Published in final edited form as:
IEEE Trans Nucl Sci. 2007 ; 54: 116–123.

Evaluation of Multiclass Model Observers in PET LROC Studies

H. C. Gifford, **IEEE Member**, **P. E. Kinahan**, **IEEE Senior Member**, **C. Lartizien**, **M. A. King**, and **IEEE Senior Member**
H. C. Gifford and M. A. King are with the Department of Radiology, University of Massachusetts Medical School, Worcester, MA 01655 USA (e-mail: howard.gifford@umassmed.edu).P. E. Kinahan is with the Department of Radiology, University of Washington, Seattle, WA 98195 USA.C. Lartizien is with CREATIS-UMR5515 CNRS-U630 INSERM, 69367 Lyon Cedex 07, France

Abstract

A localization ROC (LROC) study was conducted to evaluate nonprewhitening matched-filter (NPW) and channelized NPW (CNPW) versions of a multiclass model observer as predictors of human tumor-detection performance with PET images. Target localization is explicitly performed by these model observers. Tumors were placed in the liver, lungs, and background soft tissue of a mathematical phantom, and the data simulation modeled a full-3D acquisition mode. Reconstructions were performed with the FORE+AWOSEM algorithm. The LROC study measured observer performance with 2D images consisting of either coronal, sagittal, or transverse views of the same set of cases. Versions of the CNPW observer based on two previously published difference-of-Gaussian channel models demonstrated good quantitative agreement with human observers. One interpretation of these results treats the CNPW observer as a channelized Hotelling observer with implicit internal noise.

Index Terms

Human; model observers; image quality; localization ROC (LROC); positron-emission tomography (PET); tumor detection

I. Introduction

In a span of five years, positron emission tomography (PET) with FDG-18 has become a primary clinical modality for imaging many types of cancer [1], and continues to be the focus of researchers working to improve its diagnostic capabilities. These efforts invariably contend with the question of how to gauge diagnostic improvement without resorting to extensive human-observer studies. One solution uses mathematical model observers that can predict the tumor-detection performance of the humans [2]. Two well-known examples that have been applied in PET studies (e.g., [3], [4]) are the nonprewhitening matched filter (NPW) and channelized Hotelling (CH) observer. To date, these have generally been used in signal-known-exactly (SKE) ROC studies where lesion locations are known before-hand.

Detection tasks involving lesion localizations are oftentimes more clinically realistic for PET, yet model observers that are capable of such tasks have only recently been considered. D' Asseler *et al.* [5] used a multiclass channelized NPW (CNPW) observer in a localization ROC (LROC) study that investigated bootstrap resampling of list-mode data. Multiclass observers represent one way of extending CH-type linear observers from binary (or two-class) detection tasks to localization-detection tasks, and have been shown to correlate with human observers in SPECT LROC studies [6].

The basis for our current model-observer work was the alternative free-response ROC (AFROC) comparison of acquisition modes for whole-body PET that was recently presented by Lartizien *et al.* [7]. This AFROC study required the detection and localization of multiple targets in an image volume. While our long-term goal is to evaluate the multiclass model observers in such a task, an impediment to testing them against the human results in [7] was the clinically oriented design of the AFROC study, which allowed the human observers to adjust the image display. Our current model observers lack a mechanism for this. Thus, we considered herein a simpler single-slice LROC study with fixed display settings in order to maintain task equivalence among the observers. Nonetheless, the imaging simulation was exactly as detailed in [7], as our PET images were drawn from the reconstructed volumes used in that work.

The LROC study compared the coronal, sagittal, and transverse display formats for viewing images. The multiclass NPW and CNPW observers were applied, the latter with multiple channel models. The multiclass CH observer [8] was not included in this study because the available images were insufficient for its training.

Sec. II details the imaging simulation, model observers, and study methodology. Study results are presented in Sec. III. In Sec. IV, the results with the CNPW observer are discussed in the context of previous comparisons [6] between it and the CH observer.

II. Methods

A. Simulated Acquisition

The mathematical cardiac torso (MCAT) phantom [9] was modified for a whole-body FDG simulation by the addition of a neck and head, arms, and a bladder. The head and bladder were not part of the reconstructed volume, but did contribute out-of-field activity to the projection data. The dimensions of the phantom corresponded to a 170-cm patient with a weight of 70 kg. Spherical tumors in the liver, lungs, and background soft tissue were 1 cm in diameter and had relative tumor-to-background activities that varied with organ. Five tumor contrasts per organ were selected such that they would yield fractions of tumors correctly localized (or, more *simply, fractions found*) of approximately 0.1, 0.3, 0.5, 0.7, and 0.9 for human observers in the AFROC study [7]. The contrast assignment to the lesions was randomized. These contrasts ranged from 2.5–4.75 in the liver, 5.5–9.0 in the lungs, and 6.5–10.5 in the soft tissue. Each abnormal case contained 7 lesions, dispersed among the organs according to a multinomial probability distribution that specified means of 2.5 lesions in both the liver and lungs, and 2.0 lesions in the tissue regions. A lesion placement within a given organ was subject to the restriction that it be no closer than 1 cm from the organ surface or from other lesions, but was otherwise random. A total of 50 abnormal cases were created.

The data simulation modeled the characteristics of the Siemens/CTI ECAT HR+ scanner operated in fully 3D mode [7]. Noiseless transmission and emission sinograms were obtained using the ASIM analytic projector [10] that accounted for attenuation, scatter, and randoms. The Poisson noise added to the emission sinograms was consistent with clinical protocols utilizing a 12-mCi dose and an uptake period of 90 minutes. In addition to a noisy projection set for each of the 50 abnormal cases, 25 noisy sets of normal projection data were also generated.

B. Image Reconstruction

Image volumes were reconstructed using 4 iterations (16 subsets for 144 projection angles) of the FORE+AWOSEM algorithm [11], for which the attenuation correction applied noiseless weights. Each volume consisted of 225 transaxial slices (2.4-mm thickness), with slice dimensions of 128 × 128 (5-mm voxel width). These noncubic voxel dimensions led to oval lesion profiles when the images were viewed in the coronal and sagittal formats. Post-smoothing was performed using a 3D Gaussian filter whose 10-mm FWHM was set based on contrast-to-noise considerations [7]. Once the image slices were extracted for the LROC study, a final image processing for the study was carried out as described below.

The original multitarget AFROC purpose of these reconstructions conflicted with our one-target LROC requirement, so we worked with abnormal slices that contained a single lesion in a given organ. Observers could then be instructed to search within a particular organ in an image. To avoid problems from out-of-slice lesions, a lesion was used for a given display format only if its center and that of any other lesion in the same organ were at least five slices apart. Lesion-absent slices were extracted from both the normal and abnormal volumes, and were defined in the latter case using a similar five-slice criterion. In units of length, the minimum separation was 12 mm in the axial direction and 25 mm in the transverse plane, values that represented a trade-off between garnering enough lesions for the study and limiting the out-of-slice lesion effects. Table I shows the distribution of activities for the lesions in the LROC study.

The image slices underwent an organ-specific upper-thresholding that primarily influenced lesion detection in the lungs. To briefly explain this process, we consider the set of images that contain a lesion in one of the given organs. The maximum tumor pixel in the ith image is ψ_i. The maximum $\psi_{\max}$ and standard deviation σ_ψ were determined from the set of ψ_i, and the threshold level was set to $\psi_{\max} + \sigma_\psi$. The thresholding was followed by conversion to greyscale. The images were then zero-padded to the 256 × 256 dimensions compatible with our viewing software. Example images are shown in Fig. 1. The coronal slice in Fig. 1(d), thresholded for a tissue lesion, has a considerable amount of hot artifact at the lower extremes that is caused by the out-of-field bladder activity. Sagittal slices also suffered from these artifacts.

Not shown in Fig. 1 are the labels that were affixed in the upper left-hand corners of the images to inform the human observers as to which organ should be searched.

C. Human-Observer Studies

Two members of the medical physics group at the University of Massachusetts Medical School participated in the study. Both had extensive LROC experience reading simulated images. There were 104 lesions that satisfied the five-slice separation criterion for all three image formats, with 21 in the liver, 32 in the lungs, and 51 in the tissue. These formed the test set of lesions. A training set of 42 lesions (14 per organ) for a given format was formed with lesions that were specific to that format. With an abnormal/normal image pair for each lesion location, this amounted to 208 test images and 84 training images per format.

The observers read these images in two equal sessions, with a test subset of 104 images preceded by a training subset of 42 images. The subset reading order varied with observer. Ordinal confidence ratings were collected on a discrete, six-point scale, with rating levels 1 and 6 identified with the labels “high confidence absent” and “high confidence present,” respectively. The observers marked a suspected lesion location with a set of crosshairs that was controlled with the computer mouse. A localization was assessed as correct if it lay within a circle of radius R_{cl} centered on the true location. This *radius of correct localization* was determined empirically from the human observer data and a single value was applied for all the studies. We set R_{cl} by examining how the fraction found for each image-display format increases with the circle radius. The objective is to pick a radius within an interval where the fraction is relatively constant across the different formats. More details on the choice of R_{cl} are presented in Sec. III.A.

An observer’s data from the two image subsets were pooled for scoring purposes. An area under the LROC curve (A_L) was calculated from the pooled data using the maximum-likelihood

(ML) fitting software described in [12]. The overall score for a format was the average area computed over the two observers.

D. Model Observers

For a given study image $\hat{\mathbf{f}}$, a multiclass model observer computes a perception measurement Z_n at each image pixel n in a predefined search region Ω. The max and arg max of Z_n are then recorded as the confidence rating and localization, respectively, for $\hat{\mathbf{f}}$**.** For a "signal-known-statistically, background-known-exactly" (SKS-BKE) detection task, Z_n is given by the inner product

$$Z_n = \mathbf{w}_n^t[\widehat{\mathbf{f}} - \mathbf{b}], \tag{1}$$

involving the observer's template image $\mathbf{w}_n$ for the image pixel and the mean lesion-absent image **b** for the slice. The background subtraction primarily serves to inject observer knowledge of the patient anatomy into the detection task. The 25 normal reconstruction volumes were used to estimate **b**. Noise-free background reconstructions have also worked for this purpose [13], but none were available for this simulation.

The NPW and CNPW observer templates are constructed to be shift-invariant, allowing evaluation of (1) for all n by means of a single 2D cross-correlation operation. Both observers use the 2D lesion profile averaged over location. This mean profile **s** was estimated from the training images. The NPW template $\mathbf{w}_n$ is **s** centered on the nth pixel.

Using the average lesion profile for the template means that the detection task is SKS not only because the lesion location is unknown, but also because local variations in the lesion characteristics are not accounted for. However, these variations (such as lesion shape) are relatively slight in our images. Also, the characteristics at a given location are fixed. Eckstein *et al.* [14] have considered a more general SKS task in which the lesion could vary substantially at location n. In that case, the model observer comprises a set of templates $\mathbf{w}_{n,}\mathbf{j}$, where vector **j** indexes the range of lesion variables.

The CNPW-observer template for our task is the mathematical projection of the lesion profile onto a set of channel responses. With these responses represented by the columns of matrix $\mathbf{U}_n$, the template is

$$\mathbf{w}_{n,\mathrm{cnpw}} = \mathbf{U}_n \mathbf{U}_n^t \mathbf{s}. \tag{2}$$

The channels are defined in frequency space, and the responses are their inverse Fourier transforms. The three sets of radially symmetric, 2D channels presented in [15] were tested. One set had four square-profile (SQ) channels [Fig. 2(a)] defined as

$$\tilde{u}_c(\rho) = \begin{cases} 1 & \| \rho \| \in [\rho_0 2^c, \rho_0 2^{c+1}] \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

with the 2D frequency vector $\boldsymbol{\rho} = (\rho_x, \rho_y)$, c ε{0, 1, 2, 3}, and $\rho_0 = 0.015$ pixel^{-1}. The other two sets were difference-of-Gaussian (DOG) designs, with the definition

$$\tilde{u}_c(\rho) = \exp\left[-\frac{1}{2}\left(\frac{\| \rho \|}{Q\sigma_c}\right)^2\right] - \exp\left[-\frac{1}{2}\left(\frac{\| \rho \|}{\sigma_c}\right)^2\right]. \tag{4}$$

The channel standard deviation is $\sigma_c = \alpha^c \sigma_0$. A sparse-DOG (SDOG) model had 3 channels, with $\sigma_0 = 0.015$, $Q = 2$, and $\alpha = 2$ [Fig. 2(b)]. A dense-DOG (DDOG) model used 10 channels with $\sigma_0 = 0.005$, $Q = 1.67$, and $\alpha = 1.4$ [Fig. 2(c)].

It is important to note that unlike the CH observer, the CNPW observer is sensitive to the means of channel normalization. In Fig. 2, the spectral responses of the various channel sets have been normalized to 1.0. Other possible normalizations like integrated power or peak spatial response would produce a different observer template.

The use of noncubic reconstruction voxels caused the lesions to appear elongated in the coronal and sagittal slices [see Fig. 3(a)]. This led us to test the CNPW observer with a set of nonsymmetric SDOG channels, defined according to the formula

$$\begin{aligned}\tilde{u}_c(\rho) &= \exp\left\{-\frac{1}{2Q^2}\left[\left(\frac{\rho_x}{\sigma_{x,c}}\right)^2+\left(\frac{\rho_y}{\sigma_{y,c}}\right)^2\right]\right\}\\ &-\exp\left\{-\frac{1}{2}\left[\left(\frac{\rho_x}{\sigma_{x,c}}\right)^2+\left(\frac{\rho_y}{\sigma_{y,c}}\right)^2\right]\right\},\end{aligned} \tag{5}$$

with $\sigma_{x,c}$ the horizontal standard deviation and $\sigma_{y,c}$ the vertical standard deviation. With $\sigma_{x,0} = 0.015$ and $\sigma_{y,0} = 0.0075$, the one change from the symmetric SDOG channels was a smaller standard deviation in y. The response for $\tilde{u}_{0,}$ centered in the field-of-view, is shown in Fig. 3 (b).

E. Model-Observer Studies

The model observers read the same study images as the humans. The 84 training images per display format from the human study were used to estimate the observer templates. We compared the performances of all observers across the three display formats, and also analyzed their performances on the basis of format and organ.

We applied Swensson's fitting software to compute the model-observer performances, and matched the radius of correct localization to the value used in the human study (see Sec. III.A). Doing so maintained some consistency in scoring across our studies, although one practical modification was required to handle the continuous rating data produced by (1). In general, ML methods for fitting ordinal rating data are also valid for continuous data [16]. Ranking the continuous ratings in ascending (or descending) order defines a natural binning in terms of the run lengths of the tumor-present and tumor-absent images (see example below). Our difficulty was that having more than twelve bins led to convergence problems with the fitting software. To overcome this limitation, we applied an iterative rebinning procedure that reduced the ratings to the same six-point scale used for the human-observer study.

Fig. 4 diagrams a hypothetical 16-image example that helps explain the rebinning procedure. The ratings are sorted in ascending order to form a sequence S [Fig. 4(a)]. The sequence elements p_i and a_i are the ith largest ratings from the tumor-present and tumor-absent images, respectively. Beginning with the larger values of S, we use the consecutive runs of p's and a's to define the initial number of bins N_0 and the bin contents. We note that these run lengths also define a nonparametric, or empirical, ROC curve. Our hypothetical data produces four bins [Fig. 4(b)], based on p-run lengths of 4, 2, 1, 1, and corresponding a-run lengths of 1, 1, 3, 3. According to Metz [16], an ML fit of this ordinal data is equivalent to the ML fit of the original continuous data. At this point, an iterative rebinning step reduces this to $N_0 - 1$ bins [Fig. 4(c)] by combining the smallest bin with whichever adjoining bin has the fewer elements. The iterations continue until the desired number of bins is reached. One justification for this manner of rebinning is that assimilating the smallest bins will have the least impact on the shape of the empirical ROC curve.

This rebinning procedure was first applied by creating the sequence S from the full set of 208 images per display format. Typically, N_0 was on the order of 50. However, the different noise magnitudes and textures in the liver, lungs, and soft-tissue regions raised the question of whether the model-observer responses would be adequately normalized between these organs. Obtaining systematically higher responses in one organ than the others could skew the binning. To test this possibility, we also tried binning the ratings for each organ separately. These binned ratings were then combined prior to scoring. We refer to this approach as *partial binning* to distinguish it from the *full binning* that treated all images at once. In short, full binning defines observers that employ one consistent rating scale for all organs. With partial binning, the use of the rating scale is adjusted for organ in the sense that the discrete rating for a given image will be influenced only by the other images for that organ.

Previous experiments [6] showed little difference between using six or twelve bins with the Swensson code. As part of our current analysis of the model observers, we also computed the Wilcoxon area estimates [17] obtained from numerical integration of the empirical LROC curves. The rebinning procedure was not applied for these estimates.

All the human and model observers were made aware of the 1-cm minimum separation imposed between the lesions and the organ boundaries. For the model observers, the search region Ω in a particular organ was simply defined by the organ area minus a 1-cm margin at the boundaries. The observers were also tested with definitions of Ω that did not account for the margin.

III. Results

A. Localization Radius

Fig. 5(a) is a plot of the fractions found for the human observers as a function of the radius of correct localization R_{cl}. The six curves treat the data from the two observers using the three display formats. All these curves reach a plateau at about the same point, evidence that the accuracy of the lesion marking by the individual observers was fairly consistent across formats, and thus unaffected by the noncubic voxels. On the basis of this graph, R_{cl} was set to 15 mm. Partial-volume effects, the Gaussian post-smoothing, and the image interpolation to 256×256 pixels contributed to an R_{cl} that exceeded the actual 5-mm lesion radius.

The effect of R_{cl} on the fractions found for the model observers is summarized by the rest of Fig. 5. The search regions for the observers included the 1-cm margins at the organ boundaries. Each of the plots (b–d) treats the localization data obtained from the four observers applied to one of the display formats. The curves for the channelized observers are similar to the human curves in Fig. 5(a). Although the choice of R_{cl} was based solely on the human data, it is evident that small changes in the radius would not significantly change the CNPW-observer performances. The plateaus for the NPW observer are less distinct and show more variation across the three formats, indicating that the choice of R_{cl} will have a somewhat greater effect on this observer. Still, the localization penalties in this case, particularly with the coronal images, would have a relatively small impact on our analysis.

B. Aggregate Observer Performance

Table II lists areas under the LROC curve for the two human observers with the three image formats. A two-way ANOVA coupled with Scheffe's multiple-comparisons test [18] did not find a statistically significant interobserver effect ($\alpha = 0.05$). The differences between the coronal format and the sagittal and transverse formats were significant. The lower scores for the coronal slices are attributable in part to the larger areas that must be searched relative to the sagittal and transverse slices.

Table II also compares the average human and model-observer performances. The results for the model observers include A_L calculated with the Swensson code and using both the full and the partial binning schemes. Full binning decreased A_L by approximately 0.02 in most cases compared to the partial binning. With either scheme, each model observer attained its worst performance with the coronal format. A multiple-comparisons test of the model observers with the partial binning found significant differences between the coronal format and the other two formats and between the NPW observer and the CNPW observers.

The Wilcoxon estimates of A_L for the model observers are given in Table III, along with standard deviations calculated from sets of 10000 bootstrap resampling trials. As these estimates have only slight differences with the full-binning scores in Table II, the remainder of the analysis in this section was carried out with the Swensson software alone.

The CNPW observer was effectively indifferent to the switch from the symmetric SDOG channels to the nonsymmetric ones defined by (5). The nonsymmetric channels produced partial-binning scores of 0.53 ± 0.04 and 0.74 ± 0.04 for the coronal and sagittal images, respectively.

C. Stratified Observer Performance

Observer performances calculated on the basis of lesion location and display format are given in Table IV. Partial binning was applied for the model observers. The higher uncertainties compared to Table II reflect the relatively smaller numbers of images with which to estimate each LROC curve. With each format, human performances were lowest for the liver, while their lung and soft-tissue scores were comparable. In the volumetric AFROC study [7], the ordering of the human scores was: 1) lungs, 2) soft tissue, and 3) liver. An exact correspondence between the AFROC and LROC studies should not be expected since the current study used only a subset of the AFROC cases and its images underwent more-controlled post-processing.

From Table IV, it can be seen that the model observers usually agreed with the humans in the sense that their scores were considerably worse for liver lesions than for lung and soft-tissue lesions. The one exception was the NPW observer applied to the sagittal slices. Overall, this observer deviated the most from the humans, as can be seen by the scatter plots in Fig. 6. These plots are one-on-one comparisons of the human and model-observer data in Table IV. The linear correlation coefficients r for the DDOG, SDOG, and SQ versions of the CNPW observer were 0.91, 0.86, and 0.84, respectively. With full rebinning, the coefficients were 0.92, 0.87, and 0.88. The r-values for the NPW observer were 0.62 and 0.65, due primarily to its poor coronal scores. The critical value of r for rejecting the null hypothesis of no correlation at the $\alpha = 0.05$ significance level is 0.67.

Although these correlation coefficients offer one measure of agreement between the human and model observers, they do not account for any systematic biases that might exist between the observers. For example, Fig. 6(c) suggests a tendency of the SQ-channel CNPW observer to underperform compared to the humans. The dashed diagonal line in each scatter plot denotes the ideal model-observer behavior, and we tested the null hypothesis that this fit our experimental results. For a given scatter plot, a χ^2 statistic with 16 degrees of freedom takes the form of the weighted total-least-squares error

$$\chi^2 = \sum_{i=1}^{9} \frac{(A_i - A_{\mathrm{mod},i})^2}{\sigma^2_{mod,i}} + \frac{(A_i - A_{\mathrm{hum},i})^2}{\sigma^2_{\mathrm{hum},i}} \tag{6}$$

between the nine data pairs ($A_{hum},_i$, $A_{mod,i}$) and corresponding points (A_i, A_i) on the line. The latter points were chosen to minimize the error. An assumption underlying this test is that the observer scores obey independent Gaussian statistics, and the uncertainties listed in Table IV were adopted as the standard deviations $\sigma_{hum,i}$ and $\sigma_{mod,i}$ for the distributions. The hypothesis that the diagonal line fit the NPW-observer data was rejected for $\alpha = 0.05$. Rejection was not possible for any of the CNPW observers even when the $\sigma_{mod,i}$ values were halved.

D. Effect of Search Region

All the model-observer results to this point were based on search regions ω having the 1-cm margin at the organ boundaries. Without these margins, the model-observer performances dropped sharply: averaged over the three display formats and the three channel models, A_L for the CNPW observers decreased by 0.19 for lung lesions, 0.04 for soft-tissue lesions, and 0.03 for liver lesions. The difficulty in the lungs is that this larger search region forces the observers to consider possible lesion locations that contain considerable spill-in from the hotter soft-tissue areas.

IV. Discussion

On the basis of aggregate performance, the best human-model agreement was achieved by the CNPW observers with the DOG channels. The stratified analysis of performance indicates this correlation extended to the individual organ types. The largest discrepancy between the CNPW and human observers was obtained with the square channels applied to the transverse images, although this difference was not statistically significant.

The NPW observer has the same training requirements as the CNPW observer, as both use estimates of s and b. That the NPW observer considerably underestimated the human performances is not surprising in light of earlier SKE-task work (e.g., [19], [20]) and our own LROC results [6] that have indicated a human ability to partially prewhiten images. Burgess *et al.* [20] showed that their human-observer data could not be fit by any modified NPW-observer model. How, then, to reconcile the performance of the CNPW observer?

As implemented in the CH observer, prewhitening takes the form of an inverse weighting of the channel responses in (2), using the $C \times C$ covariance matrix **K** for a C-channel model. With a detection-localization task, **K** is indexed by search pixel. Generalized with internal noise that uses an additive scalar a to uniformly increase channel variances [15], the CH-observer template at the nth pixel is

$$\mathbf{w}_{n,\mathrm{ch}} = \mathbf{U}_n[\mathbf{K}_n + a\mathbf{I}]^{-1}\mathbf{U}_n^t\mathbf{s}, \tag{7}$$

where **I** is the identity matrix. The training for this observer is typically dominated by the estimation of $\mathbf{K}_n$. A bootstrap study of the training-image requirements based on a SPECT simulation [6] set a target of approximately 150 tumor-absent images for estimating $\mathbf{K}_n$, far exceeding the 25 normal images available for this PET study.

From (2) and (7), it is evident that the CNPW observer differs from the CH observer by a covariance-dependent internal-noise term $\mathbf{K}_{n,\mathrm{noise}}$ at the nth pixel given by $k_n\mathbf{I} - \mathbf{K}_n$, where k_n is the largest diagonal element of $\mathbf{K}_n$. In this sense, the CNPW observer has more in common with the CH observer than with the NPW observer. It is not clear what physical process might fit such an internal-noise model, especially as $\mathbf{K}_{n,\mathrm{noise}}$ is not guaranteed to be a true covariance matrix, but this formulation does suggest the potential volatility of the CNPW observer as a computationally simpler substitute for the CH observer. The substitution is obviously least apt for tasks where the prewhitening has a major impact. With white channel noise, $\mathbf{K}_n$ is proportional to **I** and the two observers are equivalent. This condition is unlikely to hold exactly for most realistic channel models and imaging applications, but even with nuclear-medicine simulations, we have obtained similar performances with the CNPW and CH observers. This was the case in our previous LROC experiments with SPECT OSEM images [6], where the channel noise was strongly uncorrelated but non-white. One explanation for this agreement relates to the relatively poor resolution of the SPECT images, whereby the white-noise requirement is weakened since only a subset of relatively low-pass channels contribute to the lesion detection. To extend this line of reasoning to the higher-resolution PET study, the effective white-noise requirement is more stringent since more channels play a role in the detection task. Given truly non-white noise, the degraded performance of the CNPW observer should then become more apparent with PET images. This would explain why no explicit internal noise was required for the CNPW observer to fit the human data, whereas it was necessary for the earlier SPECT studies.

More work is required to assess the reliability of the CNPW observer, as its performance relative to the CH observer will depend on many factors. One factor is reconstruction algorithm: CNPW and human performances with SPECT EBP images have compared poorly [6], [21]. Even with statistical reconstruction, however, more extensive ranges of parameters need to be tested. The effects for SKE tasks should also be evaluated, as this observer has rarely been applied for such tasks in emission imaging. Above all, as with any other human-model observer, the use of the CNPW observer for new applications should be validated against human data.

Our method of data analysis was not a constraint on the use of the CNPW observer. We make extensive use of the Swensson code in our LROC studies, coupled with rebinning of the continuous rating data from the model observers. Comparison with Wilcoxon estimates of A_L from the original rating data did not find appreciable differences. Working with ROC analysis, Metz and colleagues [16], [22] employed similar rebinning procedures as a means of shortening the ML computation times, and found slightly reduced statistical efficiency and increased bias in standard-error estimates as consequences [16].

V. Conclusions

The multiclass CNPW observers with DOG channels demonstrated good quantitative agreement with the human observers for this set of LROC studies. An explicit internal-noise model was not required for the data fit, although these model observers can be interpreted as versions of the CH observer containing implicit internal noise. More extensive studies are needed to test the overall applicability of the CNPW observer as a model for humans in emission-imaging studies.

Acknowledgements

The authors would like to thank the anonymous reviewers for comments that greatly improved the manuscript.

This work was supported by the National Institute for Biomedical Imaging and Bioengineering under Grant EB-02798 and by the National Cancer Institute under Grant CA-74135. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the granting agencies.

References

1. Rohren EM, Turkington TG, Coleman RE. Clinical applications of PET in oncology. Radiology 2004;231:305–332. [PubMed: 15044750]
2. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. Proc Natl Acad Sci USA 1993;90:9758–9765. [PubMed: 8234311]
3. Leahy RM, Chan MT, Cherry SR, Mumcuoglu EU, Czernin J, Chatziioannou A. Comparison of lesion detection performance for PET image reconstruction using channelized Hotelling observers. J Nucl Med 1997;38:286.
4. Fakhri GE, Badawi RD, Surti S, Holdsworth CH, Kinahan PE, Karp JS, Moore SC. Is NEC a useful surrogate for lesion detectability in whole-body PET? J Nucl Med 2004;45:164P. [PubMed: 14960631]

5. D'Asseler Y, Groiselle CJ, Gifford HC, Vandenberghe S, de Walle RV, Lemahieu IL, Glick SJ. Evaluating human observer performance for list mode PET using the bootstrap method. Proc IEEE Nuclear Science `Symp and Medical Imaging Conf Rec 2003;5:3070–3073.

6. Gifford HC, King MA, Pretorius PH, Wells RG. A comparison of human and model observers in multislice LROC studies. IEEE Trans Med Imag 2005;24:160–169.

7. Lartizien C, Kinahan PE, Comtat C. A lesion detection observer study comparing 2-dimensional versus fully 3-dimensional whole-body PET imaging protocols. J Nucl Med 2004;45:714–723. [PubMed: 15073270]

8. Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. J Opt Soc Amer A Opt Image Sci 1987;4:2447–2457.

9. Tsui BMW, Zhao XD, Gregoriou GK, Li J, Lalush DL, Eisner RL. Quantitative cardiac SPECT reconstruction with reduced image degradation due to patient anatomy. IEEE Trans Nucl Sci Dec; 1994 41(6):2838–2848.

10. Comtat C, Kinahan PE, Defrise M, Michel C, Townsend DW. Simulating whole-body PET scanning with rapid analytical methods. Proc IEEE Nuclear Science Symp and Medical Imaging Conf Rec 1999:1260–1264.

11. Comtat C, Kinahan PE, Defrise M, Michel C, Townsend DW. Fast reconstruction of 3D PET data with accurate statistical modelling. IEEE Trans Nucl Sci Jun;1998 45(3):1083–1089.

12. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys 1996;23:1709–1725. [PubMed: 8946368]

13. Gifford HC, Pretorius PH, King MA. Comparison of human-and model-observer LROC studies. Proc SPIE 2003;5034:112–122.

14. Eckstein MP, Zhang Y, Pham B, Abbey CK. Optimization of model observer performance for signal known exactly but variable tasks leads to optimized performance in signal known statistically tasks. Proc SPIE 2003;5034:123–134.

15. Abbey CK, Barrett HH. Human-and model-observer performance in ramp-spectrum noise: Effects of regularization and object variability. J Opt Soc Amer A Opt Image Sci 2001;18:473–488.

16. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in Medicine 1998;17:1033–1053. [PubMed: 9612889]

17. Hanley JA, McNeil BJ. The meaning and the use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. [PubMed: 7063747]

18. Pollard, JH. A Handbook of Numerical and Statistical Techniques. Cambridge, U.K: Cambridge Univ. Press; 1977.

19. Rolland JP, Barrett HH. Effect of random background inhomogeneity on observer detection performance. J Opt Soc Amer A Opt Image Sci 1992;9:649–658.

20. Burgess AE, Li X, Abbey CK. Visual signal detectability with two noise components: Anomalous masking effects. J Opt Soc Amer A Opt Image Sci 1997;14:2420–2442.

21. Gifford HC, Zheng XM, Boening G, Bruyant PP, King MA. An investigation of iterative reconstruction strategies for lung lesion detection in SPECT. Proc IEEE Nuclear Science Symp and Medical Imaging Conf Rec 2004;7:4241–4245.

22. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation. Acad Radiology 1997;4:298–303.

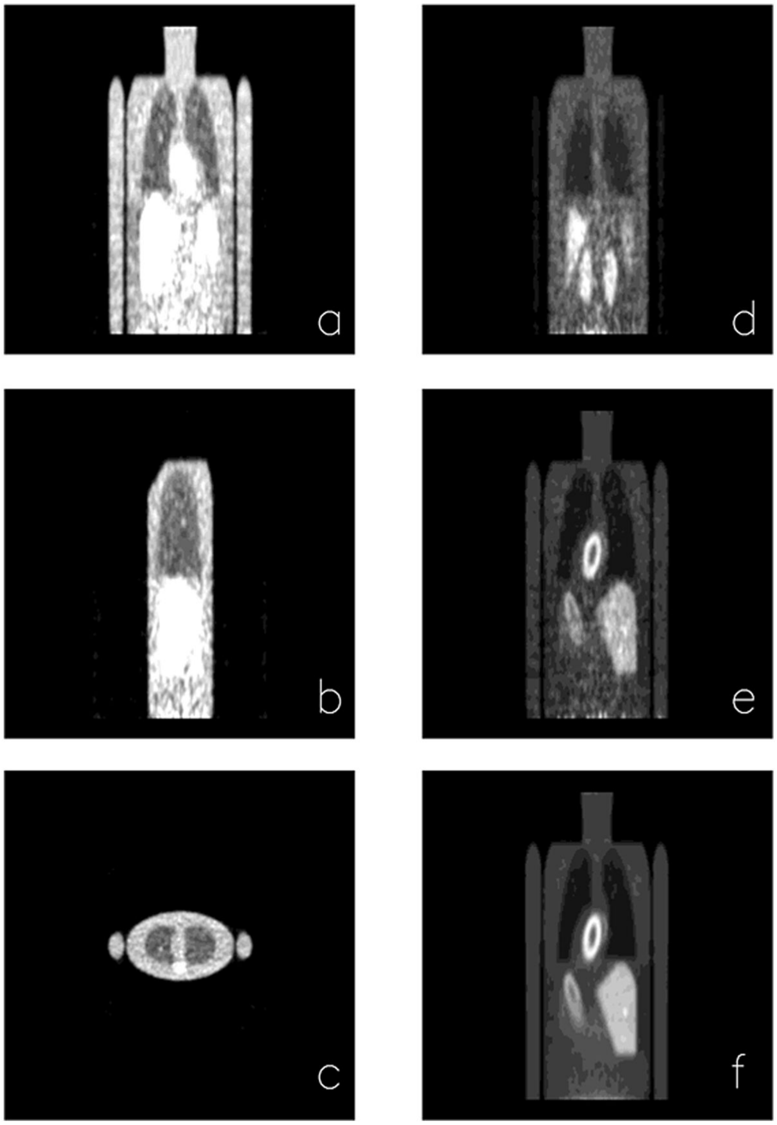

Fig. 1.

Example study images. Down the left-hand column (a—c), a case with a lesion in the right lung is shown in the coronal, sagittal, and transverse views. The coronal image (d) has been thresholded for a tissue lesion, while (e) and (f) are noisy and noise-free images of a case with a liver lesion. The stretched appearance of the phantom along the axial direction is a consequence of the noncubic reconstruction voxels.

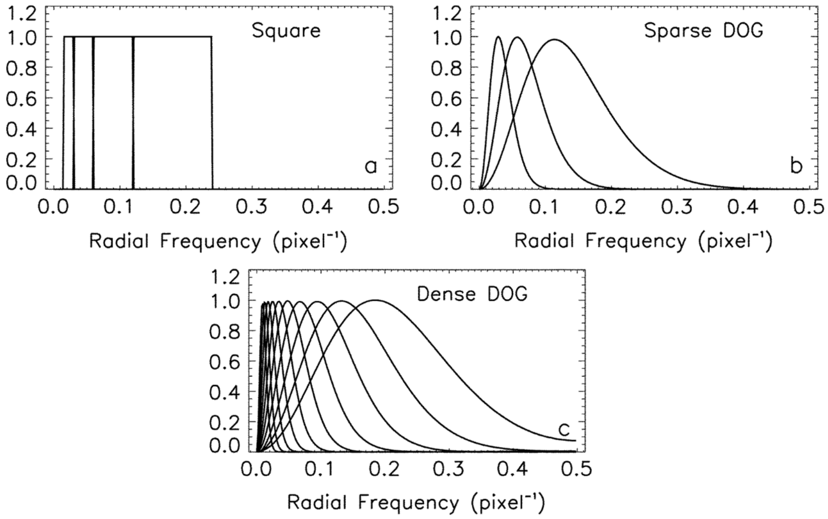

Fig. 2.

Profiles of the rotationally symmetric channels applied with the CNPW observer, (a) Four square-profile channels; (b) the three sparse-DOG channels; (c) the 10 dense-DOG channels.

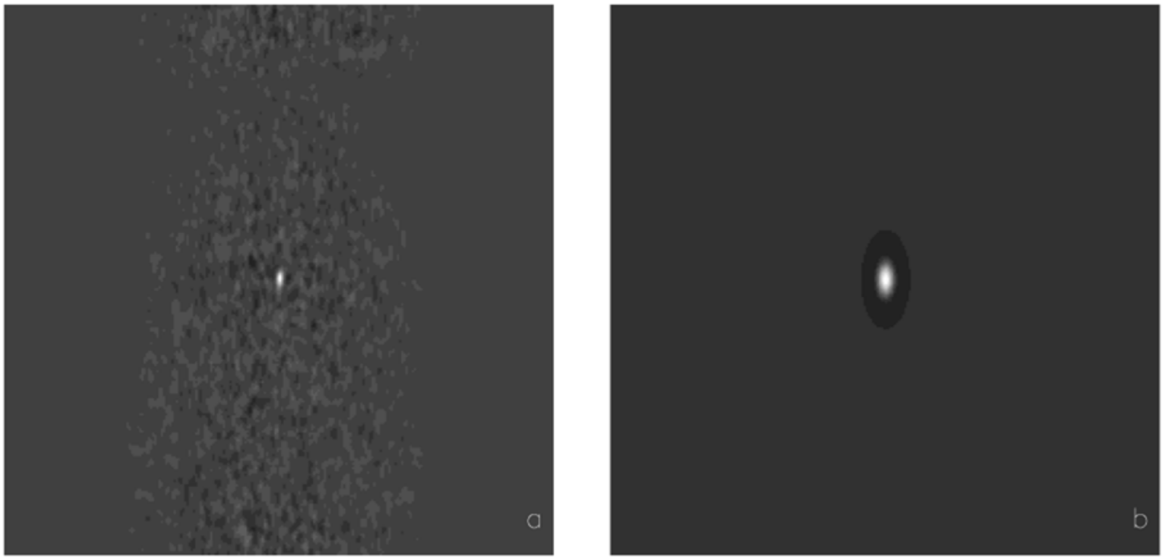

Fig. 3.
(a) The mean lesion in the coronal view as estimated from 42 pairs of tumor-present/tumor-absent images. (b) The response from a nonsymmetric SDOG channel.

$$\mathcal{S} \quad = \quad a_8, a_7, a_6, p_8 \mid a_5, a_4, a_3, p_7 \mid a_2, p_6, p_5 \mid a_1, p_4, p_3, p_2, p_1$$

(a)

	Bin #			
	1	2	3	4
p	1	1	2	4
a	3	3	1	1

(b)

$\Longrightarrow$

	Bin #		
	1	2	3
p	1	3	4
a	3	4	1

(c)

Fig. 4.

Diagram describing the binning procedure, a) Sequence *S* lists the ratings for tumor-present (p_i)and tumor-absent (a_i) images in ascending order. The vertical bars indicate the initial binning defined by the consecutive runs of *p*'s and *a*'s. b) The initial binning in tabular form. c) An iterative rebinning step decrements the number of bins by removing the one with the fewest elements.

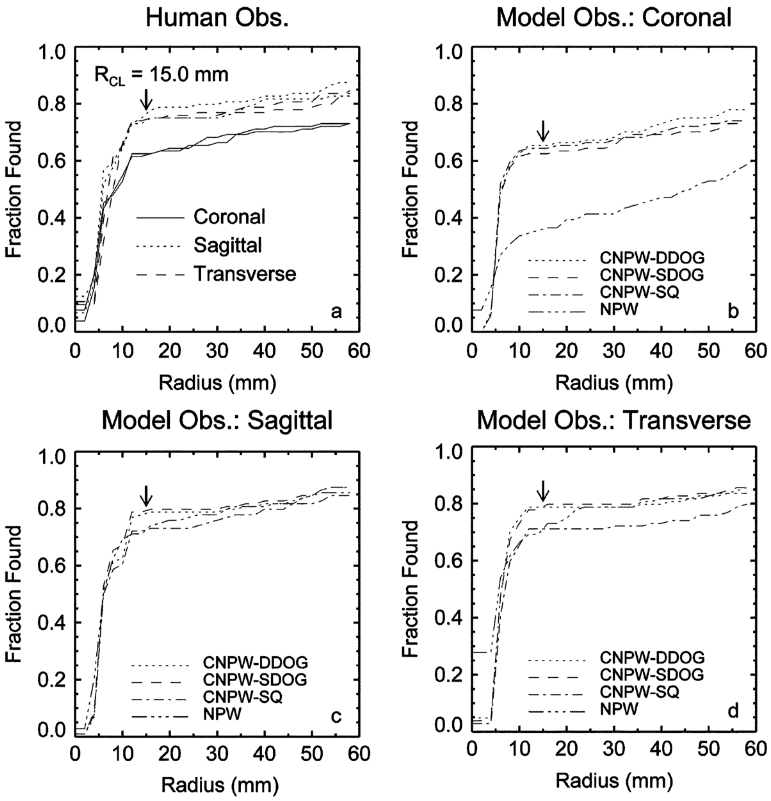

Fig. 5.

The observers' fractions found as a function of radius of correct localization R_{cl}. The selected value of R_{cl} = 15 mm is marked by the arrows. Plot (a) summarizes the localizations for the two observers and three display formats. Each of the plots (b)–(d) summarizes the model-observer localizations for one display format.

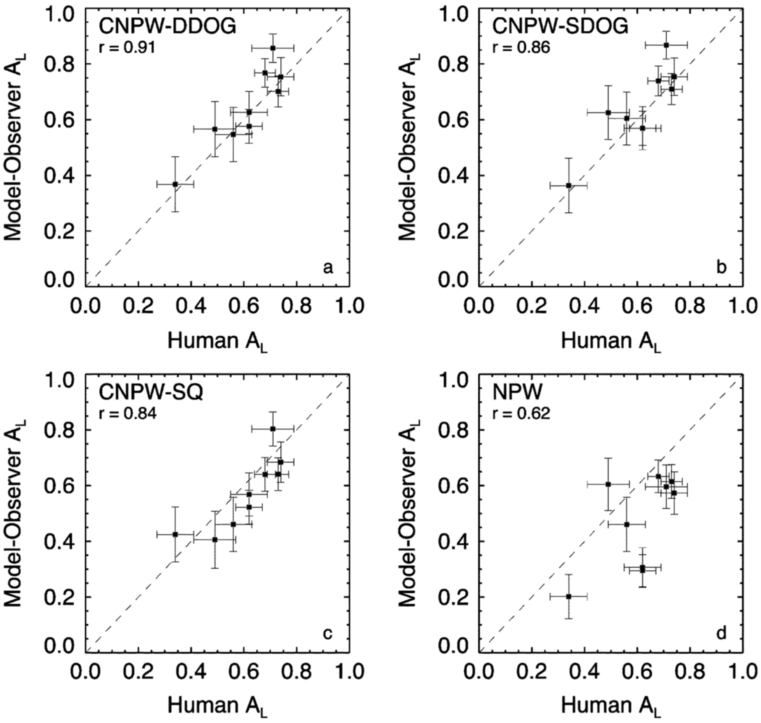

Fig. 6.

Comparison of model-observer and average human performances taken from Table IV. A linear correlation coefficient *r* is given for each set of data. The ideal model-human relationship is represented by the diagonal line.

TABLE I

Tumor Contrasts for the LROC Study. The Five Contrasts Per Organ Were Intended to Produce a Range of Fractions Found [7]. The Number of Cases With a Particular Contrast Is Given in Parentheses

	Fraction Found				
Organ	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
liver	2.5 (5)	3(2)	3.25 (6)	3.75 (3)	4.75 (5)
lung	5.5 (5)	6.5 (11)	7 (10)	7.5 (5)	9 (1)
tissue	6.5 (6)	7 (10)	8 (12)	9 (13)	10.5 (10)

TABLE II

Comparison of Model-Observer and Human-Observer Performances (A_L) With the Three Display Formats. The Uncertainties for the Average Human Scores Account for Observer Variability and Uncertainties in the Data Fitting

	Format		
Observer	Coronal	Sagittal	Transverse
Human observer 1	0.55 ± 0.04	0.69 ± 0.04	0.67 ± 0.04
Human observer 2	0.54 ± 0.04	0.65 ± 0.04	0.68 ± 0.04
Human average	0.54 ± 0.03	0.67 ± 0.04	0.67 ± 0.03
CNPW-DDOG[a]	0.54 ± 0.03	0.71 ± 0.03	0.68 ± 0.03
CNPW-SDOG	0.50 ± 0.03	0.72 ± 0.03	0.68 ± 0.03
CNPW-SQ	0.50 ± 0.03	0.65 ± 0.03	0.60 ± 0.03
NPW	0.25 ± 0.03	0.58 ± 0.03	0.57 ± 0.03
CNPW-DDOG[b]	0.54 ± 0.04	0.72 ± 0.04	0.70 ± 0.04
CNPW-SDOG	0.52 ± 0.04	0.74 ± 0.04	0.71 ± 0.04
CNPW-SQ	0.52 ± 0.04	0.65 ± 0.04	0.61 ± 0.04
NPW	0.27 ± 0.04	0.60 ± 0.04	0.57 ± 0.04

[a] full binning;

[b] partial binning

TABLE III

Wilcoxon Estimates of Model-Observer Performances (A_L) With the Three Display Formats. The Uncertainties Are Based on Resampling Trials

Observer	Format		
	Coronal	Sagittal	Transverse
CNPW-DDOG	0.54 ± 0.04	0.69 ± 0.04	0.69 ± 0.04
CNPW-SDOG	0.52 ± 0.04	0.71 ± 0.04	0.68 ± 0.04
CNPW-SQ	0.52 ± 0.04	0.63 ± 0.04	0.60 ± 0.04
NPW	0.25 ± 0.04	0.57 ± 0.04	0.58 ± 0.04

TABLE IV

Observer performance (A_L) by individual organ for the three display formats

		Format		
Observer	**Organ**	**Coronal**	**Sagittal**	**Transverse**
Human average	liver	0.34 ± 0.07	0.49 ± 0.08	0.56 ± 0.07
	lungs	0.62 ± 0.07	0.71 ± 0.08	0.74 ± 0.05
	tissue	0.62 ± 0.05	0.73 ± 0.04	0.68 ± 0.04
CNPW-DDOG	liver	0.37 ± 0.10	0.56 ± 0.10	0.55 ± 0.10
	lungs	0.62 ± 0.08	0.86 ± 0.05	0.75 ± 0.07
	tissue	0.58 ± 0.06	0.70 ± 0.06	0.77 ± 0.05
CNPW-SDOG	liver	0.36 ± 0.10	0.62 ± 0.10	0.60 ± 0.10
	lungs	0.57 ± 0.08	0.87 ± 0.05	0.75 ± 0.07
	tissue	0.57 ± 0.06	0.71 ± 0.06	0.74 ± 0.05
CNPW-SQ	liver	0.42 ± 0.10	0.41 ± 0.10	0.46 ± 0.10
	lungs	0.57 ± 0.08	0.80 ± 0.06	0.68 ± 0.07
	tissue	0.52 ± 0.06	0.64 ± 0.06	0.64 ± 0.06
NPW	liver	0.20 ± 0.08	0.60 ± 0.10	0.46 ± 0.10
	lungs	0.31 ± 0.07	0.60 ± 0.08	0.57 ± 0.08
	tissue	0.30 ± 0.06	0.61 ± 0.06	0.63 ± 0.06