LBL-30737
Part II of II
Preprint

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Engineering Division

Submitted to IEEE Transactions on Medical Imaging

Optimization of Maximum Likelihood Estimator Images for PET: II. Statistical Analysis of Human Brain FDG Studies

J. Llacer, E. Veklerov, K.J. Coakley, E.J. Hoffman, and J. Nunez

December 1991

U. C. Lawrence Berkeley Laboratory
Library, Berkeley

FOR REFERENCE

Not to be taken from this room

LBL-30737
Copy 1
Bldg. 50 Library.

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-30737
Part II of II

Optimization of Maximum Likelihood Estimator Images for PET: II. Statistical Analysis of Human Brain FDG Studies

J. Llacer, Ph.D.[1], E. Veklerov, Ph.D.[1], K.J. Coakley, Ph.D.[2],
E.J. Hoffman, Ph.D.[3] and J. Nunez, Ph.D.[4]

[1]Engineering Division, Lawrence Berkeley Laboratory
University of California, Berkeley, CA 94720
[2]Statistical Engineering Division
National Institute of Standards and Technology,
[3]Dept. of Radiological Sciences, School of Medicine,
University of California, Los Angeles, and
[4]Facultat de Fisica, Universitat de Barcelona, Spain

December 15, 1991

This work has been supported, in part, by grants from the National Institutes of Health, CA-39501 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. The work of J. Nunez has been supported, in part, by the Center of Catalan Studies, Generalitat de Catalunya, Spain, and University of California, Berkeley.

LBL-30737
Part II of II

Optimization of Maximum Likelihood Estimator Images for PET: II. Statistical Analysis of Human Brain FDG Studies

J. Llacer, Ph.D.[1], E. Veklerov, Ph.D.[1], K.J. Coakley, Ph.D.[2],
E.J. Hoffman, Ph.D.[3] and J. Nunez, Ph.D.[4]

[1]Engineering Division, Lawrence Berkeley Laboratory
University of California, Berkeley, CA 94720
[2]Statistical Engineering Division
National Institute of Standards and Technology,
[3]Dept. of Radiological Sciences, School of Medicine,
University of California, Los Angeles, and
[4]Facultat de Fisica, Universitat de Barcelona, Spain

Abstract

The work presented in this paper studies the question of what are the characteristics of Maximum Likelihood Estimator (MLE) reconstructions in Positron Emission Tomography (PET) that make that method superior to the standard Filtered Backprojection (FBP). PET data of human brain Fluoro-deoxiglucose studies have been used to evaluate the comparative statistical characteristics of regional bias and expected error of the two methods. The results are that properly reconstructed MLE images are as unbiased as those obtained by FBP and that the expected error in the MLE case is approximately proportional to the square root of the number of counts in a region. In contrast, FBP reconstructions show an expected error that is high and nearly independent of the number of counts in a region. A preliminary study shows that those statistical characteristics of MLE reconstructions translate into improved lesion detectability in regions of low activity for the case of a simple detectability task carried out by non-medical observers.

I. INTRODUCTION

Since Shepp and Vardi first introduced a Maximum Likelihood Estimator (MLE) iterative procedure [1] based on the Expectation-Maximization algorithm of Dempster et al,[2] numerous workers have investigated the method and have, in general, reached the conclusion that, handled in some specific manner, it yields "better" or "less noisy" images from Emission Tomography (ET) data than conventional Filtered Backprojection (FB) reconstruction methods. It has often been felt, however, that there is a need to understand in which manner MLE images are better than FBP and that the differences in the quality of images should be quantified in statistical terms. Llacer and Bajamonde [3] were able to demonstrate some statistical characteristics of MLE reconstructions of PET data from a Hoffman brain phantom which were a first step in the direction of providing a quantification of the difference between MLE and FBP images. The principal conclusions that were reached were:

1. By chosing properly designed methods, MLE images can be sharper than FBP images without paying a significant penalty in noise,
2. Those MLE images exhibit biases of the same order as those of FBP images, and
3. The MLE images have lower variance in the regions of low intensity than the FBP images.

In this paper we will report on a more detailed comparative statistical study of MLE and FBP reconstructions of data from Fluoro-deoxiglucose (FDG) human brain PET studies of four normal subjects. A companion paper [4] describes the mathematical development of a methodology for MLE image reconstruction in PET that has been found to result in images that exhibit a consistent and favorable set of characteristics in terms of bias and expected error. That methodology will be used in the reconstruction of the images to be reported in this paper. Section II of this paper describes the methods, data and results of the statistical analysis. Section III describes the results of a Receiver Operating Characteristics (ROC) "lesion" detectability study of a very simple nature carried out with simulated images and non-medical personnel. The ROC study supports the idea that the mathematically describable advantage of the MLE images translates itself into improved lesion detectability in specific cases. The paper will conclude with a discussion of what has been accomplished and what still needs to be done in order to establish the MLE methodology as the preferred reconstruction method for PET.

II. BIAS AND VARIANCE ANALYSIS

Bias has a well-defined meaning in Statistics and, in the imaging context, could be described as differences between the expected value of reconstructed pixel intensities and the correct values. In the case of tomographic reconstructions, in which filtering of high frequencies is a necessity, the above definition of bias is not practical, since all edges in the image would show bias. A more practical definition that we devised in Ref. [3] is that of "regional bias" as the difference between the expected average value in extended uniform regions, away from edges, and the correct average values for those regions. In this paper we will examine regional bias for images reconstructed from real FDG human brain data. MLE reconstructions will be compared with FBP reconstructions of the same data sets and to reference images obtained by reconstructing corresponding data sets with large number of counts. For the reference FBP reconstructions we have used the Shepp-Logan filter in configuration space, which can be expected to exhibit less bias than filtering in frequency domain.[5] The pixel-by-pixel standard deviation from the mean image for sets of independent data from the same subject will also be examined for MLE vs. FBP reconstructions.

A. Procedures

1. Data and Reconstruction Methods

Data from four subjects, S1 through S4, have been used in this study. FDG data were taken in six sequential time intervals for each subject when the radioisotope concentration in the brain had stabilized. The number of counts per data set was between 1.1 and 1.4 million per image plane. The six data sets were followed by high count data sets that have been added to form a reference data set. From the data sets for each subject, 3 out of 15 planes have been selected, one in the upper, one in the middle and one in the lower brain. This corresponds to 18 data sets per subject, plus the reference sets. Data for subjects S1 and S2 were obtained with separate random coincidence files, while the randoms for S3 and S4 were pre-subtracted by the hardware. The differences in reconstruction strategy for the two types of scan data are described in the companion paper.[4]

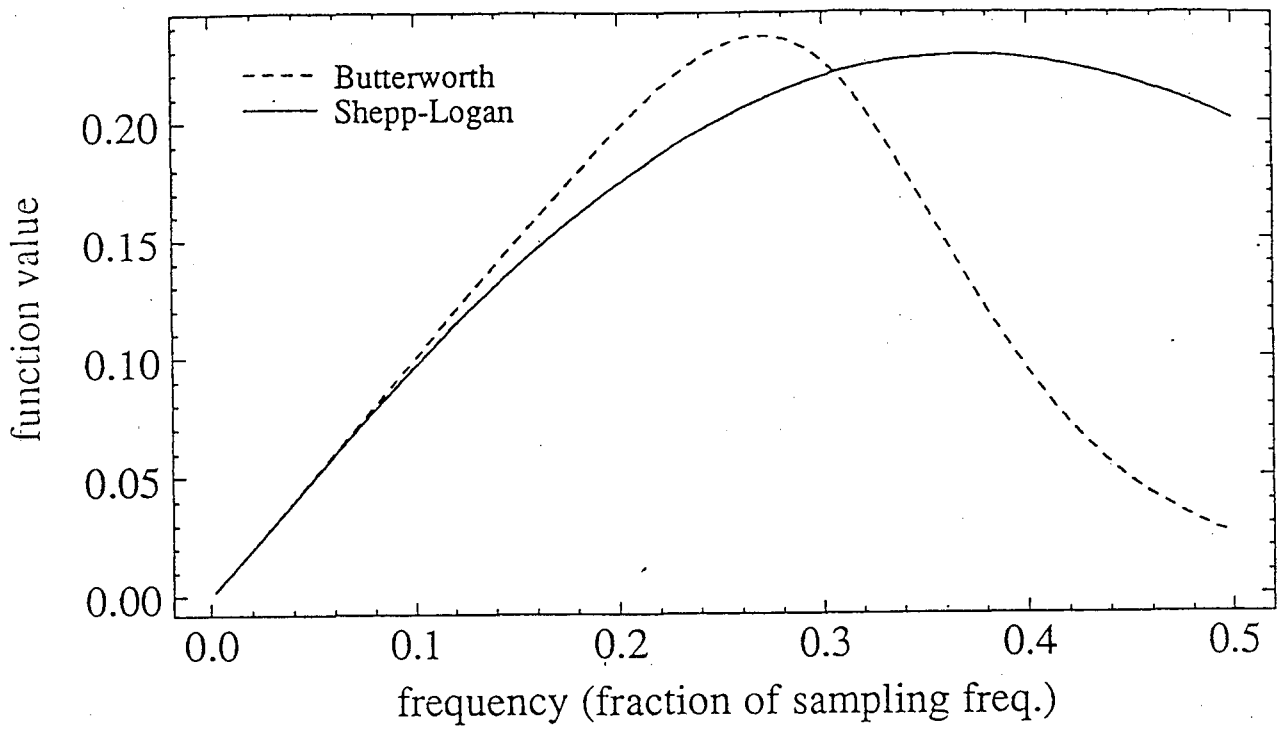


Fig. 1: Shepp-Logan and Butterworth filter functions used in the FBP reconstructions reported in this paper.

The independent data sets have been reconstructed by the FBP method with two different filters whose frequency pass characteristics are shown in Fig. 1. The Shepp-Logan filter with a cutoff frequency has been selected through the experience of several years of clinical practice with FDG brain images, and the Butterworth filter exhibits improved response in frequencies in the region between 0.15 and 0.3 of the sampling frequency (3.18 cycles/cm), while reducing higher frequency noise. The appropriateness of the parameters chosen for the Butterworth filter was confirmed by analyzing the power spectral density of a group of independent images of the same source distribution and determining the approximate cutoff frequency above which there was mostly noise. Independent visual tests carried out by physicians at the Department of Nuclear Medicine, UCLA have confirmed the appropriateness of the filter for the brain data used. The filtering and the backprojection (BIN) algorithms for the reconstructions were obtained from the Donner Algorithms Package.[6]

The data sets were also reconstructed by the "data splitting" MLE-PF method, following the methodology described in Ref. [4]. For the images resulting from MLE reconstructions to be statistically acceptable (feasible), the iterative procedure has to be stopped at some point. The stopping criterion that we have used is based on likelihood cross-validation, as described in detail in the companion paper.[4] The method contains a measure that defines the qualities of the image. In this paper, we will call this measure the Cross Validation Ratio (CVR). The CVR changes as the iterative procedure continues, with initial values of ~ 1.0 at the early iterations and a theoretical optimum of zero at the stopping point, but stopping the iterations at values between 0.5 and 0 should be studied when evaluating the reconstruction of real PET data. A final filtering with a Gaussian kernel is part of the MLE-PF procedure. We have used a filter with a standard deviation equal to 0.75 pixels. As shown in Fig. 2, the choice of this filter results in edge responses for the MLE-PF images which are as sharp as or sharper than those of the FBP reconstructions.

The reference data sets, containing between 10 and 30 million counts were reconstructed by FBP using the full bandwidth Shepp-Logan filter in configuration space. The reconstructions are somewhat noisy, but can be expected to be unbiased in the "regional" sense described above.

2. Bias analysis

Because of the lack of registration between FBP and MLE reconstructions (FBP assumes a space invariant point response function, while the transition matrix that we use for the MLE takes into consideration the narrowing tube size as we move away from the center of the image [1]), it is not possible to carry out a pixel-by-pixel comparison between the reference image and the individually reconstructed images, except within FBP reconstructions. The use of pixel value frequency histograms has also been found inadequate for the same reason. Then, bearing in mind our definition of "regional bias", we have established the following procedure:

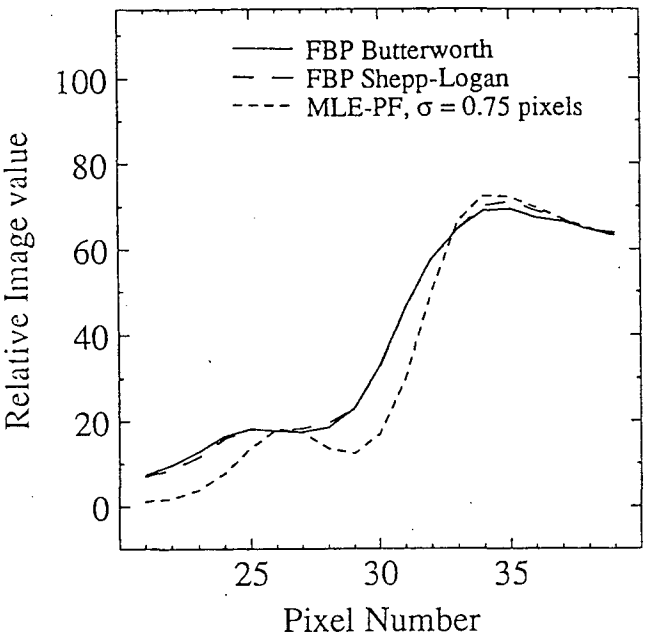


Fig. 2: Typical profile of a left edge of a brain PET image reconstructed by the FBP and MLE-PF. The edge appears sharper in the MLE-PF reconstruction than in the FBP cases. The pedestal to the left of the profiles corresponds to the activity in the skull.

a) The reference image is normalized so that the average value of a relatively large selected region of interest (ROI) with high isotope uptake corresponds to an image value of 100. The available number of image values (color levels in the display) is 128.

b) The six independent images to be analyzed are then normalized so that the high activity ROI also corresponds to an average image value of 100.

c) At least two more ROI's are selected in uniform areas of low activity, staying away from steps in activity. These ROI's can be of any shape, including long strings of pixels.

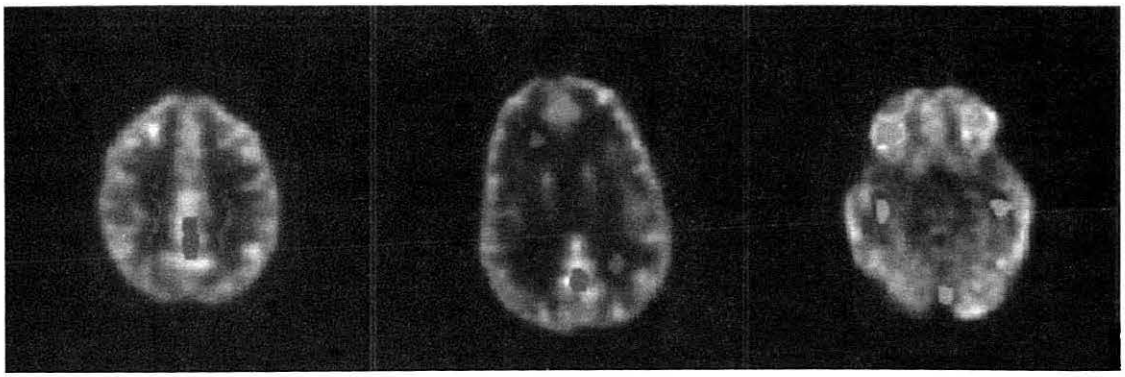
d) Ratios are obtained between the mean activity in the low activity ROI's and the high activity ROI. This is done for the reference image and for the "mean image" obtained by averaging pixel-by-pixel the six independent reconstructions of a given method.

e) The measured regional activity ratios of the reference are compared to those of the different reconstruction methods.

Figure 3 shows three sets of ROI's at different planes of different subjects, as examples of the ROI's selected for the above procedure.

3. Standard deviation analysis

Once we have established that a method of reconstruction is acceptably unbiased, the following calculations are carried out:



XBB 9112-9625

Fig. 3: Three sets of regions of interest used in the "regional bias" study. One region is in a high activity area, the others are in lower activity areas.

a) The pixel-by-pixel standard deviation with respect to the corresponding pixel mean is calculated for the six independent images. The resulting "standard deviation image" is then displayed and recorded with a color scale magnified by a factor of 10.

b) A two-dimensional histogram is generated in which the abcissa corresponds to the number of counts in a pixel of the mean image, the ordinate corresponds to the standard deviation in a pixel and the color scale corresponds to the logarithm of the number of pixels. The logarithm is taken in order to be able to see the large range of number of pixels in the different histogram bins in a single image display. The two-dimensional histogram shows the relationship between average counts in a region of the image and the standard deviation that can be expected in the reconstruction of that region, or "expected error".

B. Results of bias and standard deviation measurements

The results of the bias and standard deviation measurements on the sets of data from the four subjects can be clearly separated into two groups: a) results from reconstructions of data sets in which the randoms background was obtained separately, subjects S1 and S2, and b) results for data sets in which the background was subtracted by the hardware before reconstruction, subjects S3 and S4. The MLE reconstructions of data from S1 and S2 attained a reasonably unbiased condition and a favorable standard deviation when the iterations were stopped at the optimum point predicted by the cross-validation stopping rule, i.e., at CVR = 0. On the other hand reconstructions from from S3 and S4 data, if carried to that expected optimum point, exhibited excessive standard deviation. The detailed results will now be given.

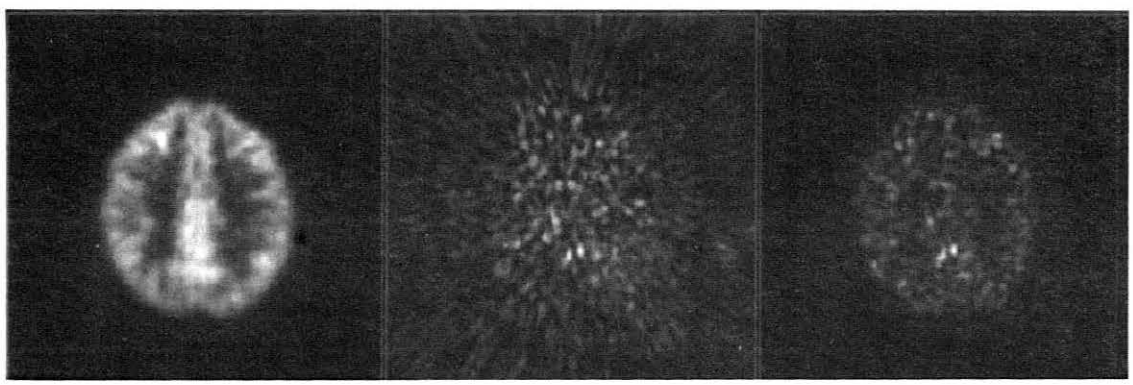
1. Bias results

Table I shows the ratios of average activity in low activity ROI's to average activity in the high activity ROI for five different reconstruction processes, for subject S1, plane 3 (near top of brain).

In all the reconstructions for all subjects, the differences between the ratios for FBP - Butterworth and FBP - Shepp-Logan filters have been small. On the other hand, stopping MLE-PF reconstructions at CVR = 0.5 and 0.3 points has always resulted in substantial bias. For those reasons, the FBP - Shepp-Logan and MLE-PF at those stopping points will not be discussed further. A summary of results of bias for subjects S1 and S2 is shown in Table 2. The results show that stopping the MLE-PF algorithm at the optimum point (CVR = 0), results in images that are as unbiased as FBP reconstructions.

TABLE 1 Bias Analysis, subject S1, estimated background	
Reconstruction method	**Activity ratio (low/high)**
Reference, FBP Shepp-Logan, conf. space	0.349
FBP - Butterworth	0.344
FBP - Shepp-Logan	0.351
MLE-EM - stop at $\Delta XL/\Delta L = 0.5$, $\sigma = 0.75$	0.386
ditto. $\Delta XL/\Delta L = 0.3$	0.368
ditto. $\Delta XL/\Delta L = 0.0$	0.354

TABLE II Bias Analysis, subjects S1 and S2, estimated background				
Patient	**Image plane**	**Ratio - Reference FBP - Shepp-Logan**	**Ratio - FBP - Butterworth**	**Ratio - MLE-PF CVR = 0.0**
P1	3 - top	0.349	0.351	0.354
	7 - middle	0.323	0.314	0.308
	11 - bottom	0.538	0.533	0.518
P2	3 - top	0.457	0.444	0.467
	7 - middle	0.266	0.259	0.268
	11 - bottom	0.429	0.424	0.421



XBB 9112-9626

Fig. 4: Standard deviation images for a brain plane near the top of patient P1. Left: reference image. Center: Standard deviation image for FBP - Butterworth reconstruction. Right: ditto for MLE-PF image. There is some large scale structure in the latter case, with lower standard deviation values where the reference image has low activity.

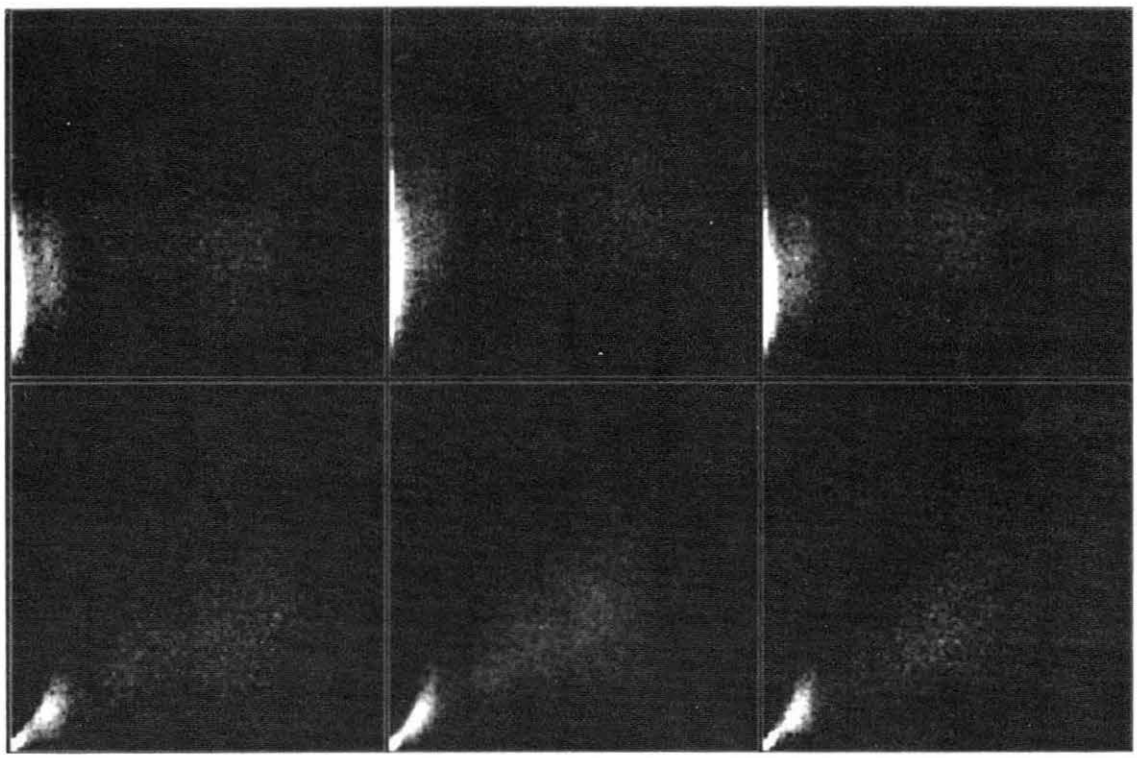
TABLE III Bias Analysis, subjects S3 and S4, pre-subtracted background					
Patient	**Image plane**	**Ratio - Ref. FBP - SL**	**Ratio - FBP-Butterworth**	**Ratio - MLE-PF at 0.2**	**Ratio - MLE-PF at 0.1**
P3	4 - top	0.457	0.443	0.449	0.437
	8 - middle	0.378	0.375	0.382	0.374
	12 - bottom	0.430	0.424	0.431	0.417
P4	4 - top	0.445	0.448	0.459	0.446
	8 - middle	0.367	0.360	0.375	0.366
	12 - bottom	0.528	0.526	0.522	0.519

For the case with background pre-subtracted by the hardware, the results of the statistical analysis are different from the above. Table 3 shows the bias results for subjects S3 and S4. Columns for MLE stopping points at CVR = 0.2 and 0.1 are shown. The case for CVR = 0 is not considered because it leads to excessive standard deviation.

2. Pixel-by-pixel standard deviation results

If we consider the MLE-EM results at CVR = 0.0 for subjects S1 and S2 acceptably unbiased and those at CVR = 0.2 or 0.1 also acceptable for S3 and S4, we can now proceed to study the pixel-by-pixel standard deviation of the reconstructions with respect to their means as a measure of how much error can be expected in a single reconstruction of one data set. Figure 4 (left) shows the reference image for subject S1, plane 3. Figure 4 (center) shows the standard deviation image for the FBP - Butterworth reconstructions, with a grey scale magnified by a factor of 10, and the same figure (right) shows the corresponding standard deviation image for the MLE-EM reconstructions at CVR = 0.0. It is evident that the expected error in the FBP image is distributed uniformly in the whole brain area, with some decrease in the regions outside the brain. In contrast, the standard deviation image for the MLE-EM results shows structure, with lower standard deviation in the regions of lower activity, and practically zero deviation outside the brain. This pattern of expected error is observed in all the standard deviation images that we have obtained and it suggests making two-dimensional histograms ("expected error" histograms) of frequency of pixels having a given mean number of counts and a given standard deviation. The dependency between mean number of counts in a region and the standard deviation can then be examined for the different reconstruction methods.

Figure 5 shows, on the top row, the expected error histograms for the three data sets (top, middle and lower brain) of subject S1, reconstructed by FBP-Butterworth. On the lower row the corresponding histograms for the MLE-PF at CVR = 0.0 are shown. There are two main regions in the histograms that correspond to the area of the brain: grey matter in the large region near the center and white matter towards the left. Further to the left there is the region of counts outside the brain (image background). Then, we can observe that the expected error in a MLE reconstruction in grey matter is not higher and often lower than the corresponding results in the FBP reconstruction, while the error in the white matter is substantially lower in the MLE. In an approximate quantitative way, one could state that the expected error in white matter is approximately equal to that of grey matter in FBP reconstructions, while it is reduced by at least a factor of the square root of the ratio of counts for the MLE results. The results for subject S2 support the same conclusion.



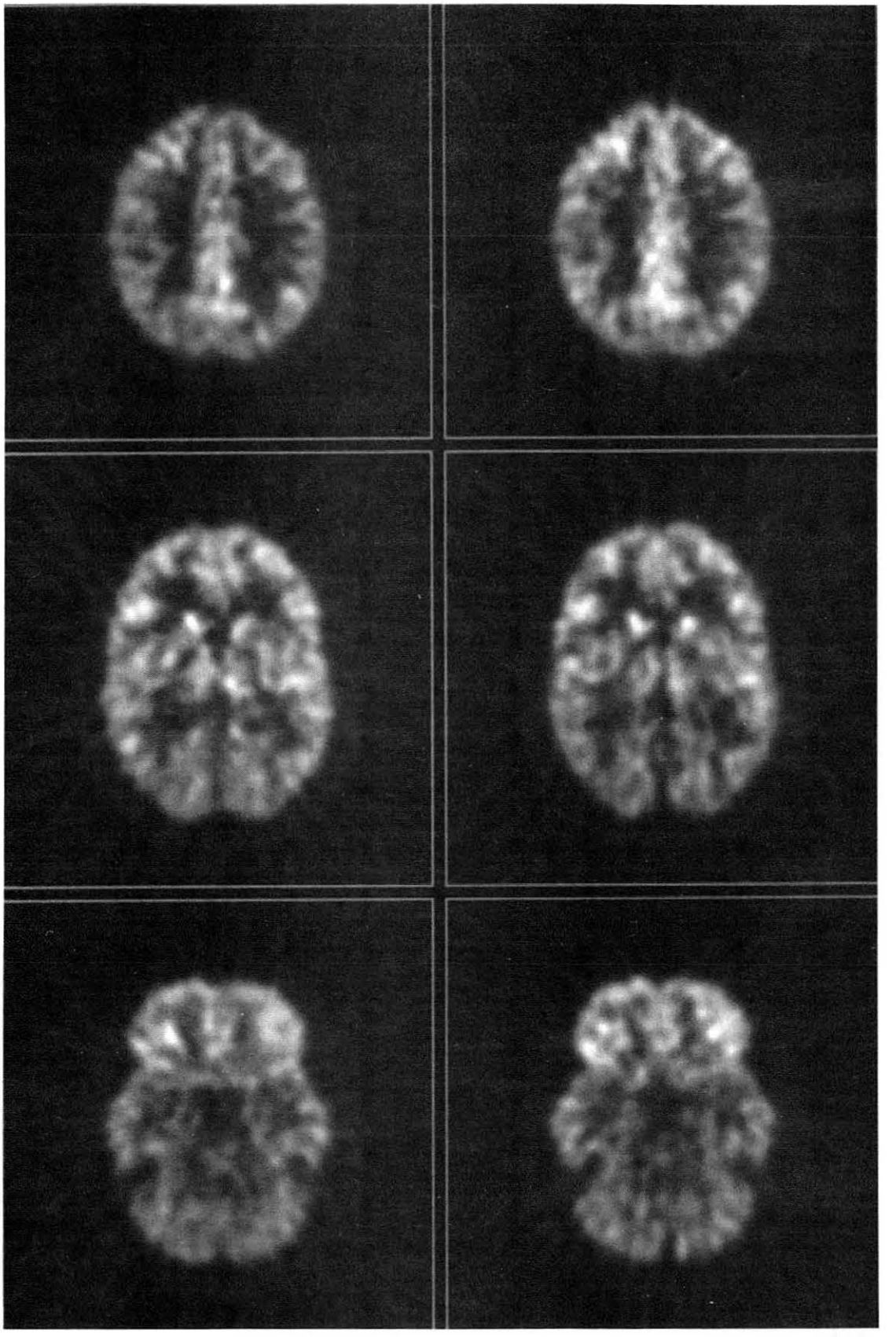
XBB 9112-9627

Fig. 5: Histograms of number of pixels having a given standard deviation (ordinate) and a given image value (abcissa). The standard deviation scale is magnified X10 with respect to the image value scale. Top row: Histograms from FBP reconstructions of top, middle and lower brain planes for patient P1. Bottom row: Histograms from MLE reconstructions of the same data. The large clusters near the middle of the histograms correspond to pixels in the grey matter, the next clusters to the left are in the white matter. The left-most clusters correspond to background, outside the brain.

The standard deviation results for subjects S3 and S4 show higher expected error in some parts of the grey matter areas than the FBP results if the iterative process is stopped at CVR = 0.1 and 0.0. Their results are similar to those of Fig. 5, however, where we stop at CVR = 0.2. It appears, then, that the justification for stopping the MLE procedure at CVR = 0.0 as being the point of highest consistency in the cross-validation process [4] does not apply well to data that have lost their Poisson characteristics by having even a small percentage of background counts subtracted (~ 3 - 5%), with negative numbers set to zero.

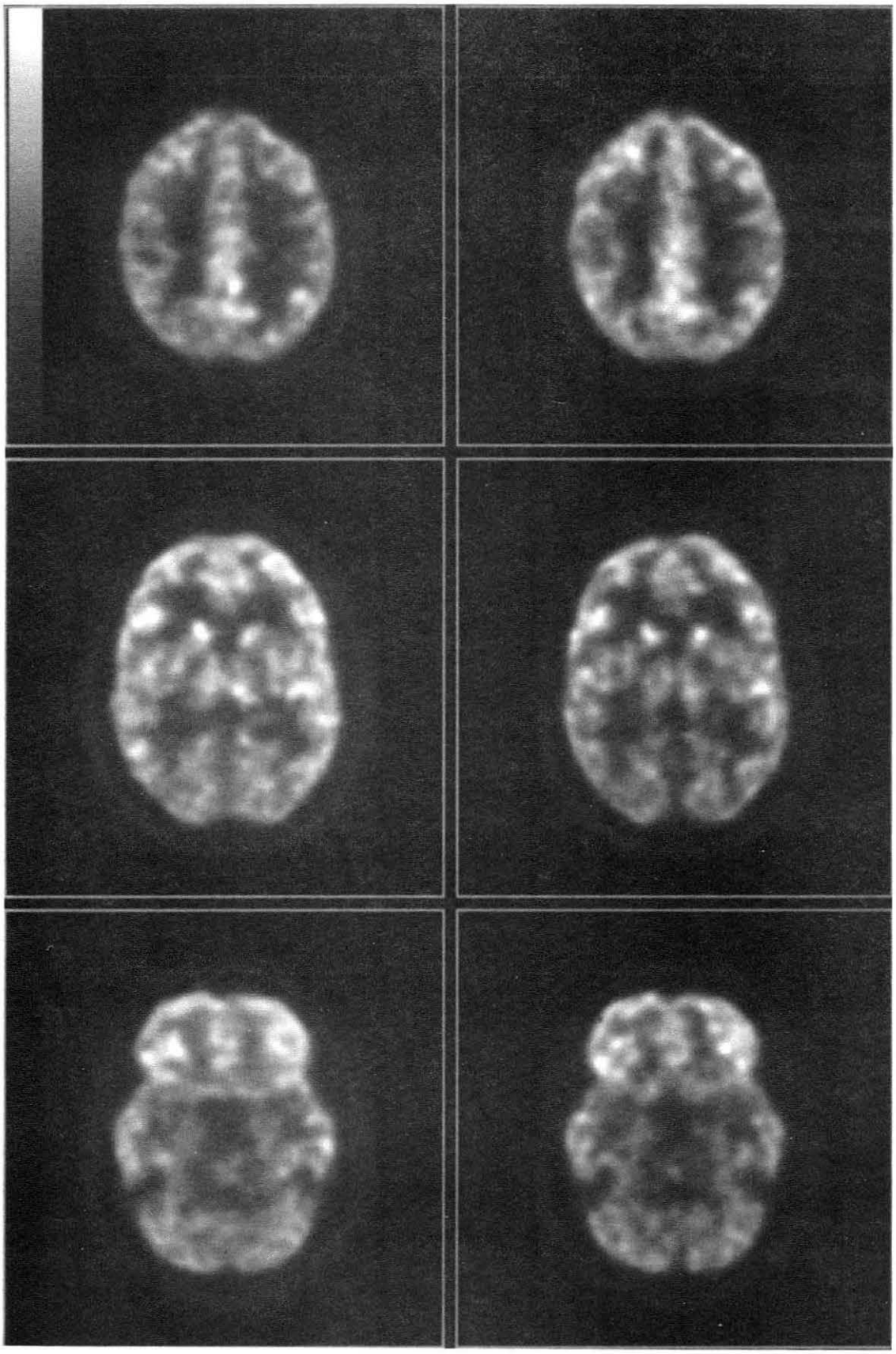
C. Effect of background subtraction methods

The above observations raise the question of whether images reconstructed from data sets in which the random background estimates have been obtained separately are "better" than those in which the background has been pre-subtracted. In order to answer this question, the data sets for subject S1 have been reprocessed by subtracting the randoms background (setting negative numbers to zero) and reconstructing accordingly. The results obtained show that regional bias and pixel-to-pixel standard deviation at CVR = 0.2 are practically indistinguishable from the results obtained originally from those data sets and so are the images. The number of iterations required by each method to arrive at the desired stopping point are similar. When the MLE is allowed to go to CVR = 0.1 with the pre-subtracted data, bias results do not change substantially, but the expected error (standard deviation histogram) shows a noticeable deterioration of the image with respect to the one obtained originally.



XBB 9112-9629

Fig. 6: Samples of FBP-Butterworth reconstructions. Two independent data sets each of top, middle and bottom brain planes are shown.



XBB 9112-9628

Fig. 7: Samples of MLE-PF reconstructions. The data correspond to the same data sets of Fig. 6.

The only conclusion that one can extract from this experiment is that, in cases with relatively low background (in the order to 5% or less), reconstruction from data in which the background data were obtained separately from the scan data allows for a more "elegantly" defined stopping point, but the results are not significantly different than reconstructions with pre-subtracted background.

D. Effect of speeding up the reconstruction by the Successive Substitutions method

The reconstructions reported above have been carried out by using the MLE iterative formulas based on the EM algorithm, as defined in Ref. [4]. The data sets for subject S1 have also been processed by the faster iterative formulas derived from the Successive Substitutions (SS) algorithm also described in Ref. [4]. Although in simulation experiments a speedup parameter of 3.0 could be used with no apparent instability, we have found that, with real data, we should stay below n = 2.0 for reliably stable solutions. This results in reconstructions of the S1 data sets that are visually indistinguishable from those of the EM algorithm. Bias results differ from the EM case usually in the third digit of the ratios of low activity to high activity regions, while the expected error histograms are not significantly different from those of the EM. The SS results have been obtained in approximately one-half the number of iterations needed by the EM method.

E. Examples of optimized MLE reconstructions

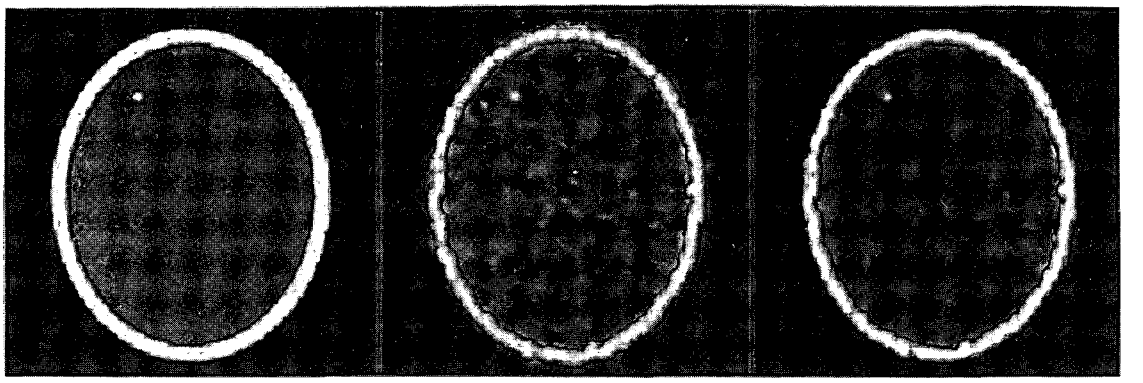
Having arrived at a point in which the characteristics of MLE images are well defined in terms of bias and expected error, it will be useful to give a comparative visual example of the differences between those images and the corresponding FBP images. Within the limitations of the printing process, it is hoped that the images of Figs. 6 and 7 will show the differences expected from the above analysis: a) MLE results that are at least as good as the FBP results in the areas of high radioisotope uptake and b) show a significantly reduced "noise" in the regions of low uptake. For this purpose we have chosen two independent data sets each of planes 3, 7 and 11 for subject S1. Figure 6 shows the FBP - Butterworth reconstructions, and Fig. 7 shows the MLE-PF at CVR = 0 results. Each one of the data sets contains between 1.3 and 1.45 million counts.

III. PRELIMINARY ROC ANALYSIS

A preliminary ROC study has been carried out in order to establish whether the lower expected error in the regions of low activity in MLE reconstructions results in an improved detection of small focal lesions in those areas, in comparison with FBP reconstructions. The ROC methodology, although it can only be applied to relatively simple observer tasks, is now well established as a reliable method of statistically determining the differences in performance of medical procedures that combine human observers and technology in carrying out medical diagnostics tasks.[7,8,9,10,11] This preliminary ROC study did not involve medical personnel, but was done with the assistance of ten scientists and engineers in different disciplines, carrying out a detection task. By designing a very simple task, it has been hoped that the use of non-medically trained observers would still yield meaningful results. This preliminary test was not meant to replace ROC tests with medical personnel, which are under way.

A. Design of preliminary study

The ROC study was carried out with a computer generated phantom consisting of an outer elliptical ring of 100% activity and an internal region of 25% activity. The dimensions of the phantom were similar to those of a human brain. From the source phantom, statistically independent data sets of 500k



XBB 9112-9624

Fig. 8: Examples of images used for a preliminary ROC study of detectability. The simulated phantom on the left contains one simple "lesion" which is to be detected in the FBP reconstructions (middle) or MLE-PF (right). The case shown is for a lesion of 85% intensity which should be easily detected in the reconstructions.

counts were generated according to the Poisson distribution by computer simulation. The CTI-831 Neuro PET tomograph was the model use for the data set.

"Abnormal" data sets were obtained by adding one lesion to the above described "normal" data. The lesion consisted of a hot region of 7.5 mm diameter with activity levels ranging from 55% to 85%, straddling the range of detectability by human observers in the presence of reconstruction noise. The lesion was placed at random in the internal 25% activity region. A total of 70 normal and 80 abnormal independent data sets were eventually generated and used in the study, although not all the observers saw all the sets. Each data set was reconstructed by the FBP and MLE-PF methods, except that the MLE reconstructions were stopped by the feasibility criterion described in Ref. [4]. The total number of images generated was 300. Figure 8, left to right, shows the source phantom, the FBP and the MLE reconstructions of a typical data set corresponding to an abnormal case with 85% activity level in the lesion, a reasonably easy case to detect correctly. The interior 25% region has been adjusted in the display so that it corresponds to the same average grey level in the three images.

Up to ten observers, readers r1 through r10, looked at each normal and abnormal image. The images were presented to the reader in random order on an image display station (color and b&w), and the reader was asked to respond to the question: "Does this image have a lesion?", after using any viewing methodology that they wished to establish, according to the following scale:

1 - almost definitely not
2 - probably not
3 - perhaps yes
4 - probably yes
5 - almost definitely yes

The results of the readings were processed by ROC evaluation programs supplied by Charles E. Metz of the University of Chicago. In addition to fitting ROC curves, the programs calculate a number of statistical parameters to test the validity of the hypothesis being tested.

B. Results of preliminary ROC tests

After processing the results for the ten readers, it was found that eight out of ten readers gave ROC curves similar to the one of Fig. 9, which corresponds to reader r4, indicating a substantial advantage in lesion detectability for the MLE reconstructions. One reader indicated a smaller advantage and one

reader showed no significant difference between the two reconstruction methods. The ordinate of the ROC curves corresponds to the true positive fraction (TPF), or the fraction of positive decisions in actually positive cases. The abcissa corresponds to the false positive fraction (FPF), or the fraction of positive decisions in actually negative cases.

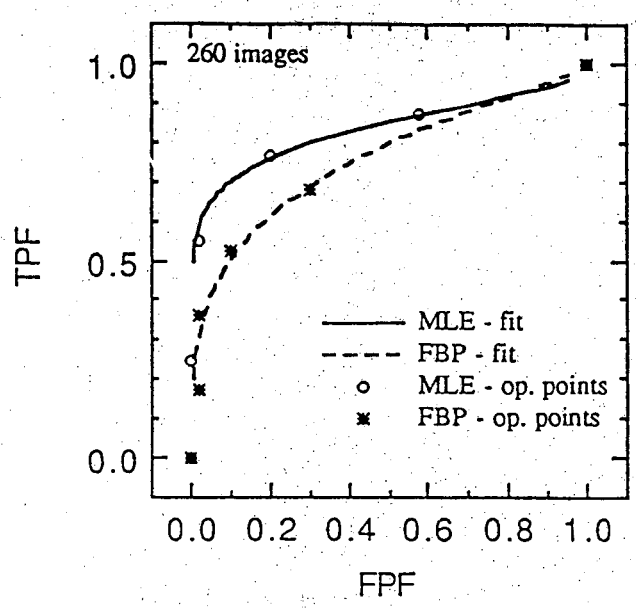


Fig. 9: ROC curve obtained from one of the observers (r4) typical in 8 of the 10 participants in the preliminary ROC study of detectability.

An interpretation of the above results can be given in the following manner: for the specific task of detecting a hot lesion on a cool field in a smooth phantom with lesion activity level near the threshold of detectability, we can state that the fraction of lesions detected correctly by eight out of ten observers, operating at a FPF = 0.15 (a region of reasonable desired performance), increased on the average from ~ 0.51 to 0.71 when using MLE-PF reconstructions, compared to FBP reconstructions. Table 4 summarizes the results of the preliminary study. Some derived values are not available due to accidental data loss in the early part of the study. The p-value, in our context, is defined as the calculated probability that the TPF for the MLE would be so much higher than that of the FBP if the MLE and FBP reconstructions were equally effective in lesion detection. In order to define the power, let us set an acceptance threshold $\alpha = 0.05$, i.e., we postulate that when p-value < 0.05 the modalities are accepted as having different diagnostic value, and when p-value > 0.05 we consider the modalities undistinguishable. Then power is a measure of the separability of the two methods at that value of acceptance threshold. More precisely, it is the probability of arriving at the decision that the two methods are different if they are indeed different. Values of $\alpha \approx 0.05$ and Power ≥ 0.80 are recommended for good statistical confidence in ROC studies. The p-values and powers shown in Table 4 were calculated by the programs CORROC2 and ROCPWR2 of the Metz package, using correlated pairs of responses given by the observers to the FBP and MLE reconstructions of each data set.

We conclude from the above analysis that there is a definite improvement in small lesion detection that corresponds to the decrease in pixel-to-pixel standard deviation in feasible MLE reconstructions. The question of whether detectability tests carried out with human data by physicians would result in results as favorable as the ones shown above cannot be answered at this time. The design of a clinical ROC study has been completed and the data generation and image reconstruction are underway at the time of this writing. Five M.D.'s at the Department of Nuclear Medicine, UCLA, are participating in the clinical study, which should be completed by the end of 1991.

IV. Discussion and Conclusions

Since Shepp and Vardi published their initial work on the MLE method of reconstruction for Positron Emission Tomography,[1] it has been known that MLE images exhibit "low noise", although what was actually meant by that has remained unclear for some time. In the last few years a number of

TABLE IV Statistical parameters derived from the preliminary ROC study See text for definition of parameters					
Reader	Number of images read	TPF at FPF = 0.15		p-value	Power at $\alpha = 0.05$
		MLE	FPB		
r1	226	0.65	0.46	Data loss	
r2	205	0.70	0.57	Data loss	
r3	220	0.78	0.74	0.58	≈ 0.08
r4	260	0.73	0.56	0.005	≈ 0.83
r5	160	0.57	0.31	Data loss	
r6	260	0.71	0.50	0.004	≈ 0.82
r7	300	0.67	0.58	0.13	≈ 0.35
r8	300	0.77	0.57	0.0005	≈ 0.94
r9	300	0.74	0.52	0.0004	≈ 0.95
r10	300	0.79	0.60	0.0004	≈ 0.96

workers have published attempts at characterizing MLE images in terms of the accuracy with which radioisotope uptake in specific regions of interest can be determined, by comparison with FBP images using their standard filtering methods. Typically, the comparison has been made with MLE images in which the stopping point of the iterative procedure has been determined arbitrarily. Still, other workers have developed variations on the iterative formulas with the aims of speeding up convergence, modifying the frequency contents of the resulting images, etc. by changing the iterative formulas in reasonable, but *ad hoc* ways, with the result that, typically, it is no longer known what function is being maximized or minimized by the process. Other workers, among them some of us, have presented algorithms based on Bayesian considerations that use some prior knowledge for the solution of the problem, in addition to the measured data. The use of the prior knowledge, principally entropy or a smoothness constraint, requires the incorporation of parameters that can be set in an educated way, but also with a large degree of arbitrariness. A similar problem exists in the process of limiting the set of images from which MLE estimates can be accepted (method of sieves). The result of all this activity, in too many papers to refer to here, is one of confusion in the minds of possible users of the method. The principal question: "Is the MLE method good for something specifically?" has not been addressed before in a consistent way, to our knowledge.

What we have attempted to do in this paper and its companion, is to present a methodology for solving the Emission Tomography image reconstruction problem that:

1) uses Maximum Likelihood as the target function to maximize, but recognizes that the iterative process must be stopped before the resulting images become unusable by the method attempting to follow noise in the data too closely,

2) uses a cross-validation method for determining that stopping point, with little or no arbitrariness when handling real tomographic data,

3) yields images that have similar statistical characteristics in the regions of high radioisotope uptake as nearly optimum FBP images, so that uptake measurements can be expected to be at least as good as those obtained with the latter images,

4) yields images that are less variable in the regions of low activity than FBP images, with an apparent improvement in lesion detectability in those regions, and

5) has excellent possiblities of being implemented in parallel processing structures at a reasonable price in the near future.

We believe that the work presented here goes some way in the direction of answering the above question: properly obtained MLE images can provide better information about regions of low activity without giving up accuracy of uptake measurements or anatomical information in the regions of high activity. This has been determined for the case of FDG studies in the human brain, but can be expected to extend to studies with all other radioisotopes.

Two areas, at least, need further verification and study:

1) quantification of radioisotope uptake as a function of time and

2) detectability of lesions in real data by medical personel.

In both cases, it would be expected that MLE-PF reconstructions are as good as FBP for ROI's in high activity areas, while information from ROI's placed in regions of low activity, in the presence of high activity regions, would be much improved with the MLE-PF reconstructions. We are proceeding to verify those expectations.

V. ACKNOWLEDGMENTS

The authors would like to thank Charles E. Metz of the University of Chicago for his assistance in setting up and interpreting the results of the preliminary ROC studies. This work has been supported, in part, by grants from the National Institutes of Health, CA-39501 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. The work of J. Nunez has been supported, in part, by the Center of Catalan Studies, Generalitat de Catalunya, Spain, and University of California, Berkeley. 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.

VI. REFERENCES

[1] L.A. Shepp and Y. Vardi, "Maximum Likelihood reconstruction for emission tomography", *IEEE Trans. Med. Imaging,* MI-1, No.2, 113-121 (1982).

[2] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Soc.*, B39, 1-37 (1977).

[3] J. Llacer, A.C. Bajamonde, "Characteristics of feasible images obtained from real PET data by MLE, Bayesian and Sieve methods", *SPIE Proc., Digital Image Synthesis and Inverse Optics,* 1351, 300-312 (1990).

[4] J. Llacer, E. Veklerov, K.J. Coakley, E.J. Hoffman and J. Nunez, "Optimization of Maximum Likelihood Estimator Images for PET: Algorithm Implementation", this issue.

[5] C.R. Crawford, "CT Filtration Aliasing Artifacts", *IEEE Trans. Med. Imaging*, MI-10, No. 1, 99-103 (1991).

[6] R.H. Huesman, G.T. Gullberg, W.L. Greenberg and T.F. Budinger, *Donner Algorithms for Reconstruction Tomography*, LBL Publication 214 (1977).

[7] J.A. Swets and R.M. Picket, *Evaluation of Diagnostic Systems*, Academic Press (1982).

[8] C.E. Metz, "Basic Principles of ROC Analysis", *Seminars in Nuclear Medicine*, VIII, No. 4, 283-298 (1978).

[9] C.E. Metz and H. Kronman, "Statistical Significance Tests for Binormal ROC Curves", *J. Mathematical Psychology*, 22, 218-243 (1980).

[10] C.E. Metz, "ROC Methodology in Radiologic Imaging", *Investigative Radiology*, 21, No. 9, 720-733 (1986).

[11] C.E. Metz, "Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies", *Investigative Radiology*, 24, No. 3, 234-245 (1989).

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
TECHNICAL INFORMATION DEPARTMENT
BERKELEY, CALIFORNIA 94720