ROC Study of Maximum Likelihood Estimator Human Brain Image Reconstructions in PET Clinical Practice: a Progress Report*

Jorge Llacer[1], Eugene Veklerov[1] and Deborah Nolan[1,2]
[1]Engineering Division, Lawrence Berkeley Laboratory
and [2]Department of Statistics, University of California
Berkeley, CA 94720
and
Scott T. Grafton[3], John C. Mazziotta[3,4], Randall A. Hawkins[3],
Carl K. Hoh[3] and Edward J. Hoffman[3]
[3]Division of Nuclear Medicine and Biophysics
and [4]Department of Neurology, School of Medicine
University of California
Los Angeles, CA 90024

LBL--29711

DE91 005371

Abstract

This paper will report on the progress to date in carrying out Receiver Operating Characteristics (ROC) studies comparing *Maximum Likelihood Estimator (MLE)* and Filtered Backprojection (FBP) reconstructions of normal and abnormal human brain PET data in a clinical setting. A previous stastical study of reconstructions of the Hoffman brain phantom with real data indicated that the pixel-to-pixel standard deviation in feasible MLE images is approximately proportional to the square root of the number of counts in a region, as opposed to a standard deviation which is high and largely independent of the number of counts in FBP. A preliminary ROC study carried out with 10 non-medical observers performing a relatively simple detectability task indicates that, for the majority of observers, lower standard deviation translates itself into a statistically significant detectability advantage in MLE reconstructions. The initial results of ongoing tests with four experienced neurologists/nuclear medicine physicians are presented. "Normal" cases of ^{18}F - fluorodeoxyglucose (FDG) cerebral metabolism studies and "abnormal" cases in which a variety of lesions have been introduced into "normal" data sets have been evaluated. We report on the results of reading the reconstructions of 90 data sets, each corresponding to a single brain slice. It has become apparent that the design of the study based on reading single brain slices is too insensitive and we propose a variation based on reading three consecutive slices at a time, rating only the center slice.

* This work has been supported, in part, by grants from the National Institutes of Health, CA-39501, 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.

I. INTRODUCTION

After a substantial amount of work carried out by our group (J. Llacer and co-workers) in developing an adequate understanding of the Maximum Likelihood Estimator (MLE) method of image reconstruction and the concept of feasible images for Positron Emission Tomography (PET) [1-5], it has become evident that the time has come for an assessment of the possible benefits that such reconstructions can give to medical diagnosis, in comparison with the established reconstruction technique, Filtered Backprojection (FBP). Theoretical "figures of merit" exist to assess the differences between two sets of images coming from different reconstruction techniques, for example, based on the "ideal observer" or the Hotelling trace. These indices can provide an objective measurement of those differences by analytic methods. Because of the differences between ideal observers and human observers, however, we felt it would be more reliable to carry out a Receiver Operating Characteristics (ROC) study in which trained medical observers would determine the differences between the reconstruction modalities to be examined. Although ROC methodology can only be applied to relatively simple observer tasks, it 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 medical diagnostics tasks[6-9].

This paper describes the reconstruction methods used and the process leading to the choice of reconstruction parameters, a preliminary ROC study carried out at Lawrence Berkeley Laboratory and finally, our ongoing ROC experiment with real PET data from FDG studies, giving the current status of the results and their interpretation.

II. DATA AND RECONSTRUCTION METHODS

All the patient data used in the tests to be described below were obtained from the CTI-831 Neuro PET system at UCLA and have 1.5 to 2.5 million counts per slice in the mid-brain. The FBP reconstructions for FDG studies have been traditionally obtained at that institution by the use of a Shepp-Logan filter whose characteristics are shown in Fig. 1 (slf1). Selection of the filter parameters has evolved over a period of years of clinical practice. An analysis of the instrument parameters and of the frequency spectra of reconstructions indicated that a Butterworth filter, Fig. 1 (bw0), could be used to improve the response at frequencies in the region of 0.5 to 1.0 Hz/cm while reducing higher frequency noise, although it can result in some "ringing" in narrow structures[5]. The appropriateness of the choice of parameters for the Butterworth filter was confirmed independently by tests carried out by the group at UCLA. This Butterworth filter, which we consider close to optimum for the task to be done, was used for all the FBP reconstructions in the ROC study. It was initially suggested to us by S. Derenzo and R. Huesman.

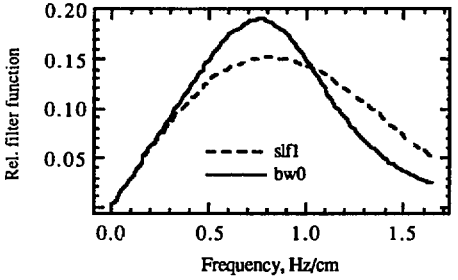


Fig. 1: Frequency spectra of the filters used for the FBP reconstructions. Curve "slf1" corresponds to the traditional Shepp-Logan filter used for images similar to those of this study. Curve "bw0" corresponds to a Butterworth filter with improved performance.

For MLE reconstructions, we employed "feasible" images. They are defined as those images that, if they were true activity distributions in a patient, could have given the initial data by a Poisson process[3]. Of the different feasible images that can be obtained, we have used those resulting from iterating with the EM algorithm past the onset of feasibility (~ 50% more iterations than minimum for feasibility, as determined by our test in [3]) and returning to feasibility by filtering with a two-dimensional Gaussian kernel of $\sigma = 0.65$ pixels (1 pixel = 0.185 cm). The total number of iterations was 40 in all cases. Statistical tests carried out on reconstructions of real data obtained from the Hoffman brain phantom indicate that those feasible images exhibit no significant regional bias, better high frequency response than the FBP images, and a pixel-by-pixel standard deviation that is approximately proportional to the square root of the number of counts in the region[5]. In contrast, FBP reconstructions exhibit a pixel-by-pixel standard deviation which is large and practically independent of the number of counts in the region considered.

III. PRELIMINARY ROC STUDY

A preliminary ROC study has been carried out in order to help us understand the relationships between the different statistical parameters involved in a clinical ROC study and to preview what may be the benefits of the favorable pixel-to-pixel standard deviation of the MLE reconstructions. This preliminary ROC study involved 10 scientists and engineers in different disciplines.

A. Design of preliminary study

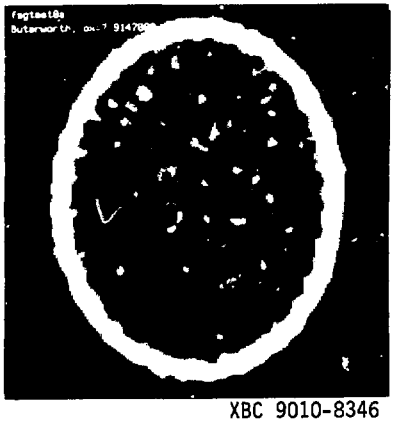
Our computer generated phantom consisted of an outer elliptical ring of 100% activity, with an internal region of 25% activity and dimensions similar to a human brain. From the phantom, statistically independent data sets of 500k counts were generated according to the Poisson distribution by computer simulation. The CTI-831 was the model used for the data set.

"Abnormal" data sets were obtained by adding one lesion to the above described "normal" data, which contained no lesions. 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% 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 methods, essentially as indicated in Sect. II, for a total of 300 images. Figures 2a and 2b show, respectively, the FBP and 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 grey level in both reconstructions.

Ten observers, readers r1 through r10, looked at each normal and abnormal image. The images were presented to the reader in random sequence on an image display station (color and b&w). The observers were asked to respond to the question, "Does this image have a lesion?" according to the following scale:

1 - almost definitely no
2 - probably no
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 U. of Chicago. In addition to fitting ROC curves, the programs calculate a



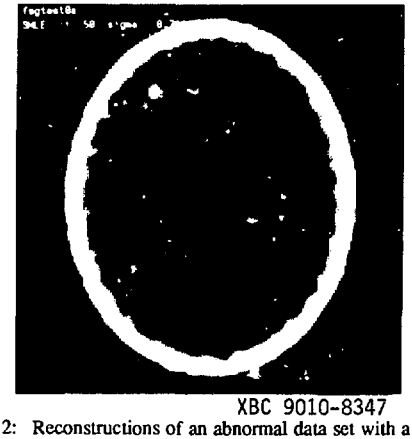


Fig. 2: Reconstructions of an abnormal data set with a 75% lesion. a) FBP with Butterworth filter, b) MLE feasible image. Gray levels are normalized to same value in the interior 25% region.

number of statistical parameters to test the validity of the hypothesis being tested.

B. Results of preliminary ROC tests

Data from eight of the ten readers produced ROC curves very similar to those of Fig. 3, indicating a susbtantial advantage of the MLE reconstructions. In addition, one reader indicated a smaller advantage and one reader showed no appreciable 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.

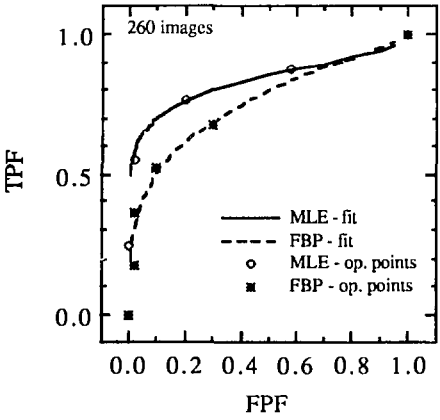


Fig. 3: ROC curves for reader r4 in the preliminary study. The operating points correspond to the results of the actual measurements.

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 brain-like phantom 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 feasible MLE reconstructions, compared to FBP reconstructions. Table 1 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 are equally effective in lesion detection at the .15 FPF level. In order to define the power, let's 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 $\geq$ 0.80 are recommended for good statistical confidence in ROC studies. The p-values and powers shown in Table 1 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. Each member of a pair was evaluated at different times.

TABLE 1

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	FBP		
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

IV. CLINICAL ROC STUDY

Encouraged by the above results, an ROC experiment involving initially four trained observers has been designed. The objective of this study is to determine whether feasible MLE images present an advantage over FBP images in PET of the human brain, specifically for FDG studies. FDG studies are the most common PET procedure carried out in the clinic at this time, and it is possible to expect the accumulation of enough data for meaningful statistical results in a reasonable period of time. On the other hand, the physiology of the human brain is such that FDG studies only show medically meaningful information in regions of high and moderate uptake (grey and white matter). There are no interesting regions with low counts. Because the advantages of MLE reconstruction may be primarily in low count regions, FDG studies cannot be expected to be optimal for the purpose of showing the benefits of the MLE method. Other radioisotope studies in which there are regions of interest with very low numbers of counts in the presence of regions of high counts would utilize the characteristics of the MLE to higher advantage. FDG studies are, nevertheless, of current clinical importance and it is felt necessary to ascertain the value of the MLE method in that clinical environment.

A. Design of ROC clinical study

The study starts with single slices of FDG-PET scans of normal volunteers or patients that have been declared normal. Approximately 45% of the slices remain as "normals" for the ROC study, while the remaining 55% of the slices are modified by the introduction of one focal (single spot) lesion. There are two families of lesions: a) lesions in the grey matter and b) lesions in the white matter. Within each family there are two types: subtractive or additive lesions. Each type has two intended levels of severity: reasonably high visibility and lower visibility. The lesions are introduced interactively on a normal image by "thinning" the projection data (for subtractive lesions) or by adding Poisson distributed counts to the projection data (for additive lesions), thus maintaining the Poisson character of the original data at all times. The modified data are then reconstructed by FBP and MLE methods to generate the "abnormal" images. The design of useful lesions for our study (near the threshold of detectability) is not an easy task and, in fact, it has turned out to be virtually impossible if the study was to be based on presenting single slices to the trained observers, as will be described below.

The following conditions were established for the initial design:

1) Each patient scan yields 15 slices. In order to decrease correlation effects between individual slices from a given patient, only alternating slices are being used for the study. Only odd or even slices are selected at random from a given patient. All available slices from the different patients form a pool of data sets.
2) Each image file to be presented for evaluation consists of 15 images. Of those images, 7 are "normal" and 8 are "abnormal". The "abnormal" images consist of one of each of the 8 different kinds of lesions: 2 families x 2 types x 2 severities. The selection of which scan data sets are destined to serve as normals and which as lesions is done at random from the pool of data sets.
3) Every image file contains 7 reconstructions by FBP and 8 by MLE, and a complementary image file contains 8 FBP and 7 MLE reconstructions, in such a manner that those data sets reconstructed by FBP in one image file are reconstructed by MLE in the complementary file. Complementary files are evaluated by the observers sometime after the initial files are evaluated.

4) Once the images forming one file have been defined and reconstructed, they are placed in random order in the final file structure and presented for evaluation.
5) The order of file presentation is such that if a particular type of lesion is first seen in a MLE reconstruction, for example, the same type of lesion will be seen first as a FBP reconstruction in a different data set.
6) Observers can use any tools available to them or any methodology that they decide upon for the evaluations, and are expected not to discuss the results of evaluations until completion of the study. The rating of the images is done in the same manner described in Sect. III A.

From our experience in the preliminary ROC study, we expect to need 150 different data sets for good statistical power. At the time of this writing, we have used 90 data sets, for a total of 180 images.

B. Results to date

The results from 90 data sets have been separated into the two main families: a) lesions in the grey matter and b) lesions in the white matter, although the results can also be pooled for an overall result or broken down in a variety of ways. The ROC curves that have resulted have, in general, a common characteristic, relatively independent of the method of reconstruction or lesion type: they are near 45° lines, i.e., the area under the ROC curve is not much above 0.5. This indicates the existence of a factor that, in effect, makes the observers' performance similar to a random rating. We have traced this effect to the observers' difficulty in determining whether an image is normal or abnormal when only a single slice in an FDG brain study is viewed. Lesions introduced by the test administrator which appear quite evident when the abnormal image is compared side-by-side with the corresponding normal image, become non-readable to an observer when no adjacent slices are available for comparison. Indeed, the variability of human anatomy is such that the abnormality that was so apparent to the test administrator could be quite normal anatomically when taken out of context. Lesions that are sufficiently large to be totally evident in isolation will not be useful in the ROC study, because both methods of reconstruction will show them equally clearly.

Figure 3 shows a pooled histogram of ratings (1 to 5) given by the four observers for images reconstructed by FBP. Three observers have rated 90 data sets each, one has rated 30 sets. It is observed that both actually negative and actually positive cases have a similar distribution. This result is anomalous since the weight of the distribution for the actually positive cases should be towards the higher ratings. There is only a small transfer of weight in the histogram in going from actually negatives to positives. MLE results are only slightly better.

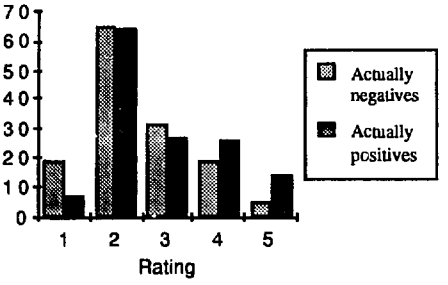


Fig. 3: Histogram of number of images receiving a given rating separated by actually negative and positive data sets, normalized to the same number of images for each class.

The above conditions correspond to a very insensitive ROC study and it is clear that the experiment needs to be redesigned.

C. Proposed new design

We propose to correct the experimental ROC design by providing normal and abnormal images in groups of 3 adjacent slices. The central slice will be rated for the existence or non-existence of a lesion, while the other two will be used for anatomical comparison. A new set of lesions will be generated that will appear principally in the central slice of the abnormal sets, but may extend into one of the adjacent slices to some extent. By providing the groups of 3 slices we expect to obtain two benefits: 1) the ratings for the actually normal slices will be weighted more strongly towards lower numbers, since the observers will be more secure in their decisions, and 2) the lesions introduced artificially will be more identifiable even in sizes and intensities that are just above marginal detection. The presentation of multiple slices will make the experiment closer to the reality of medical practice, though more complex to implement.

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 and current ROC studies and Landis K. Griffeth of the Washington University Medical Center for his valuable comments on the process of reading FDG studies, his assistance in generating realistic lesions, and evaluating those that we have generated. 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] J. Llacer, E. Veklerov and E.J. Hoffman, "On the Convergence of the Maximum Likelihood Estimator Method of Tomographic Image Reconstruction," *SPIE Proceedings*, Vol. 767, pp. 70-76, 1987.

[2] E. Veklerov and J. Llacer, "Stopping Rule for the MLE Algorithm Based on Statistical Hypothesis Testing," *IEEE Trans. Med. Imaging*, Vol. MI-6, pp. 313-319, 1987.

[3] J. Llacer and E. Veklerov, "Feasible Images and Practical Stopping Rules for Iterative Algorithms in Emission Tomography," *IEEE Trans. Med. Imaging*, Vol. MI-8, No. 2, pp. 186-193, 1989.

[4] J. Llacer, "On the Validity of Hypothesis Testing for Feasibility of Image Reconstructions," *IEEE Trans Med. Imaging*, Vol. MI-9, No. 2, pp. 226-230, 1990.

[5] J. Llacer and A. Bajamonde, "Characteristics of Feasible Images Obtained From Real PET Data by MLE, Bayesian and Sieve Methods," pres. SPIE 1990 Intl. Symp. on Optical and Optoelectronic Applied Science and Engineering, San Diego, July 1990; to be publ. *Conf. Proc.* LBL- 29150.

[6] J.A. Swets and R.M. Picket, *Evaluation of Diagnostic Systems,"* Academic Press, 1982.

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

[8] C.E. Metz, "ROC Methodology in Radiologic Imaging," *Investigative Radiology*, Vol. 21, No. 9, pp. 720-733, 1986.

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