Results of a Clinical Receiver Operating Characteristic Study Comparing Filtered Backprojection and Maximum Likelihood Estimator Images in FDG PET Studies

J. Llacer, E. Veklerov, L.R. Baxter, S.T. Grafton, L.K. Griffeth, R.A. Hawkins, C.K. Hoh, J.C. Mazziotta, E.J. Hoffman and C.E. Metz

Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, California; Department of Neurology and Division of Nuclear Medicine, UCLA School of Medicine and Laboratory of Nuclear Medicine, Los Angeles, California; Departments of Neurology and Radiology, University of Southern California, Los Angeles, California; Mallinckrodt Institute of Radiology, Washington University Medical Center, St. Louis, Missouri; and Department of Radiology, University of Chicago, Chicago, Illinois

The results of a receiver operator characteristic (ROC) study comparing maximum likelihood estimator (MLE) reconstructions of human FDG PET brain scan data to filtered backprojection reconstructions of the same data are reported. The purpose of the study was to determine whether MLE reconstructions would result in higher detectability of small focal lesions introduced artificially into otherwise normal scan data. One physician assisted in defining the location and intensity of the lesions and five physicians read the final images. Data from 90 datasets were used for the study. Of those, 42 were left in their original "normal" condition and 48 were modified by added lesions. All datasets were reconstructed by the two methods and submitted to the five physicians for evaluation. The results show an increase in the area under the ROC curve from approximately 0.65 for filtered backprojection to approximately 0.71 for the maximum likelihood reconstructions for four of the five observers with good statistical significance.

J Nucl Med 1993; 34:1198–1203

The maximum likelihood estimator (MLE) method of image reconstruction for emission tomography has been under study by research groups for several years because it promises lower noise and its consequent higher effective sensitivity when compared with standard filtered backprojection (FBP) methods (*1–13*). After questions related to behavior of MLE algorithms at high iteration numbers have been resolved and the effects controlled, studies indicate that MLE reconstructions of positron emission tomography (PET) data exhibit lower noise in regions of low radioisotope uptake than FBP reconstructions. The noise in regions of high isotope concentrations are comparable in both methods (*14–18*) for similar resolution. By using standard statistical techniques, our group has been able to quantify that improvement in phantoms and in real PET data from ^{18}F-2-fluoro-2-deoxyglucose (FDG) human brain studies for the case of standard (nontime-of-flight) PET. The analysis shows that the expected error in the estimation of uptake in regions of low uptake drops by approximately one-third in MLE reconstructions by comparison with FBP reconstructions of the same data (*19*).

The reduced noise in low uptake regions raises the expectation that detectability of small focal lesions in those regions would be better with MLE than with FBP reconstructions. A number of figures of merit or confidence factors have been devised for the purpose of predicting the performance of human observers in carrying out well specified tasks under controlled conditions. For the PET case, with correlated noise resulting from a nonlinear reconstruction method (MLE), and in the very complex task of detection in real PET FDG images, the state of the art is still far from being able to predict human performance. We are then left with the time-consuming but proven receiver operator characteristic (ROC) methodology to verify the correctness of our expectation. We focused on lesions in both grey and white matter of a nature and contrast level that make them borderline in detectability. Lesions that are easy to detect by FBP would also be easily detectable by MLE and those that are impossible to detect by FBP may also be undetectable by MLE. It is in borderline cases where the MLE method can be expected to yield better results.

METHODS

ROC methodology is now well established as a reliable way of statistically determining the differences in performance of medical procedures that combine human observers and technology in

Received Nov. 23, 1992; revision accepted Mar. 18, 1993.
For correspondence and reprints contact: Jorge Llacer, PhD, Bldg. 46A, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720.

medical diagnostic tasks (*20–23*). A well-defined protocol that considers a number of possible pitfalls in those studies is essential if the final results are to have statistical validity (*24,25*). We present here the procedures we followed in a very strict manner in order to ensure such statistical validity.

Data Selection and Organization

FDG PET scan data from 15 individuals were utilized for this study. All 15 were either normal volunteers or patients who yielded PET studies that were considered normal. Data for each individual consisted of 15 planes through the brain in a number of time frames adding up to totals from 30 min to more than 1 hr obtained with a CTI-831 tomograph (Knoxville, TN). Data collection started approximately 45 min after FDG injection. A preliminary study showed that in the absence of additional anatomical information (i.e., provided by planes adjacent to the one under study), physicians were not able to detect focal lesions that would have been considered to be quite obvious (based on local contrast ratios and on nonblinded comparison with the original image) to the physicists who placed them into the datasets (*25*). In order to provide that anatomical information, all images were prepared as sets of three consecutive planes, with the center plane being evaluated for presence or absence of a lesion. These sets of three images are referred to as "image sets." Ninety image sets were formed from the 225 available planes, with most of the peripheral planes being reused in different image sets (bottom plane of one set becoming top plane of another set). The central planes for all the sets were different and were never used as peripheral planes. Planes 1 and 15 of most of the patients were removed from consideration for the datasets because they often showed little grey or white matter that was suitable for our study.

Statistical power in ROC studies is maximized when approximately half of the images are normal. From the 90 image sets, we selected 42 sets at random to be left as "normals," i.e., no lesion was added to them. The remaining 48 image sets received lesions as described below. The cumulative time frame for each of the 90 image sets to be used in the ROC evaluation contained approximately 1.4 million counts in the center plane. This number was selected qualitatively as yielding images that are noisy, but not excessively so. It corresponds to a typical 5-min time frame for many of the patient studies that we have seen. The selection of noisy images is consistent with our aim of comparing MLE reconstructions with FBP reconstructions through detectability experiments in borderline cases, although it may not be a standard clinical FDG procedure. The experiment corresponds to a general situation in many other PET studies with count-limited images.

Introduction of Lesions

In addition to preparing the 1.4-million count image sets, all the available time frames for each image set were totaled to provide "high-count" image sets. These were used for the introduction of artificial lesions in order to decouple the problem of generating medically plausible lesions from that of detecting lesions in low-count datasets.

Three kinds of lesions were introduced: (1) additive lesions in grey matter, (2) subtractive lesions in grey matter and (3) additive lesions in white matter. Subtractive lesions in white matter were not considered for several reasons. First, small subtractive white matter lesions would almost certainly be considered as an effect due to noise in our images. For example, PET images of noise and resolution characteristics similar to ours would commonly be unable to delineate boundaries between white matter, where little uptake is expected, from ventricles, where no uptake is expected. Second, the applicability of PET for such lesions (especially by comparison with that of structural imaging such as computed tomography or magnetic resonance imaging) would be expected to be quite limited in routine clinical use. The choice of which dataset gets what type of lesion was done at random. From a medical point of view, the lesions corresponded to plausible cases of small focal lesions found in clinical practice. The intensity, size and local contrast were chosen so that in the high-count datasets used for that purpose, the lesions were reasonably easy to detect by experienced physicians. It was expected, however, that a range of difficulty would come naturally from the process. In approximately 30% of the cases, lesions were allowed to extend to the peripheral plane above or below the center plane to be evaluated, as would occur in practice.

Once a proposed lesion was found to be acceptable, it was transferred to the 1.4-million count image set, preserving location and relative contrast or intensity, in the following manner: A "perfect" lesion was first projected into the data space by multiplication with the response matrix of the tomograph. For additive lesions, the new counts were added to the original dataset in a Poisson fashion. For subtractive lesions, the projected counts were removed from the scan data by the thinning process which preserves their Poisson characteristics. The modified datasets were finally reconstructed by MLE and FBP for presentation to the observers. Figures 1, 2 and 3 show examples of high-count image sets, both before and after the introduction of lesions. Also shown are the FBP and MLE reconstructions of the corresponding low-count image sets (labeled ROC) with lesions for the cases with an additive lesion in grey matter, a subtractive lesion in grey matter and an additive lesion in white matter.

Reconstruction Methods

The normal and modified image sets were reconstructed by FBP using a Butterworth filter with characteristics shown in Figure 4. This filter represents an improvement over the Shepp-Logan filter used routinely for FDG images with approximately 1.4 million counts also shown in the same figure. The Butterworth filter enhances the middle frequencies and cuts off the high frequencies more strongly where there is a predominant contribution from noise. The choice of parameters for the Butterworth filter was verified by the five observers as yielding images with the optimal information for that number of counts.

All the datasets were also reconstructed by the maximum likelihood estimator with cross-validation (MLE-CV) method with a small amount of Gaussian postfiltering, yielding images with a resolution equivalent to the FBP images. The reconstruction procedure and resolution evaluation have been described in detail (*19*). All the images were reconstructed on a 128 × 128 grid of pixels with a side dimension of 0.18 cm.

Image Set Presentation

Images were submitted to the five observers in groups of 15 image sets containing a balanced set of normals and lesions as follows:

1. There were seven normal and eight lesion image sets in each group.
2. Of the eight lesions, four were in grey matter and four were in white matter.
3. The grey matter lesions were two additive and two subtractive.
4. The white matter lesions were all additive.

5. In what we call "direct" sets, three normal image sets were reconstructed by FBP and four by MLE. The image sets with lesions were reconstructed (four by FBP and four by MLE) in random assignments.
6. In the corresponding "reverse" sets, the methods of reconstruction of exactly the same datasets were reversed from the above.
7. The 15 image sets in a group were presented in random order to the observers.

There were 12 groups of 15 image sets, for a total of 180 image sets. The "direct" sets were presented first and the "reverse" sets followed. The observers knew only that approximately 50% of the images in a 15-image set group were normal and that there were no subtractive white matter lesions. The images reconstructed by MLE or by FBP were immediately obvious to the observers but posed no problem for the ROC procedure.

A training session preceded the study. The nature of the study was explained to the individual physicians. Sample images with and without lesions were shown to them and they were instructed to respond to the question: "Does the image show an abnormality?" for the center plane, with a 5-point rating scale corresponding to: 1 = definitely or almost definitely not; 2 = probably not; 3 = possibly yes; 4 = probably yes; and 5 = definitely or almost definitely yes.

The physicians had the freedom to choose any method they preferred for viewing the images, although the ability to manipulate the color or intensity scale in order to extract the maximum amount of information from the images was emphasized. They all chose to work interactively with no time limit on good quality black and white or color image display workstations. There was the possibility that physicians would give an affirmative answer in response to either an added lesion or to a normal anatomical feature that they perceived to be a lesion which would lead to an error. By designing the study as a correlated one, i.e., with each image being evaluated for both methods of reconstruction, errors would be made with nearly equal probability in both modalities with results that cancel out in the final ROC analysis.

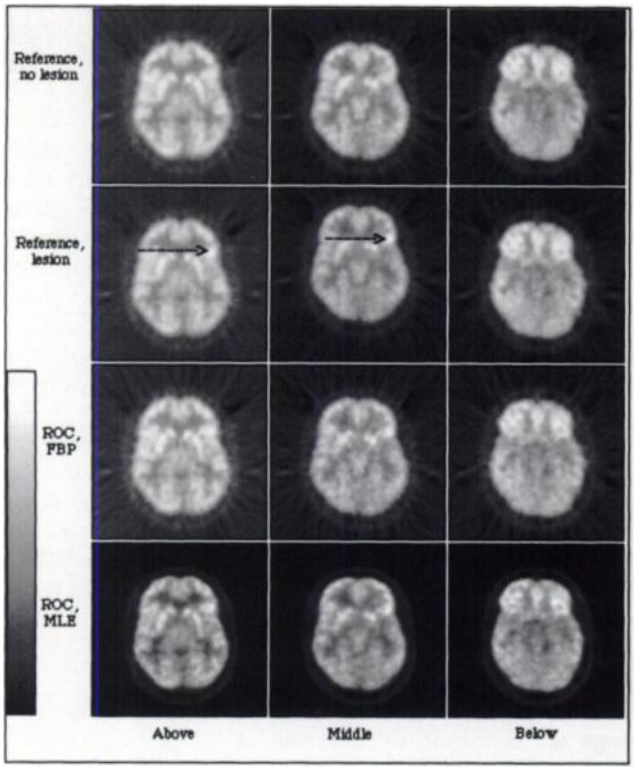

FIGURE 1. High-count (reference) images reconstructed by FBP before adding a lesion (top row), after placing an additive lesion in grey matter indicated by an arrow (second row), FBP reconstructions of a 1.4M dataset corresponding to the second row images (third row) and MLE reconstruction of the same 1.4M dataset. The images on the left and right columns are submitted together with the center image to provide anatomical information to the physician who has to rate the center image. The added lesion was extended to the plane above the center.

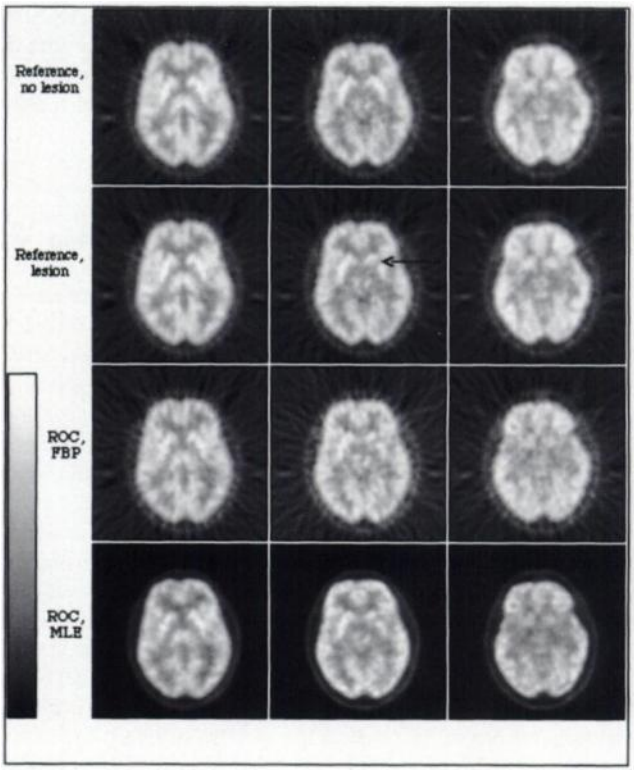

FIGURE 2. An example similar to Figure 2, with the lesion being subtracted in grey matter indicated by an arrow.

RESULTS OF THE STUDY

ROC Curves

The ROC methodology used for this work is based on fitting the data to bi-normal distributions, plotting the data in a ROC curve and evaluating the significance of the results, bearing in mind the correlated nature of the study (i.e., each image set was evaluated by each observer in the two modalities being compared). The underlying assumption for the bi-normal distribution is that an observer faced with evaluating an image for the presence or absence of some characteristic will give numerical results that are normally distributed about some mean for positive cases and normally distributed about a different mean for negative cases. A detailed discussion of the basic process of ROC analysis has been given previously (*21*). The significance of the results was evaluated by using the CORROC2 program developed at the University of Chicago for correlated data, based on the work of Metz, Wang and Kronman (*23,26*). The resulting five pairs of ROC curves are shown

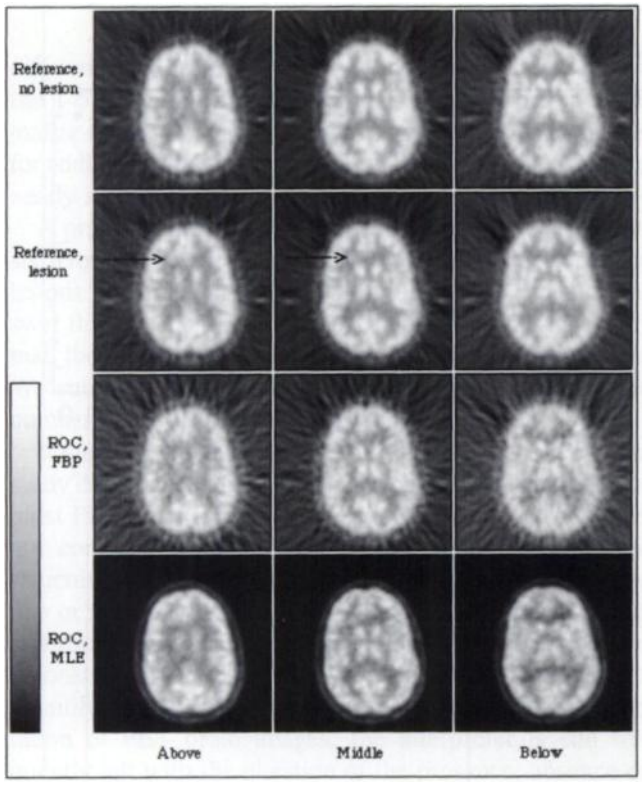

FIGURE 3. An example similar to Figure 2, with the lesion being added in white matter indicated by an arrow. This lesion extends to the plane above center.

in Figures 5A–5E (one pair for each of the observers in the study). The solid or broken lines correspond to the binormal fitting, whereas the dots correspond to the actual data points obtained from the data. As indicated in Figure 5A, the solid line represents the MLE results.

Table 1 shows the area A_Z under the ROC curve for each of the observers for the two methods of reconstruction and the difference between the two. In the first four cases, the values of A_Z for the MLE are larger than those for the FBP by approximately 10% or more. For physician p5, no significant difference is observed between the two methods, but it must be noted that the actual A_Z values are barely above 0.5 which indicate a performance not too different from chance for both reconstruction methods. This exercise is very different from a diagnostic procedure in nuclear medicine so the above result for p5 must be taken in the proper context.

Statistical Significance

The ROC curves for the cases of p1–p4 are not sufficiently separated to establish statistical significance individually. We estimate that approximately 300 independent datasets (600 different image sets) would have been needed for that purpose instead of the 90 datasets used. Taken collectively, however, the probability that four observers would find those similar results by chance is very small, as will be shown below. A collective statistical analysis was performed by using Student's t-test for paired data for testing the hypothesis that the two methods of reconstruction yield images with equal detectability. The test was carried out in two ways: (1) assuming that physicians p1–p4 are representative of the nuclear medicine physician population with PET experience and (2) assuming that the five physicians (including p5) are representative of that population. This assumption for the second case is, however, not strictly correct; the t-test assumes that the data being analyzed are normally (Gaussian) distributed about a mean we wish to estimate. It implies that the individual participants in the test have some characteristic in common that justifies the assumption of normality of the data with a unique mean. The areas A_Z under the ROC curves for p5 were very near 0.5, indicating a performance for the ROC task which was very near random for both methods of reconstruction. The requirement that physicians p1–p5 have similar characteristics for the performance of the ROC task was not met and therefore the assumption of normality of the data cannot be supported. Nevertheless, we have evaluated the data with and without the inclusion of p5 to show that in the worst case the hypothesis being tested can be rejected.

Case 1: Taking Only the Data from Observers p1–p4. The t-test establishes that ΔA_Z (the true difference in the areas A_Z) observed by four physicians taken from a group of physicians with characteristics similar to the p1–p4 group, which replicates the experiment we have carried out with the same image sets, is bound by the 95% confidence interval:

$$0.0685 - 0.0017 < \Delta A_Z < 0.0685 + 0.0017,$$

in favor of the MLE method.

We then define the p level as the probability of finding an absolute difference in area as large or larger than the one observed by a group of four physicians with characteristics similar to the p1–p4 group, replicating the same experiment if the two methods of reconstruction were equally effective in demonstrating lesions. The p level is <0.00001, i.e., the hypothesis that the two methods are equally effective can be rejected.

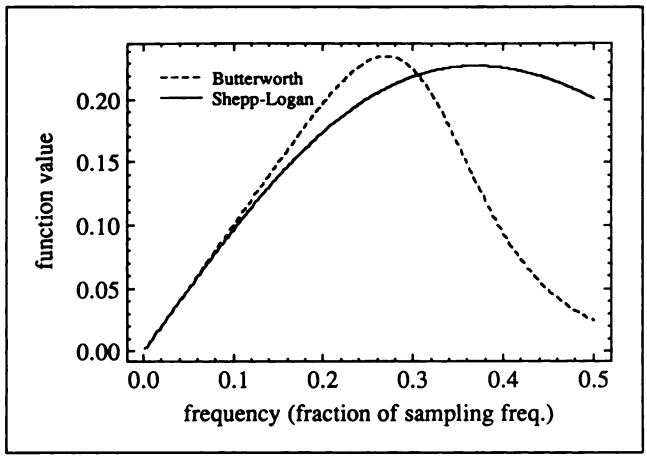

FIGURE 4. Butterworth filter used in the FBP reconstructions shown with the Shepp-Logan filter used routinely for the reconstruction of 1.4M count datasets in the UCLA Dept. of Nuclear Medicine.

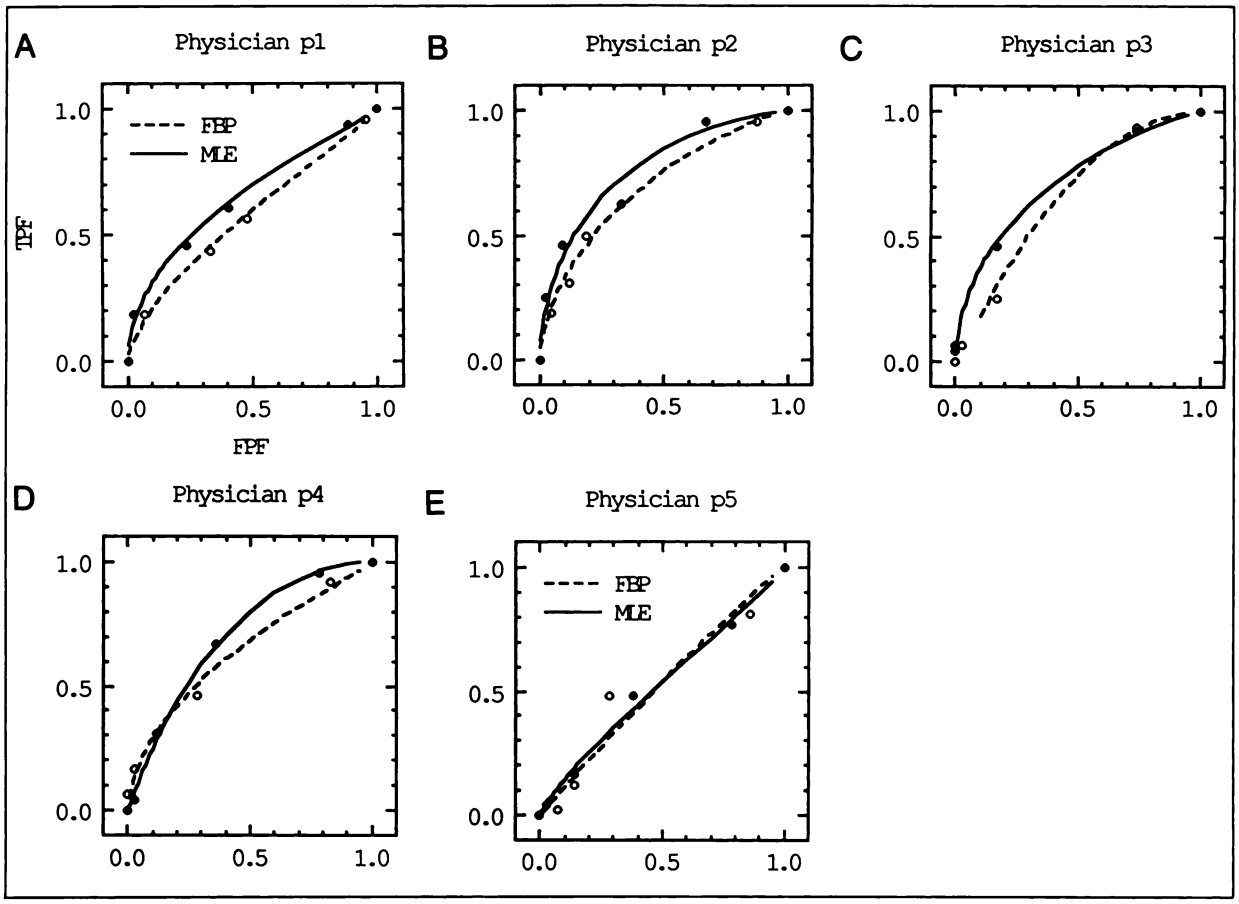

FIGURE 5. (A–E) ROC curves for the five individual physicians. Solid and dashed lines correspond to a bi-normal fitting of the measured data, which are represented by circles. The MLE results are those shown by the solid lines. The abcissa corresponds to the false-positive fraction (FPF, as shown in A), i.e., the fraction of negative cases diagnosed as positive. The ordinate corresponds to the true-positive fraction (TPF in A), i.e., the fraction of positive correctly diagnosed as positive. An excellent diagnosis would correspond to curves with high values of TPF at very low values of FPF. Such curves would have an area A_Z near 1.0.

Case 2: Taking the Data from All Five Observers. If we assume that the five observers (p1–p5) are truly representative of the population of experienced PET nuclear medicine physicians, then the true difference in the areas A_Z for any five physicians replicating the same experiment is bound between:

$$0.055 - 0.038 < \Delta A_Z < 0.055 + 0.038,$$

in favor of the MLE method. The p level obtained is <0.20, i.e., the hypothesis that the two methods of reconstruction are equally effective can be rejected with at least 80% probability of being correct.

TABLE 1
Comparison of ROC Curve Results

Physician	Area under ROC curve, A_Z		
	MLE	FBP	Difference
p1	0.657	0.576	0.081
p2	0.773	0.699	0.074
p3	0.722	0.663	0.059
p4	0.706	0.646	0.060
p5	0.527	0.526	0.001

DISCUSSION

The above results establish the fact that the MLE method of image reconstruction is better than the FBP for the task of detecting small focal lesions at the threshold of detectability. We note that our intended goal of evaluating the difference between the methods of reconstruction in borderline cases has been fulfilled. Except for observer p5, the area A_Z for the FBP ranges between 0.576–0.699, while for MLE it ranges between 0.657–0.773. It is generally accepted that values for A_Z in the vicinity of 0.65–0.75 result in useful sensitivity to demonstrate differences between procedures. In our case, lower values would indicate lesions that are very difficult for that group of physicians to detect while substantially larger numbers would probably indicate that the lesions are too easily detected. The lesions placed on the high-count images were all judged to be detectable by an experienced observer, though some were relatively subtle. When transferred to the low-count images, these lesions became considerably more difficult to detect because of statistical noise. By a simple analysis of the responses of different observers to different lesions, we have established qualitatively that the

difference in the ROC curves is due to higher ratings (on the 1–5 scale) in MLE images for additive lesions in white matter and for subtractive lesions in grey matter. The results for additive lesions in grey matter and in normal images were nearly identical for both methods of reconstruction.

A practical question arises as a result of the above analysis: How often are physicians faced with having to detect lesions in borderline cases? Unfortunately, we cannot answer that question. Every time a study is diagnosed as normal, the physician may have been facing a borderline case. We submit that MLE reconstructions should decrease the number of false-negative readings of PET studies.

It might also be argued that the situation devised for this study differs substantially from the clinical interpretation of most PET brain imaging studies in which the PET images are compared directly with corresponding images from structural imaging techniques such as computed tomography or magnetic resonance imaging. To a large extent, this assertion is true and the influence of potentially improved reconstruction methods may be difficult or impossible to quantify in such cases. However, in the clinical interpretation of PET brain images, the interpreter is still frequently left with the question of the presence, absence or level of activity within a discrete focus. For example, a common clinical application of FDG-PET brain imaging is the distinction between radiation necrosis and recurrent brain tumor based on the level and/or spatial distribution of FDG accumulation. This dilemma frequently entails the assessment of activity levels within a small lesion or a small portion of a structurally heterogeneous lesion. The current results suggest that the MLE reconstruction technique may offer clinically significant advantages in such situations.

One of the objections often raised to the use of MLE reconstructions in a clinical setting is the requirement for high computational times. The MLE images shown in Figures 1, 2 and 3 were obtained in approximately 10 min per plane using a readily available Hewlett-Packard-730 workstation and our software, which is not particularly optimized for speed. The advent of powerful, reasonably priced multiprocessors will increase the speed of computation significantly. The performance of a Hyper-Cube structure of 8 iPSC/860 processors was evaluated recently by our group and was found to permit reconstruction of one plane in approximately 30 sec, bringing the MLE method in the range of clinical utility. With rapidly decreasing prices for such processors, clinical trials of the MLE algorithm are now indicated.

ACKNOWLEDGMENTS

The above work has been supported, in part, by grants from the National Institutes of Health, CA-39531, NS-15654 and MH-37916, and by the Director, Office of Energy Research, Office of Health and Environmental Research, Physical and Technological Division, of the U.S. Department of Energy under contract nos. DE-AC03-76SF00098 and DE-FC03-87ER-60615. Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

REFERENCES

1. Shepp LA, Vardi Y. Maximum likelihood reconstruction for emission tomography. *IEEE Trans Med Imaging* 1982;1:113–121.
2. Shepp LA, Vardi Y, Ra JB, Hilal SK, Cho ZH. Maximum likelihood PET with real data. *IEEE Trans Nucl Sci* 1984;31:910–913.
3. Kaufman L. Implementing and accelerating the EM algorithm for positron emission tomography. *IEEE Trans Med Imaging* 1987;6:37–51.
4. Snyder DL, Miller MI, Thomas LJ, Politte DG. Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. *IEEE Trans Med Imaging* 1987;6:228–238.
5. Politte DG, Snyder DS. Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron emission tomography. *IEEE Trans Med Imaging* 1991;10:82–89.
6. Floyd CE, Jaszczak RJ, Coleman RE. Inverse Monte Carlo: a unified reconstruction algorithm for SPECT. *IEEE Trans Nucl Sci* 1985;32:779–785.
7. Llacer J, Veklerov E, Hoffman EJ. On the convergence of the maximum likelihood estimator method of tomographic image reconstruction. *SPIE Medical Imaging* 1987;767:70–76.
8. Llacer J, Veklerov E. The maximum likelihood estimator method of image reconstruction: its fundamental characteristics and their origin. In: de Graaf CN, Viergever MA, eds. *Information processing in medical imaging*. New York: Plenum; 1988:201–216.
9. Veklerov E, Llacer J. Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Trans Med Imaging* 1987;6:313–319.
10. Veklerov E, Llacer J, Hoffman EJ. MLE reconstruction of a brain phantom using a Monte Carlo transition matrix and a statistical stopping rule. *IEEE Trans Nucl Sci* 1988;35:603–607.
11. Llacer J, Veklerov E. Feasible images and practical stopping rules in iterative image reconstruction. *IEEE Trans Med Imaging* 1989;8:186–193.
12. Llacer J, Veklerov E, Nunez J. Stopping rules, Bayesian reconstructions and Sieves. In: Ortendahl DA, Llacer J, eds. *Information processing in medical imaging*. New York: Wiley-Liss; 1991:81–93.
13. Herman GT, Odhner D. Performance evaluation of an iterative image reconstruction algorithm for positron emission tomography. *IEEE Trans Med Imaging* 1991;10:336–346.
14. Rosenqvist G, Dahlbom M, Eriksson L, Bohm C, Blomqvist G. Quantitation of PET data with the EM reconstruction technique. *IEEE Trans Nucl Sci* 1989;36:1113–1116.
15. Tanaka E. A fast reconstruction algorithm for stationary positron emission tomography based on a modified EM algorithm. *IEEE Trans Med Imaging* 1987;6:98–104.
16. Holte S, Schmidlin P, Linden A, Rosenqvist G, Eriksson L. Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems. *IEEE Trans Nucl Sci* 1990;37:629–635.
17. Llacer J, Bajamonde AC. Characteristics of feasible images obtained from real PET data by MLE, Bayesian and sieve methods. *SPIE proceedings, digital image synthesis and inverse optics* 1990;1351:300–312.
18. Liow JS, Strother SC. Practical tradeoffs between noise, quantization and number of iterations for maximum likelihood-based reconstructions. *IEEE Trans Med Imaging* 1991;10:563–571.
19. Llacer J, Veklerov E, Coakley KJ, Hoffman EJ, Nunez J. Statistical analysis of maximum likelihood estimator images of human brain FDG PET studies. *IEEE Trans Med Imaging* 1993:in press.
20. Swets JA, Picket RM. *Evaluation of diagnostic systems*. New York: Academic Press, 1992.
21. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;4:283–298.
22. Metz CE. ROC Methodology in radiologic imaging. *Invest Radiol* 1986;21:720–733.
23. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989;24:234–245.
24. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jacknife method. *Invest Radiol* 1992;27:723–731.
25. Llacer J, Veklerov E, Nolan D, et al. ROC study of maximum likelihood estimator human brain image reconstructions in PET clinical practice: a progress report. *Conf Record of the 1990 IEEE Nucl Sci Symp* 1990;2:1556–1561.
26. Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, ed. *Information processing in medical imaging*. The Hague, The Netherlands: Martinus Nijhoff; 1984:432–445.