comment: Definite Lung Tumor

ode: ○ None ● Tick Marks ○ Full Cursors

ription: wbdy emission WB Volume

ription: Post–inject

FIGURE 4. Illustration of dual/fusion display software and linked observer tool. Top 3 images are orthogonal views of reconstructed whole-body PET oncology scan using ^{18}F-FDG as tracer. Bottom 3 images are orthogonal views of corresponding reconstructed transmission scan. Linked observer tool is controlled from additional window overlaid (for purposes of this illustration only) on bottom left region of dual-volume viewer.

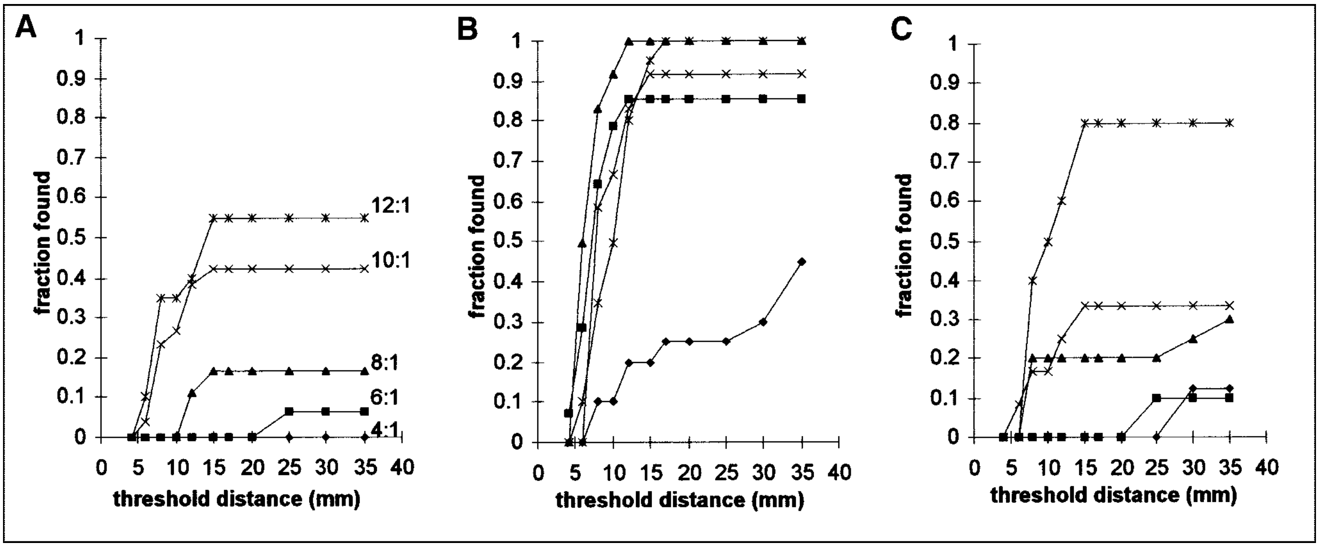


FIGURE 5. Fractions of actual targets found for different target contrasts (4:1, 6:1, 8:1, 10:1, and 12:1) as a function of threshold distance in different organs: lungs (A), liver (B), soft tissue (C).

Figure 5 shows the variations of the fraction-found averaged over observers for different target contrasts and for the different organs as a function of the distance threshold. The most significant changes occurred between 5 and 15 mm. The distance threshold was thus set to the upper limit of 15 mm for subsequent observer studies. Assuming equal localization errors in all 3 dimensions, the 3D distance threshold of 15 mm corresponds to an average 2D tolerance of approximately 12 mm in a single image plane. This value of 15 mm corresponds to approximately 3 voxels of the reconstructed image.

The measured fractions of targets found increased as a function of the target contrast, but at different rates for lesions in the liver, in the lungs, and in the soft tissues. Accordingly, separate sets of 5 contrast levels were selected for each organ (Table 1), corresponding to anticipated fraction-found values of 0.1, 0.3, 0.5, 0.7, and 0.9. These values were chosen to sample the range of the fractions of actual targets found as indicated.

Observer Detection Performance Study

For the main observer study, we simulated 3D acquisitions of the extended MCAT phantom over a 41-cm axial extent corresponding to 3 bed positions. Spheric 1-cm-diameter targets were inserted at randomly generated locations within the phantom, respecting a minimal distance of 1 cm from the edge of an organ or from another target. The number of targets per volume was set to 7, but the number of targets per organ varied randomly on the basis of a multinomial probability distribution with a mean of 2.5 targets each for the lungs and the liver and 2.0 targets for the other soft-tissue regions. On the basis of the ratio of the target and organ volumes, the probability of finding a target by chance was estimated to be 2 × 10^{-4} in the liver, 10^{-4} in the lungs, and 2 × 10^{-5} in the soft tissues. The contrast of each target within a given organ was randomly chosen from among the predetermined set of 5 values selected by the preliminary calibration study. There were 50 whole-body volumetric images (containing 7 targets each) leading to approximately 25 targets of each contrast level for the liver and the lungs (2.5 targets per organ × 50 volumes/5 contrast levels). The actual number of targets per contrast is reported in Table 1 for the 3 organ types. The 50 noisy projection sets were reconstructed with each of the 3 reconstruction algorithms (FORE+FBP, FORE+OSEM, and FORE+AWOSEM), resulting in a final set of 150 whole-body volumetric images.

Five observers participated to the study (1 nuclear medicine physician and 4 physicists working in PET facilities). Four observers were experienced in reading PET images and 1 observer was highly experienced in psychophysical studies, but less familiar with PET imaging. All observers were trained using the images generated for the calibration study. Observers practiced by rating some targets and comparing their results with the noise-free images using the dual/fusion volume viewer to display the locations they marked on both images. No time limits were imposed within either the training or the actual studies, and observers could "window" and control the display color scales as in clinical practice.

SOCIETY OF NUCLEAR MEDICINE

TABLE 1
Contrast Values of Targets Inserted in MCAT Phantom

Fraction-found	0.1	0.3	0.5	0.7	0.9
Lungs	7 (21)	8.5 (26)	9.5 (26)	11 (26)	13 (22)
Liver	3.75 (26)	4.25 (28)	4.75 (32)	5.5 (28)	6.25 (26)
Soft tissue	7 (15)	9.5 (13)	11 (23)	12 (18)	12.5 (20)

Numbers in parentheses refer to actual numbers of targets per contrast used in final observer study.

Reading sessions were interrupted and later resumed whenever necessary. Observers were told that each volume would contain exactly 7 targets but they did not know the precise number of targets within each organ. They were required to report 7 separate locations per volume, rating each report on a 5-point ordinal scale of confidence (5 = definite strong target, 4 = medium-strong target, 3 = medium-weak target, 2 = weak target, and 1 = probably not a target). This method was aimed at forcing all observers to adopt an identical criterion for the number of possible targets reported per volume. The analysis assigned a default rating of zero to all unreported targets. Two of the 5 observers read the complete set of 150 whole-body volumetric images from all 3 reconstruction algorithms, and the other 3 observers each read 100 volumetric images from 2 of the 3 conditions. The studies were arranged such that each set of 50 images from each algorithm was read by 4 separate readers. The reconstructed images were split into subsets of approximately 17 volumes that were presented to the observers in random order to reduce reading order effects. The subset approach was used to reduce the observer's need for constant readjustment to differing physical image characteristics and to balance the effects of any systematic changes in observers' strategies or performance over the course of the experiment.

Observer Detection Performance Analysis

The fraction of actual targets found and reported (fraction-found) by each observer was calculated as a function of the target contrast for each of the 3 organ types in which targets were inserted and for each of the 3 reconstruction strategies. Error bars were calculated, given the number of sampled targets, by assuming that the estimated fraction-found followed a binomial law. The overall fraction-found, for all targets within each organ (pooled across independent target samples at all 5 contrasts), was used for statistical tests of differences in performance among the 3 reconstruction algorithms. These intermodality comparisons used *z*-score tests of the difference between the overall fractions of correctly reported targets.

A goal of this observer detection study was to incorporate the major characteristics of clinical interpretations performed for whole-body PET scan. To that purpose, the set of simulated whole-body data contained multiple targets per volume. Observers viewed a volumetric display of the data and had to locate and rate the targets in their reports. Because nontarget volumetric images were not included in the experiment, the acquired data were not compatible with a conventional ROC analysis but were compatible with an AFROC analysis.

Free-response methods have been proposed to evaluate observer detection performance in more realistic tasks by measuring localization accuracy with multiple targets per image (*29*). The AFROC curve plots the probability of a correct target report at each rating cutoff as a function of the probability that the observer will also report 1 or more false targets at the same rating cutoff. Thus, the AFROC curve measures an observer's combined detection and localization performance in detection tasks that present multiple targets, either with or without the use of nontarget cases. The area below an AFROC curve may be interpreted as the probability that a specified target would be rated higher than the most suspicious nontarget location (*29*) or correctly localized (by first choice) on an image containing only that single target (*28*). Swensson (*28*) proposed a mathematic model that represents the observer's likelihood of making 1 or more false reports of nontarget locations in terms of an assumed latent perceptual variable for the most suspicious (maximum-value) nontarget location in the relevant area or volume. In the case of an AFROC analysis, this model implies that the AFROC data can be fitted with the same statistical software used for the conventional (binormal) ROC methodology.

AFROC curves in this study were generated from an observer's rating data for the different tissue types and different reconstruction strategies. Each set of these AFROC data consisted of the rating frequencies across the 6 ordinal categories (0–5) for (a) the targets of all contrast levels within a given organ and (b) the highest-rated false reports within that organ in each of the 50 separate volumes. As was done for all unreported targets, the analysis assigned the default rating of zero to any volume that had no false reports within the given organ (because the most suspicious normal finding had failed to reach the observer's minimum threshold for an explicit report).

These sets of AFROC data were fitted using the CORROC program developed by Metz et al. (*35*) for pairwise comparisons of the separate, but correlated, ratings from 2 conditions that presented the same cases. The CORROC program assumes bivariate normal distributions for the 2 distributions of an observer's latent perceptual variables that underlie ratings of the same cases in the 2 separate conditions. For AFROC data, those cases refer either to the locations of actual targets or to the maximum-value nontarget locations within the sampled areas or volumes. In this study, the same target locations and nontarget organ volumes were assigned ordinal ratings on the basis of separate interpretations of the images reconstructed by different algorithms. Rating correlations would be induced by the targets of different contrast levels. When the rating judgments in 2 separate modes are positively correlated, a CORROC analysis permits more sensitive statistical tests for intermodal differences in such estimated indices as areas below their 2 fitted AFROC curves. Our intermodality comparisons used *z*-score tests of the null hypothesis that the difference in areas below the 2 fitted AFROC curves arose from binormal AFROC curves with equal areas.

RESULTS

Coronal sections of images reconstructed from 1 of the simulated 3D datasets are shown in Figure 6 for the 3 algorithms: FORE+FBP, FORE+OSEM, and FORE+AWOSEM. The top row of Figure 7 plots the detectability curves (fraction-found vs. target contrast) for 1 observer for each reconstruction strategy and for each organ type. The ranking of algorithms was similar for all observers and the absolute levels of performance were also similar for all but the least experienced observer. The bottom row of Figure 7 plots the fraction-found averaged over all observers for each reconstruction strategy and organ type. Tables 2, 3, and 4 show values of the fraction-found in the lungs, liver, and soft tissues for each observer, averaged over target contrast, as well as results of *z*-score tests for comparing each pair of reconstruction strategies.

In Figure 7 there are noticeable cases of apparent nonmonotonic behavior of the fraction-found as a function of contrast level. This is particularly evident in the fraction-found in liver using FORE+OSEM (Figs. 7B and 7E) between the second and third contrast levels. To determine

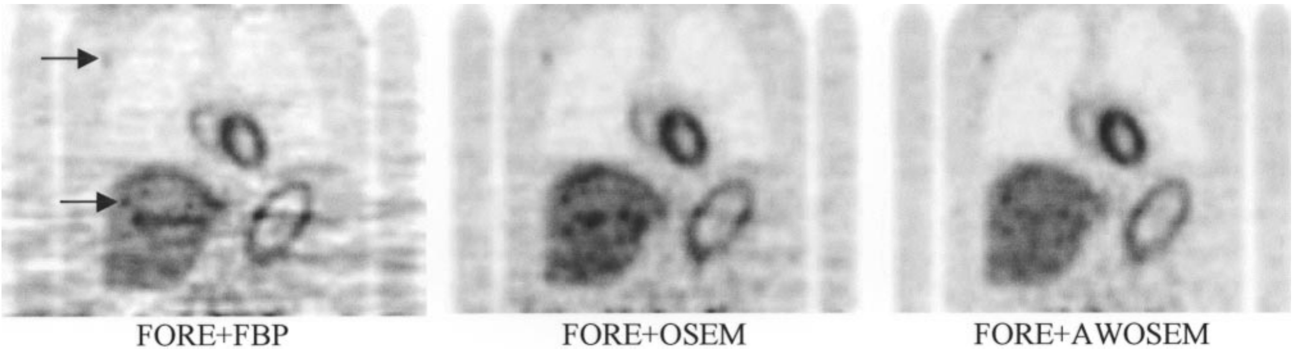


FIGURE 6. Coronal sections of typical 3D simulated dataset reconstructed by each of 3 algorithms: FORE+FBP, FORE+OSEM, and FORE+AWOSEM. Data were simulated based on an extended MCAT phantom, which in this example shows 2 targets (arrows).

if this could be explained by the necessarily small sample size (~25 targets per contrast level), we estimated the 95% confidence intervals for the measured difference in fraction-found between adjacent contrast levels assuming the true fraction-found increases with contrast. If the confidence intervals drop below zero then there is a possibility of a decrease in the measured fraction-found values. The underlying fraction-found ("truth") value FF_i at each contrast level i was based on a simple linear fit to the measured values of Figure 7, which was then used to generate the expected difference $E\{FF_{i+1} - FF_i\}$ between adjacent contrast values i and $i + 1$. The measurement error of the fraction-found was assumed to be based on a binomial model, yielding $s(FF_i) = \sqrt{p_i(1 - p_i)/N_i}$, where $p_i = R_i/N_i$ is the probability of a correct report, and R_i and N_i are the number of correct reports and targets for contrast level i. With these assumptions the 95% confidence interval for the measured difference in fraction-found between adjacent contrast levels is given by $(E\{FF_{i+1} - FF_i\} \pm 1.96\sqrt{p_i(1 - p_i)/N_i + p_{i+1}(1 - p_{i+1})/N_{i+1}})$, where the factor of 1.96 corresponds to the 95% confidence interval of (mean ± 1.96 × SD). These intervals were estimated for the 4 contrast changes for all 9 combinations of algorithms and organs, yielding 36 contrast changes. Recalling that the possible fraction-found ranges from 0 to 1, the average 95% confidence interval for the measured difference in fraction-found between the 36 adjacent (increasing) contrast levels was 0.00 to 0.27. In addition, 17 of the 95% confidence intervals included negative values, albeit marginally. For the specific case of the change from the second to third contrast levels in liver, the 95% confidence interval for the measured difference was −0.04 to 0.21 for FORE+OSEM, 0.00 to 0.26 for FORE+FBP, and −0.02 to 0.24 for FORE+AWOSEM. These results indicate that the apparent nonmonotonic behavior of the fraction-found as a function of contrast level is within measurement error and is a consequence of the limited sample size at each contrast level.

SOCIETY OF NUCLEAR MEDICINE

Figure 7 shows that, with the levels of target contrast chosen for each tissue type, the fraction-found varied across a wide range between 0 and 1.0 using images from the 3 reconstruction algorithms. For the fraction-found among targets pooled across all contrast levels (Tables 2–4), the binomial SE of the fraction-found was about 0.05 for all 5 observers and for all 3 organs.

The mean fraction-found was higher with images reconstructed with the FBP algorithm than with images reconstructed using the standard OSEM algorithm for all organs. This difference of detection performance was region specific, with larger differences in the liver and smaller differences for targets located in the soft tissues (indicated by the P values in Tables 2–4). Indeed, this difference was not statistically different for any observer in the soft tissues, whereas it was significant for all observers in the liver.

Comparisons of the FORE+AWOSEM and the FORE+OSEM algorithms in Figure 7 indicate that the AW of OSEM led to an improved fraction-found as a function of target contrast for all observers. This difference was statistically significant for all observers for targets located in the liver and for 2 of 3 observers for targets located in the lungs and the soft tissues.

Finally, comparisons of the FORE+AWOSEM and the FORE+FBP algorithms indicate that the iterative algorithm tended to increase the average fraction-found, although this difference was not consistently significant across observers or organs.

The top row of Figure 8 plots the estimated AFROC curves from 1 observer for the 3 reconstruction algorithms and for the 3 organ types. The bottom row of Figure 8 shows similar plots of the representative AFROC curves obtained by averaging across observers the estimated linear parameters of their individual fitted AFROC curves. The AFROC curves from all individual observers had similar rank orderings for the 3 algorithms, and the averaged AFROC curves in the bottom row of Figure 8 show little crossover. This means that the relative ranking of the 3 algorithms remained consistent at all levels of specificity and, thus, the A_L, the area below the AFROC curve (as estimated by the CORROC procedure), was a reasonable figure of merit for comparing the different reconstruction strategies. Tables 5-7 give the estimates of A_L for each individual observer, each organ,

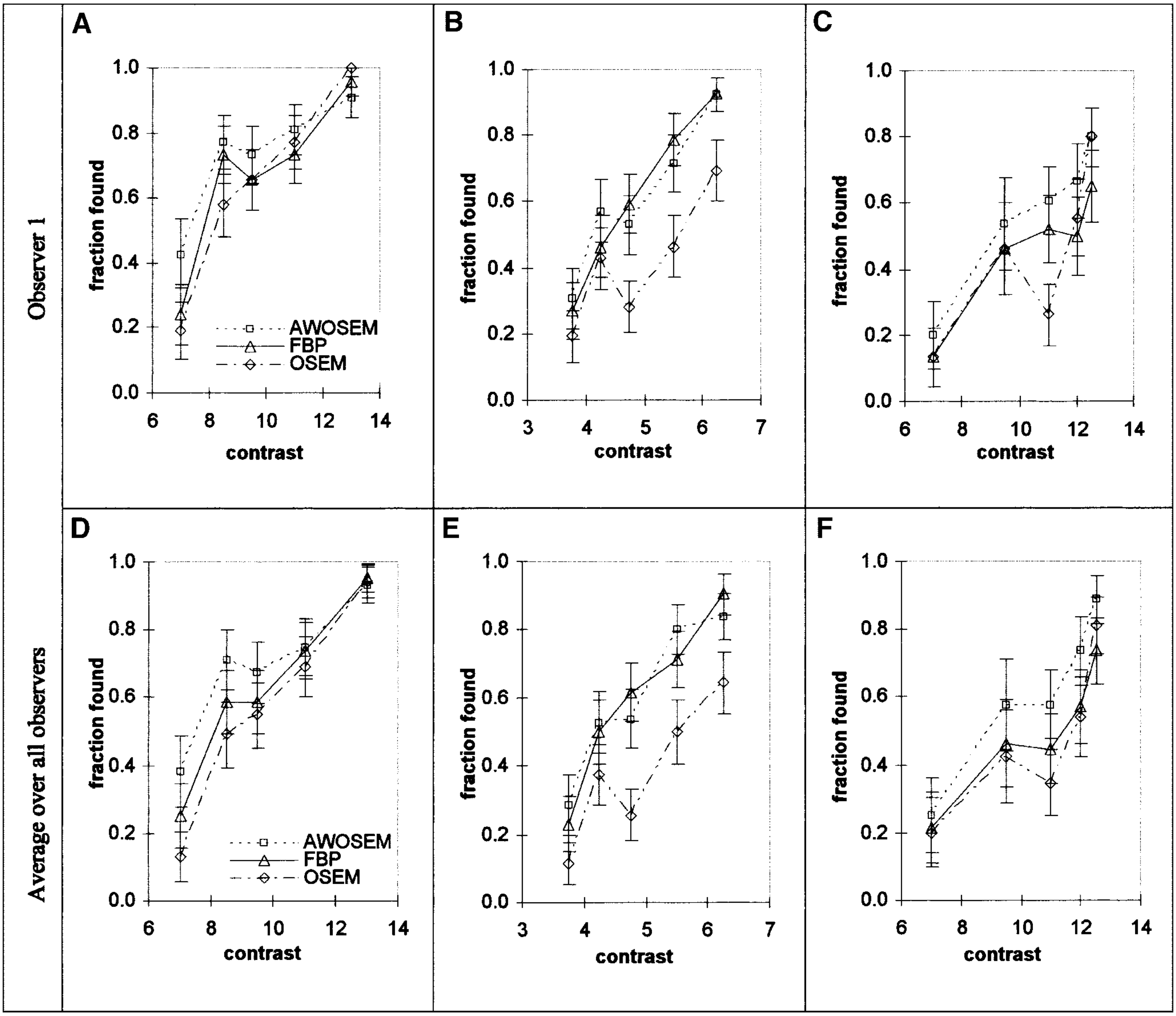


FIGURE 7. Fractions of actual targets found obtained for 1 observer (A–C) and averaged over all observers (D–F) for 3 reconstruction strategies as function of target contrast in lungs (A and D), liver (B and E), and soft tissues (C and F).

and each reconstruction strategy, together with the results from correlated *z*-score tests of differences in A_L for each pair of reconstruction strategies. The SEs for individual estimates of A_L were obtained from the ML fitting procedure and were similar for all observers and all organs with values of about 0.04.

The AFROC curves for the lungs (Fig. 8D) and the soft tissues (Fig. 8F) had similar shapes and were ranked in

SOCIETY OF NUCLEAR MEDICINE

TABLE 2
Fraction-Found Results for Lungs, Averaged Over Contrast Levels

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	FF	0.67	0.66	0.67	0.51	
OSEM	FF	0.64	0.52	0.52		0.58
AWOSEM	FF	0.74	0.67		0.66	0.71
FBP vs. OSEM	*P*	0.342	0.012	0.009		
AWOSEM vs. FBP	*P*	0.130	0.446		0.009	
AWOSEM vs. OSEM	*P*	0.062	0.009			0.015

Obs. = observer; FF = fraction of targets found; *P* = probability value of *z*-score test for each intermodality comparison.

TABLE 3
Fraction-Found Results for Liver, Averaged Over Contrast Levels

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	FF	0.61	0.69	0.58	0.51	
OSEM	FF	0.41	0.39	0.28		0.42
AWOSEM	FF	0.61	0.63		0.49	0.66
FBP vs. OSEM	*P*	<0.001	<0.001	<0.001		
AWOSEM vs. FBP	*P*	0.500	0.156		0.406	
AWOSEM vs. OSEM	*P*	<0.001	<0.001			<0.001

Obs. = observer; FF = fraction of targets found; *P* = probability value of *z*-score test.

similar orders. They showed that the FORE+AWOSEM algorithm substantially improved the target detectability in these organs**,** as compared with FORE+FBP, and that FORE+FBP was ranked higher than FORE+OSEM. For the liver, FORE+FBP and FORE+AWOSEM led to similar levels of performances (Fig. 8E), and the relative decrease in performance of FORE+OSEM was markedly larger compared with that measured in the lungs and soft tissues. Statistical tests of the differences in area under the CORROC fitted AFROC curves showed general decreases in the *P* values, compared with those for the *z*-score test of the fraction-found. For example, although comparisons of the fraction-found between images reconstructed by the FORE+FBP and FORE+OSEM algorithms failed to demonstrate any significant difference for targets in the soft tissues, the differences in estimated A_L indicated significant improvements with the FORE+FBP algorithm for 2 of the 3 observers. Similar comments apply to the comparison of the FORE+AWOSEM and FORE+FBP algorithms in the lungs and in the soft tissues.

In summary, observer detection performances were significantly higher with images reconstructed with the FORE+AWOSEM algorithm than with images reconstructed using the FORE+OSEM algorithm for all organs. FORE+FBP allowed higher target detectability than FORE+OSEM for all organs, although the difference was not significant for targets located in the soft tissues. Finally, FORE+AWOSEM produced better overall target detection and localization as compared with FORE+FBP, although detection performances were equivalent for targets located in the liver.

DISCUSSION

For this study an important decision was the choice of reconstruction parameters for each algorithm because the human observer studies could not be repeated for choice of parameters. Instead, we relied on numerical observers to select optimal parameters for each algorithm. We then compared these optimal implementations of each reconstruction algorithm. We chose to set the reconstruction parameters (the cutoff frequency of the Hann-windowed ramp filter for FBP and the FWHM of the 3D gaussian smoothing kernel for OSEM and AWOSEM) that would optimize the CNR, which is equivalent to the nonprewhitening matched filter (NPWMF) numerical observer (*34*). For some cases, this image-based criterion is correlated with the detection performance of human observers for signal-known-exactly and for background-known-exactly detection tasks (*36*), although its correlation with human observers has not been explored for whole-body PET images. The selected reconstruction parameters optimized the CNR for the target size and the level of statistical noise used in the human observer studies.

With these parameter settings, all observers showed improvements in detectability for images reconstructed with FORE+AWOSEM and FORE+FBP, as compared with FORE+OSEM, although these differences in detectability

SOCIETY OF NUCLEAR MEDICINE

TABLE 4
Fraction-Found Results for Soft Tissue, Averaged Over Contrast Levels

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	FF	0.47	0.60	0.57	0.36	
OSEM	FF	0.45	0.48	0.48		0.49
AWOSEM	FF	0.58	0.67		0.58	0.65
FBP vs. OSEM	*P*	0.382	0.065	0.114		
AWOSEM vs. FBP	*P*	0.065	0.137		0.001	
AWOSEM vs. OSEM	*P*	0.035	0.004			0.016

Obs. = observer; FF = fraction of targets found; *P* = probability value of *z*-score test.

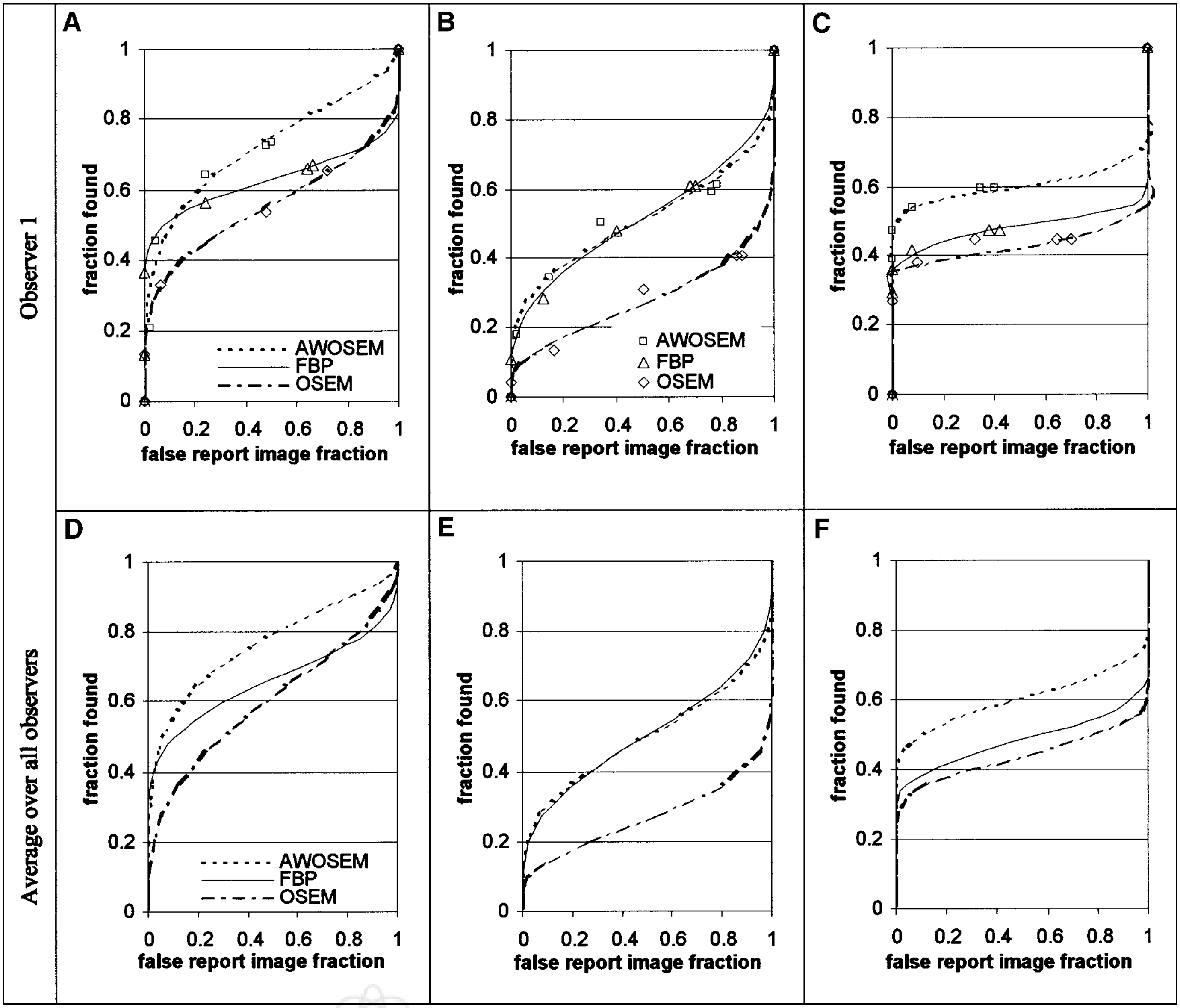


FIGURE 8. AFROC curves obtained for 1 observer (A–C) and averaged over all observers (D–F) for 3 reconstruction strategies as function of target contrast in lungs (A and D), liver (B and E), and soft tissues (C and F).

SOCIETY OF NUCLEAR MEDICINE

were region specific. This rank ordering was clearly seen in the performance averaged across all observers, the mean fractions of actual targets found and reported (bottom row of Fig. 7), and the parameter-averaged fitted AFROC curves (bottom row of Fig. 8). The numerical results of Tables 2–7 are summarized in Table 8 by averaging the fraction-found and the areas under the AFROC curves for all observers and also for all regions.

One result of this study is that the iterative FORE+OSEM algorithm produced the poorest detection performance. This is despite what seems to be a growing consensus that iterative methods generally produce images that are superior in some sense, compared with those reconstructed by analytic methods such as FORE+FBP. Our contrary results reaffirm the need for caution when attempting to predict diagnostic utility from subjective judgments of the image quality. One possible explanation of the poorer performance of FORE+OSEM is that the rebinned projection data no longer have Poisson characteristics, so that the statistical model of the OSEM algorithm is no longer valid. The violation of this assumption may result in artifacts that reduce the observer ability to detect and localize targets. In addition, the postreconstruction smoothing with a 10- to 12-mm 3D gaussian filter may well increase false detection reports of 1-cm targets because the smoothing introduces 3D noise correlations that produce "lumps" similar to the expected size of those targets.

It should be noted that, although the FORE+OSEM algorithm had significantly worse detection performance

TABLE 5
Areas Under Fitted AFROC Curves for Targets Located in Lungs

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	A_L	0.64	0.78	0.68	0.49	
OSEM	A_L	0.55	0.58	0.53		0.70
AWOSEM	A_L	0.73	0.76		0.72	0.84
FBP vs. OSEM	*P*	0.042	0.001	0.011		
AWOSEM vs. FBP	*P*	0.033	0.749		<0.001	
AWOSEM vs. OSEM	*P*	<0.001	0.025			0.035

Obs. = observer; A_L = area below AFROC curve; *P* = probability value of *z*-score test.

than FORE+FBP and FORE+AWOSEM (Table 8), the CNRs were similar for the 3 algorithms (Fig. 3). This indicates that the CNR may not be a good model for predicting human detection performance for whole-body PET imaging for different algorithms. The channelized Hotelling observer (*37*) may offer a more appropriate estimation of the trade-off between correct and false reports for human observers and, thus, a better figure of merit than the CNR (or NPWMF) to guide the optimal choice of regularizing parameters.

From the observer studies, the maximum fraction-found for the FORE+OSEM algorithm in the liver was of 0.65 instead of the approximate value of 0.9 expected from the calibration study. These differences may be attributed to the different reconstruction parameters used in the calibration study and in the comparison study. The calibration study was based on images reconstructed with the FORE+OSEM algorithm using 16 subsets, 2 iterations, and a 7-mm FWHM gaussian smoothing. In the comparison study, we used 16 subsets, 4 iterations, and a smoothing of 12-mm FWHM based on the results of the contrast-to-noise optimization. These differences may have limited the maximum performance with liver targets in the OSEM images, but the maximum fraction-found exceeded 0.9 for both of the other 2 reconstruction algorithms (FORE+AWOSEM and FORE+FBP). The levels of target contrast in the lungs and soft tissues successfully varied observer performance over a wide range for all 3 algorithms.

The relatively large number of iterations (*4*) and subsets (*16*) for the FORE+OSEM and FORE+AWOSEM algorithms was chosen so that regularization would be controlled by the kernel width of the postreconstruction 3D gaussian smoothing filter. These algorithms are typically regularized by a combination of controlling the number of iterations and subsets and the postreconstruction gaussian smoothing. It has been heuristically observed by ourselves and others that, with a small number of iterations, the resolution or contrast recovery depends on the true image value and that this undesirable coupling reduces with increasing iterations. To avoid this coupling we used a large number of iterations and controlled the image smoothness with the width of the 3D postreconstruction gaussian kernel. In a separate test using the same whole-body phantom we verified that the pixel SD and the contrast recovery did not change significantly with small changes in the number of iterations for both FORE+OSEM and FORE+AWOSEM (data not shown).

Results from this study are in good agreement with results from previous comparisons**,** demonstrating an improvement in SNR with FORE+AWOSEM as compared with FORE+OSEM (*16*,*20*), and they indicate the usefulness of the AWOSEM for the specific diagnostic task studied here. We also note that detection performance with FORE+FBP was actually better than or equivalent to that with FORE+AWOSEM in the liver, as indicated on Figure 8E. Barrett et al. (*38*) showed that the variance in images

SOCIETY OF NUCLEAR MEDICINE

TABLE 6
Areas Under Fitted AFROC Curves for Targets Located in Liver

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	A_L	0.52	0.53	0.59	0.37	
OSEM	A_L	0.28	0.28	0.25		0.28
AWOSEM	A_L	0.51	0.48		0.42	0.56
FBP vs. OSEM	*P*	<0.001	<0.001	<0.001		
AWOSEM vs. FBP	*P*	0.899	0.241		0.323	
AWOSEM vs. OSEM	*P*	<0.001	<0.001			<0.001

Obs. = observer; A_L = area below AFROC curve; *P* = probability value of *z*-score test.

TABLE 7
Areas Under Fitted AFROC Curves for Targets Located in Other Soft Tissues

Algorithm	Parameter	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
FBP	A_L	0.49	0.60	0.51	0.35	
OSEM	A_L	0.44	0.45	0.41		0.47
AWOSEM	A_L	0.61	0.71		0.56	0.63
FBP vs. OSEM	*P*	0.238	<0.001	0.039		
AWOSEM vs. FBP	*P*	0.030	0.025		<0.001	
AWOSEM vs. OSEM	*P*	<0.001	<0.001			<0.001

Obs. = observer; A_L = area below AFROC curve; *P* = probability value of *z*-score test.

reconstructed with the ML-EM is related to the local signal intensity, in contrast to FBP reconstructions, whose variance is dominated by the overall photon attenuation at each point. We thus expect significantly increased variance in the liver, relative to the lungs, for both AWOSEM and OSEM compared with FBP. This may explain the relatively better performance of the analytic reconstruction algorithm in the liver and also why the incorrect statistical model of OSEM has its most deleterious effects there.

The error bars and estimated *P* values in Tables 2–4 are based on an assumption that the fraction-found could be described by a binomial model. These values are thus minimal estimates, which should be considered when interpreting the results of the *z*-score tests of statistical significance of the difference between the 3 reconstruction strategies. The binomial model assumes that the underlying probability of a success remains the same for all N samples (i.e., no dependence on target position within organ). If that probability is itself a random variable, however, then the true SE will be larger than the binomial calculated value. This would suggest using a lower null-hypothesis probability in testing for significance with fraction-found. This does not, however, affect the consistency that was found between the rank ordering achieved with both types of performance measurements, the fraction-found and the area below the AFROC curve (Table 8).

This study chose to evaluate observer detection performance in the clinical setting of whole-body PET image interpretation. A recent study by Wells et al. (*39*) suggests that the mode of image display may affect the observers' ability to detect small lesions in thoracic SPECT scans. In particular, they showed a reduced significance of the differences in performance between different reconstruction strategies in going from a single 2D display to multiple contiguous 2D images. The authors hypothesized that this is the result of noise correlations being introduced between transverse image planes by 3D image reconstruction methods. A full volumetric display might magnify the effects seen between different reconstruction strategies, over those measured in an ROC study based on planar displays. Such effects could depend on the 3D noise correlations and artifacts that may be introduced by standard data correction and image reconstruction techniques.

In developing a method of assessing observer detection performance using a clinical volumetric display, we chose to study a multiple-target detection task. In addition to the advantage of approximating the clinical task more closely, this multiple-target task substantially increased the target sample sizes without increasing the number of image volumes that had to be read by observers. This is an important consideration, because the reading of volumetric displays requires large amounts of observer time estimated to be about 5–15 min per volume, depending on the observer. For the same reason, nontarget images were not included in our study. This study design did not permit a conventional ROC analysis of the observer data, but it was compatible with an AFROC analysis.

In this study there were exactly 7 targets per volume, which was known by the observers, although an unknown random number of targets per volume would have been more realistic.

SOCIETY OF NUCLEAR MEDICINE

TABLE 8
Summary of Tables 2–7 Averaged Over All Observers and All Regions

	FF				A_L			
Algorithm	Lungs	Liver	Soft tissue	Average over all regions	Lungs	Liver	Soft tissue	Average over all regions
FBP	0.63	0.60	0.50	0.58	0.65	0.50	0.49	0.55
OSEM	0.56	0.37	0.47	0.47	0.59	0.27	0.44	0.43
AWOSEM	0.69	0.60	0.62	0.64	0.76	0.49	0.63	0.63

FF = fraction of actual targets found; A_L = area under fitted AFROC curves.

The use of a fixed number of targets per volume, however, forced all observers to adopt an identical criterion for the number of suspicious findings (possible targets) reported per volume. We deliberately simplified the clinical detection task to control variables unrelated to an observer's ability to find and report target lesions on the images reconstructed with different algorithms. The number of targets per organ (lungs, liver, soft tissues) did vary randomly from 1 volume to another, however, and observers did not know the precise number within each organ. Although we did not simulate all details of the clinical interpretation situation, we believe that the use of even a fixed number of multiple targets, and using the same display software as that used in practice, is a significant step in that direction.

To justify the use of nonphysician observers, we note that the effects of interest were how the human observer's ability to detect simple focal densities (e.g., tumor nodules) varied with using different reconstruction algorithms. Although experienced physician observers perform realistic clinical tasks at much higher levels of accuracy than nonphysicians, for simple detection tasks it has been our experience that physician and experienced nonphysician observers have similar performance, although the physicians typically complete the readings more efficiently. This was the case in the current study.

To justify the required pooling of data from volumes with multiple targets and multiple reports, we checked for the stationarity of observer performance, as assumed by the model proposed by Swensson (*28*) for AFROC target detection and localization. This model assumes that the criteria an observer uses for reporting and assigning ratings to findings remains stationary across all multiple choices made during an image interpretation. This assumption was tested by plotting the mean number of an observer's correct and false reports, as a function of the actual number of targets (0–5) in the organ considered (data not presented). Those results generally supported the hypothesis of stationary observer performance. The mean numbers of correct reports increased in close proportion to the number of actual targets, whereas the false reports either remained nearly constant (in the liver and soft tissues) or slowly declined with an increasing number of targets (in the lungs). Further investigations are needed, however, to explore the applications of AFROC methodology in whole-body PET volume imaging. The results obtained by the AFROC analysis were similar to those for the nonparametric fraction of actual targets found and reported (fraction-found), as shown in Table 8, which supports the validity of this AFROC approach.

SOCIETY OF NUCLEAR MEDICINE

To explain the differences in shape between the AFROC curves shown here and classical ROC curves, we first note that for ROC studies the CORROC binormal fitting software uses a statistical model that assumes the observer has rated normally distributed samples drawn from 2 different populations of cases. When this binormal model is applied to fit ROC data, the 2 presumed types are (a) abnormal cases (usually containing 1 single lesion) and (b) normal cases (containing no lesion). The ROC curve measures the fraction of abnormal cases rated above a given threshold (true-positive fraction) against the correspondingly rated fraction of normal cases (false-positive fraction). These measured true-positive rates almost always exceed their corresponding false-positive rates, so that fitted ROC curves usually (but not always) lie above the chance-performance diagonal. The area below such a fitted ROC curve is >0.5 because (on average) a sampled abnormal case is likely to be rated higher than a sampled normal case.

When that same binormal model is applied to fit AFROC rating data, the 2 (presumed normal) populations are assumed to reflect an observer's degree of suspicion for (a) the image locations of actual targets and (b) the location of the maximally suspicious nontarget (normal) finding on the image(s) from each sampled case. The AFROC curve plots the measured fraction of actual targets reported at each rating criterion against the fraction of cases in which the most-suspicious nontarget finding also exceeded that same rating criterion (The latter fraction would correspond to the false-positive rate for cases that contained no lesions.). Fitted AFROC curves may have a wider range of shapes than fitted ROC curves, and the AFROC curve for a weak target may lie substantially below the diagonal. The area below the AFROC curve gives the probability that a detected (true) target will be rated higher than the most-suspicious normal finding, which may be <0.5 for a low-contrast target (as in our data). In summary, although the same binormal statistical model and fitting software can be applied to both the AFROC and the ROC curves, the 2 types of measured data have different empiric constraints. The fitted ROC curve is usually constrained by the data to lie above the chance-diagonal line, whereas the locus and shape of a fitted AFROC curve depends entirely on the target's strength or conspicuity.

Finally, we note that the absolute observer detection performances measured in this study are valid for the specific acquisition parameters, patient size, PET scanner geometry, and level of statistical noise used in this study. We reproduced standard clinical settings, but for substantial changes of scanner design, acquisition protocol, or patient size, these results may need to be reevaluated.

CONCLUSION

This study measured observer detection performance to evaluate the diagnostic impact of 3 reconstruction strategies currently used in 3D PET whole-body oncology, the analytic FORE+FBP algorithm and the 2 iterative strategies, FORE+OSEM and FORE+AWOSEM. The observers used a volumetric display of the PET whole-body data to replicate the clinical diagnostic setting.

This initial investigation shows that a methodology based on volumetric display of the images is a promising tool for performing comparison studies to design optimal whole-body PET acquisition and reconstruction protocols. In this study, we performed both an AFROC and a simple fraction-found analysis to assess changes in human observer detection performances across the images produced by 3 different reconstruction algorithms currently used for 3D whole-body clinical protocols in PET. Results show that the FORE+AWOSEM algorithm produced the best overall target detection and localization, compared with FORE+FBP and FORE+OSEM, although the performance was region specific and FORE+FBP was similar to FORE+AWOSEM for liver targets. The FORE+OSEM produced significantly worse target detection performance, especially in the liver. We speculate that this is the result of using an incorrect statistical model in FORE+OSEM and that the incorporation of AW in AWOSEM (by far the dominant data correction term) compensates for this model inaccuracy.

ACKNOWLEDGMENTS

We thank Laurence Laurier, Michèle Ottaviani, and Maria Joao Ribeiro who helped with the calibration studies; Christian Michel and Michel Defrise for advice and help with the reconstruction software; Ben Tsui and Karen Lacroix for help with the MCAT phantom; and Michael A. King, Bernard Bendriem, Harrison H. Barrett, and Irène Buvat for helpful discussions. We note the passing of our colleague Richard G. Swensson and express our appreciation for the opportunity to work with him. This work was supported by a grant from Siemens-France and by the National Cancer Institute (grant CA-74135). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

REFERENCES

1. Rigo P, Paulis P, Kaschten BJ, et al. Oncological applications of positron emission tomography with fluorine-18 fluorodeoxyglucose. *Eur J Nucl Med.* 1996;23:1641–1674.
2. Dahlbom M, Hoffman EJ, Hoh CK, et al. Whole-body positron emission tomography: Part I. methods and performance characteristics. *J Nucl Med.* 1992;33:1191–1199.
3. Leahy R, Qi J. Statistical approaches in quantitative positron emission tomography. *Stat Comput.* 2000;10:147–165.
4. Ollinger JM, Fessler JA. Positron emission tomography. *IEEE Signal Process.* 1997;14:43–55.
5. Leahy R, Byrne C. Recent developments in iterative image reconstruction for PET and SPECT. *IEEE Trans Med Imaging.* 2000;19:257–260.
6. Hudson H, Larkin R. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans Med Imaging.* 1994;13:601–609.
7. Meikle SR, Hutton BF, Bailey DL, Hooper PK, Fulham MJ. Accelerated EM reconstruction in total-body PET: potential for improving tumor detectability. *Phys Med Biol.* 1994;39:1689–1704.
8. Cherry SR, Dahlborn M, Hoffman EJ. High sensitivity, total body PET scanning, using 3D data acquisition and reconstruction. *IEEE Trans Nucl Sci.* 1992;39:1088–1092.
9. Cutler PD, Xu M. Strategies to improve 3D whole-body PET image reconstruction. *Phys Med Biol.* 1996;41:1453–1467.
10. Kinahan PE, Rogers JG. Analytic 3D image reconstruction using all detected events. *IEEE Trans Nucl Sci.* 1989;36:964–968.
11. Ollinger JM, Goggin AS. Maximum likelihood reconstruction in fully 3D PET via the SAGE algorithm. *Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference*. Anaheim, CA; 1996:1594–1598.
12. Kinahan PE, Matej S, Karp JS, Herman GT, Lewitt RM. A comparison of transform and iterative reconstruction techniques for a volume-imaging PET scanner with a large axial acceptance angle. *IEEE Trans Nucl Sci.* 1995;42:2281–2288.
13. Matej S, Herman GT, Narayan TK, Furuie SS, Lewitt RM, Kinahan PE. Evaluation of task-oriented performance of several fully 3-D PET reconstruction algorithms. *Phys Med Biol.* 1994;39:355–367.
14. Qi J, Leahy RM, Hsu C, Farquhar TH, Cherry SR. Fully 3D Bayesian image reconstruction for the ECAT EXAT HR+. *IEEE Trans Nucl Sci.* 1998;45:1096–1103.
15. Kinahan PE, Michel C, Defrise M, et al. Fast iterative image reconstruction of 3D PET data. *Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference*. Anaheim, CA; 1996:1918–1922.
16. Comtat C, Kinahan PE, Defrise M, Michel C, Townsend DW. Fast reconstruction of 3D PET data with accurate statistical modelling. *IEEE Trans Nucl Sci.* 1998;45:1083–1089.
17. Defrise M, Kinahan PE, Townsend DW, Michel C, Sibomana M, Newport DF. Exact and approximate rebinning algorithms for 3-D PET data. *IEEE Trans Med Imaging.* 1997;16:145–158.
18. Liu X, Comtat C, Michel C, Kinahan PE, Defrise M, Townsend DW. Comparison of 3D reconstruction with 3D-OSEM and with FORE+OSEM for PET. *IEEE Trans Med Imaging.* 2001;20:804–814.
19. Hebert T, Leahy RM. Fast methods for including attenuation correction in the EM algorithm. *IEEE Trans Nucl Sci.* 1990;37:754–758.
20. Michel C, Sibomana M, Bol A, et al. Preserving Poisson characteristics of PET data with weighted OSEM reconstruction. *Proceedings of the 1998 IEEE Nuclear Science Symposium and Medical Imaging Conference* [CD-ROM]. Toronto, Ontario, Canada: IEEE; 1998.
21. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol.* 1986;21:720–733.
22. Wester C, Judy PF, Polger M, Swensson R, Feldman U, Seltzer SE. Influence of visual distractors on detectability of liver nodules on contrast-enhanced spiral computed tomography scans. *Acad Radiol.* 1997;4:335–342.
23. Swensson RG, Judy PF, Wester C, Seltzer SE. Nodule polarity effects on detection and localization performance in liver CT images. *Proceedings of the International Society for Optical Engineering: Medical Imaging Conference*. San Diego, CA: 1997;3036:85–93.
24. Llacer J, Veklerov E, Baxter LR, et al. Results of clinical receiver operating characteristics study comparing filtered backprojection and maximum likelihood estimator images in FDG PET studies. *J Nucl Med.* 1993;34:1198–1203.
25. Li J, Jaszczack RJ, Turkington TG, et al. An evaluation of lesion detectability with cone-beam, fanbeam, and parallel-beam collimation in SPECT by continuous ROC study. *J Nucl Med.* 1994;35:135–140.
26. De Vries DJ, King MA, Soares EJ, Tsui BMW. Evaluation of the effect of scatter correction on lesion detection in hepatic SPECT imaging. *IEEE Trans Nucl Sci.* 1997;44:1733–1740.
27. Farquhar TH, Llacer J, Hoh CK, et al. ROC and localization ROC analyses of lesion detection in whole-body FDG PET: effects of acquisition mode, attenuation correction and reconstruction algorithm. *J Nucl Med.* 1999;40:2043–2052.
28. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys.* 1996;23:1709–1725.
29. Chakraborty DP, Winter LH. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology.* 1990;174:873–881.
30. Comtat C, Kinahan PE, Defrise M, Michel C, Townsend DW. Simulating whole-body PET scanning with rapid analytical methods. *Proceedings of the 1999 IEEE Nuclear Science Symposium and Medical Imaging Conference*. Seattle, WA; 1999:1260–1264.
31. Lacroix KJ. *Evaluation of an Attenuation Compensation Method with Respect to Defect Detection in Tc-99m-Sestamibi Myocardial SPECT* [PhD thesis]. Chapel Hill, NC: The University of North Carolina at Chapel Hill; 1997.
32. Brix G, Zaers J, Adam LE, et al. Performance evaluation of a whole-body PET scanner using the NEMA protocol. *J Nucl Med.* 1997;38:1614–1623.
33. Defrise M, Kinahan P. Data acquisition and image reconstruction for 3D PET. In: Townsend DW, Bendriem B, eds. *The Theory and Practice of 3D PET*. Dordrecht, The Netherlands: Kluwer Academic; 1998:11–54.
34. Wagner RF, Brown DG. Unified SNR analysis of medical imaging systems. *Phys Med Biol.* 1985;30:489–518.
35. Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. *Proceedings of the Information Processing in Medical Imaging*. The Hague, The Netherlands; 1984: 432–445.
36. Loo L-ND, Doi K, Metz CE. A comparison of physical image quality indices and observer performance in the radiographic detection of nylon beads. *Phys Med Biol.* 1984;29:837–856.
37. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. *Proc Natl Acad Sci USA.* 1993;90:9758–9765.
38. Barrett HH, Wilson DW, Tsui BMW. Noise properties of the EM algorithm: I. theory. *Phys Med Biol.* 1994;39:833–846.
39. Wells RG, King MA, Gifford HC, Pretorius PH. Single-slice versus multi-slice display for human-observer lesion-detection studies. *IEEE Trans Nucl Sci.* 2000; 47:1037–1044.

SOCIETY OF NUCLEAR MEDICINE