*Med. Phys.*, vol. 35, no. 12, 2008, in print.

# Model for the detection of signals in images with multiple suspicious locations

Lucreţiu M. Popescu

*University of Pennsylvania, Department of Radiology*

*423 Guardian Drive, 4-th floor Blockley Hall, Philadelphia, PA 19104-6021, USA*[*]

(Dated: September 19, 2008)

We present a model for signal detection that combines a signal model and a noise model providing mathematical descriptions of the frequency of appearance of the signals, and of the signal-like features naturally occurring in the background. We derive expressions for the likelihood functions for the whole ensemble of observed suspicious locations, in various possible combinations of signals and false signal candidates. As a result, this formalism is able to describe several new types of detection tests using likelihood ratio statistics. We have a global image abnormality test and an individual signal detection test. The model also provides an alternative mechanism in which the combination of signal and noise features candidates that has the maximum likelihood is selected. These tests can be analyzed with a variety of operating characteristic curves (ROC, LROC, FROC, etc.). In the mathematical formalism of the model, all the details characterizing the suspicious features are reduced to a single scalar function, which we name the signal specificity function, representing the frequency that a signal takes a certain value relative to the frequency of having a false signal with same value in an image of given size. The signal specificity function ranks the degree of suspiciousness of the features found, and can be used to unify into a single score all the suspicious feature characteristics, and then apply the usual decision conventions as in the Swensson's detection model (*Med. Phys. 23, pp. 1709–1725, 1996*). We present several examples in which these tests are compared. We also show how the signal specificity function can be used to model various degrees of accuracy of the observer's knowledge about image noise and signal statistical properties. Aspects concerning modeling of the human observer are also discussed.

## I. INTRODUCTION

The detection of small abnormal regions in medical images represents a common radiological task, and studying the detectability of small signals in images has become a standard image quality evaluation procedure. However, the problem of detectability evaluation of small signals at unknown locations in random fields has proven to be difficult to address theoretically. Due to the complexity of the way in which images are acquired in modern imaging devices, involving reconstruction algorithms and complex data correction procedures, to make such evaluations directly from the physical characteristics of the imaging device has become very complex. Even for the most simplified theoretical cases, this remains a hard problem, requiring various approximations[1].

A more pragmatic approach for image quality evaluation, that circumvents the theoretical difficulties of the direct estimation methods, involves performing detectability tests using numerical image scanning procedures, or in more clinically realistic conditions, using human observers. Without relying on a specific model of the detection mechanisms used by observers, the results of such tests can be analyzed by using nonparametric methods[2–4]. However, for designing numerical observers tests, or to better understand the decision mechanisms that may be employed by the human observers, we still need models that capture the relevant statistical properties of images and provide means to evaluate the likelihoods for the suspicious feature appearances in the image as ensembles of true and/or false signals.

In this paper, we present such a theoretical model for multiple signal detection in images that adds new results and generalizes our previous work on this topic[5]. As precursors of this work we cite Bunch *et al.*[6] who originally proposed a Poisson statistic for the number of false signals appearing in an image background in their free response operating characteristic (FROC) model, and the works of Swensson[7–9] which proposed decision mechanisms based on the score of the most suspicious location appearing in an image as the basis of a receiver operating characteristic (ROC) and a localization ROC (LROC) model. Here we will describe a generalization to the case when the observed suspicious locations are characterized by multidimensional variables, and we show alternative decision mechanisms that use the likelihoods of the ensemble of suspicious locations observed in an image.

In certain formal aspects, our model is related to the recent work of Chakraborty[10,11] and Edwards *et al.*[12]. However, there are significant differences in scope and interpretation, between our proposal and these works, since the models in these cited papers are limited to specific parameterizations and their decision mechanisms involve latent decision variables whose existence and distributions are postulated. Also, Chakraborty in support of his search model use arguments from a perceptual point of view, while in our derivations bellow, we assume a number statistical properties of the random fields that are of a more general nature. Some aspects concerning these models will be discussed in more detail in section IV of this paper.

## II. THEORY

In an image one can identify zero, one or more suspicious locations. These can be either signals or only signal-like features (false signals) due to image noise. We assume that each such suspicious location can be characterized by a set of numerical values, denoted here by $\mathbf{z}$. In the simplest case, $\mathbf{z}$ can be a single scalar value, for instance the contrast of a small region relative to the surrounding background. In more refined applications, $\mathbf{z}$ may represent a set of measurable values such as contrast, size, shape, and texture pattern. We assume that the suspicious locations, true or false signals, are small in size compared with the image size, and occur independently of each other. One can also assume that the $\mathbf{z}$ values represent the subjective suspicion scores assigned by a human observer. However, the measures in which these scores can be assigned independently of each other is questionable, and will be discussed later in this paper in light of the theoretical results presented here.

In order to study the detectability of the signals, we need to know the frequency of appearance and the distribution of the true signals and of the false, signal-like features, naturally occurring in the background. That is, we need to have a signal and noise model of the image.

### A. The signal model

Let $f(\mathbf{z})$ be the distribution of the signals, where $\mathbf{z}$ can be multidimensional. We denote by $\mathcal{D}_\infty$ the full domain in which $\mathbf{z}$ takes values. We have $\int_{\mathcal{D}_\infty} f(\mathbf{z})\mathrm{d}\mathbf{z} = 1$. Often due to practical reasons the variable $\mathbf{z}$ is observable only if it exceeds certain thresholds. Let $\mathcal{D} \subset \mathcal{D}_\infty$ be such a restricted domain for $\mathbf{z}$, the probability for a signal to have values in $\mathcal{D}$ is $\mu = \int_{\mathcal{D}} f(\mathbf{z})\mathrm{d}\mathbf{z}$. In the case of multiple independent and identical signals per image, the probability of having $l \geq 0$ signals with $\mathbf{z} \in \mathcal{D}$, out of a total of $m$ ($m \geq l$) signals present in an image, is given by the binomial distribution

$$S_m(l;\mathcal{D}) = \binom{m}{l}\mu^l(1-\mu)^{m-l}\,. \tag{1}$$

In general, for a population of signal-present images in which the number of signals per image is distributed following $\phi(m)$, the probability of observing $l \geq 0$ signals with values in $\mathcal{D}$ in an arbitrary signal-present image is

$$S_\phi(l; \mathcal{D}) = \sum_{m \geq l} \phi(m) S_m(l; \mathcal{D}) \,, \tag{2}$$

with $\sum_{m \geq 1} \phi(m) = 1$. We also assume that the signals may occur with equal probability anywhere in the image search area in which we are interested.

### B. The noise model

In conditions of stationarity (the images are acquired in the same conditions and have the same size), the false signals (noise nodules) with values $\mathbf{z} \in \mathcal{D}$ will occur with a constant rate. Assuming that the images are large compared with the signal size and that they occur independently, we can use a Poisson model[6] for the probabilty of having in an image $k$ false signals with values in $\mathcal{D}$:

$$P(k; \mathcal{D}) = \frac{\nu^k}{k!} \mathrm{e}^{-\nu} \,, \tag{3}$$

where $\nu$ is the average number of false signals with value in $\mathcal{D}$. If $\rho(\mathbf{z})$ is the density of these features in the image area (or volume) and in the $\mathbf{z}$-space, then, for an image of size $\Omega$, we have $\nu = \Omega \int_{\mathcal{D}} \rho(\mathbf{z}) \mathrm{d}\mathbf{z}$.

The probability of having at least one false signal with values in $\mathcal{D}$ is

$$Q(1; \mathcal{D}) = 1 - P(0; \mathcal{D}) = 1 - \mathrm{e}^{-\nu} \,. \tag{4}$$

The probability of having at least $k$ false signals with $\mathbf{z} \in \mathcal{D}$ is $Q(k; \mathcal{D}) = 1 - \sum_{i=0}^{k-1} P(i; \mathcal{D})$.

### C. Likelihoods for ensemble of suspicious image locations

Let us assume that we have $k$ suspicious locations with values $Z_k \equiv \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$ in $\mathcal{D}$. If no signal is present in $Z_k$, then the probability density, the likelihood, of such an observation is

$$L(Z_k; 0) = P(k; \mathcal{D}) \prod_{j=1}^{k} p(\mathbf{z}_j) \,, \tag{5}$$

where $p(\mathbf{z})$ is the false signal distribution with $\mathbf{z}$ normalized to the domain $\mathcal{D}$, $\int_{\mathcal{D}} p(\mathbf{z}) \mathrm{d}\mathbf{z} = 1$. We have $p(\mathbf{z}) = \frac{\Omega}{\nu} \rho(\mathbf{z})$.

If, out of the total $k$ nodules in $Z_k$, $l$ are true signals ($l \leq k$) with indices $I_l^k \equiv \{i_1, \ldots, i_l\}$, then the likelihood of such observation is

$$L(Z_k; I_l^k) = P(k-l; \mathcal{D}) \prod_{i \in I_l^k} f(\mathbf{z}_i) \prod_{j \in I_k^k - I_l^k}^{k} p(\mathbf{z}_j)\,, \tag{6}$$

where $f(\mathbf{z})$ is the signal distribution with $\mathbf{z}$, also renormalized for the domain $\mathcal{D}$, $\int_{\mathcal{D}} f(\mathbf{z}) \mathrm{d}\mathbf{z} = 1$.

If $l$ signals are present in $Z_k$, but their indices are unknown, then the likelihood of such observation is given by the average probability density over all undetermined states, that is, all $\binom{k}{l}$ possible combinations in which $l$ signals are present

$$L(Z_k; l) = \frac{P(k-l; \mathcal{D})}{\binom{k}{l}} \sum_{I_l^k} \prod_{i \in I_l^k} f(\mathbf{z}_i) \prod_{j \in I_k^k - I_l^k} p(\mathbf{z}_j)\,, \tag{7}$$

where the sum iterates over all $I_l^k$ combinations.

The likelihood, or the probability density, of $Z_k$, if the suspicious location $i$ is a signal part of any combination of $l$ signals in the $Z_k$ set, is

$$L(Z_k; i, l) = \frac{P(k-l; \mathcal{D})}{\binom{k-1}{l-1}} f(\mathbf{z}_i) \sum_{I_{l-1}^{k,i}} \prod_{j \in I_{l-1}^{k,i}} f(\mathbf{z}_j) \prod_{n \in I_{k-1}^{k,i} - I_{l-1}^{k,i}} p(\mathbf{z}_n)\,, \tag{8}$$

where $I_{l-1}^{k,i}$ is a particular combination of $l-1$ indices $\{i_1, \ldots, i_{l-1}\}$ taken from the set of all $k$ indices minus the $i$ index. The sum in the above equation iterates over all $\binom{k-1}{l-1}$ possible combinations of this kind.

Similarly, the likelihood function when feature $i$ is not among the $l$ signals present in $Z_k$ is

$$L(Z_k; \neg i, l) = \frac{P(k-l; \mathcal{D})}{\binom{k-1}{l}} \sum_{I_l^{k,i}} \prod_{j \in I_l^{k,i}} f(\mathbf{z}_j) \prod_{n \in I_{k-1}^{k,i} - I_l^{k,i}} p(\mathbf{z}_n)\,. \tag{9}$$

By expressing the likelihoods in equations (6), (7), (8) and (9) relative to the likelihood for the signal absent case $L(Z_k; 0)$ given by (5), we obtain the simplified forms

$$\Lambda(Z_k; I_l^k) = \frac{L(Z_k; I_l^k)}{L(Z_k; 0)} = \frac{k!}{(k-l)!} \prod_{i \in I_l^k} \gamma(\mathbf{z}_i)\,, \tag{10}$$

$$\Lambda(Z_k; l) = \frac{L(Z_k; l)}{L(Z_k; 0)} = l! \sum_{I_l^k} \prod_{i \in I_l^k} \gamma(\mathbf{z}_i)\,, \tag{11}$$

$$\Lambda(Z_k; i, l) = \frac{L(Z_k; i, l)}{L(Z_k; 0)} = k \cdot (l-1)! \gamma(\mathbf{z}_i) \sum_{I_{l-1}^{k,i}} \prod_{j \in I_{l-1}^{k,i}} \gamma(\mathbf{z}_j)\,, \tag{12}$$

$$\Lambda(Z_k; \neg i, l) = \frac{L(Z_k; \neg i, l)}{L(Z_k; 0)} = \frac{k \cdot l!}{(k-l)} \sum_{I_l^{k,i}} \prod_{j \in I_l^{k,i}} \gamma(\mathbf{z}_j)\,, \tag{13}$$

where $\gamma(\mathbf{z})$ is defined as

$$\gamma(\mathbf{z}) = \frac{f(\mathbf{z})}{\nu p(\mathbf{z})} = \frac{f(\mathbf{z})}{\Omega \rho(\mathbf{z})} \,. \tag{14}$$

We will call $\gamma(\mathbf{z})$ the signal sensitivity function and it represents the frequency a signal takes the value $\mathbf{z}$ relative to the frequency of having a false signal with the same value in an image of size $\Omega$. The specificity of a signal value decreases with the image search area (or volume). In the relative forms (10)–(13) of the likelihood functions, all the details relating to the suspicious locations are reduced to the $\gamma(\mathbf{z}_i)$ values.

For an arbitrary signal-present image with the number of signals $m \geq 1$ drawn from the distribution $\phi(m)$, the number of signals $l$ with values in the restricted domain $\mathcal{D}$ is a random variable. When $l \geq 0$ is random, the likelihoods for the cases corresponding to equations (6)–(9) are obtained by multiplying them with the number of signals probability distribution $S_\phi(l;\mathcal{D})$ defined in (2). Expressing them relative to $L(Z_k;0)$, we have

$$\Lambda_\phi(Z_k; I_l^k) \;=\; S_\phi(l;\mathcal{D})\Lambda(Z_k; I_l^k) \,, \tag{15}$$

$$\Lambda_\phi(Z_k; l) \;=\; S_\phi(l;\mathcal{D})\Lambda(Z_k; l) \,, \tag{16}$$

$$\Lambda_\phi(Z_k; i, l) \;=\; S_\phi(l;\mathcal{D})\Lambda(Z_k; i, l) \,, \tag{17}$$

$$\Lambda_\phi(Z_k, \neg i, l) \;=\; S_\phi(l;\mathcal{D})\Lambda(Z_k; \neg i, l) \,. \tag{18}$$

### D. Detection tests

#### 1. *Image abnormality test*

The probability density of observing the values in the $Z_k$ set if at least one signal is present in the image is obtained by summing over all possible $l \geq 0$ outcomes each taken with its own probability. Expressing it relative to the likelihood for no signal present case, and using (16), we obtain

$$\Lambda_\phi(Z_k) = \sum_{l=0}^{\infty} \Lambda_\phi(Z_k; l) \,. \tag{19}$$

Equation (19) represents the likelihood ratio statistic for the global image abnormality test case. We have the following decision procedure: for a threshold $\Lambda_d$, if $\Lambda_\phi(Z_k) \geq \Lambda_d$, then the image is positive (at least one signal present), otherwise the image is negative (no signal present). This likelihood ratio statistic gives a global score of image abnormality by taking into account the number of suspicious features and their values, not just the value of the most suspicious feature, as in Swensson's decision mechanism[7]. As shown previously[5] and exemplified again in section III, this likelihood ratio test is more sensitive when multiple signals are present and when the signal values are comparable with those of the background features.

### 2. *Individual signal detection*

The global image abnormality test does not provide a classification of the individual suspicious locations. Such a test statistic is given by comparing the likelihood for the case when a suspicious location $i$ is a signal, with the likelihood for the case when $i$ is just a false signal regardless of whether or not signals are present in the image. By using equations (17) and (18) and by taking into account all possible cases of the number of signal per image, we obtain the following expression of this likelihood ratio

$$\lambda_i(Z_k) = \frac{\sum_{l=1}^{\infty} \Lambda_\phi(Z_k; i, l)}{1 - \eta + \eta \sum_{l=0}^{\infty} \Lambda_\phi(Z_k; \neg i, l)} \,, \tag{20}$$

where $\eta$ is the fraction of abnormal images in the total population of images. For the case when the abnormal images are only a small fraction of the population tested ($\eta \ll 1$), the likelihood ratio is

$$\lambda_i(Z_k) \approx \Lambda_\phi(Z_k; i) = \sum_{l=1}^{\infty} \Lambda_\phi(Z_k; i, l) \,. \tag{21}$$

A test procedure based on the statistic in (20) or (21) uses a constant threshold, all suspicious locations with greater likelihood ratio are considered positive, the remaining being considered negative. An image with no positive result is declared as normal while an image with one or more positive results is declared as abnormal.

### 3. *The most likely image diagnostic*

Instead of sweeping the image using a uniform criterion for signal detection, an alternative approach is given by returning only the most likely answer. That is, out of the $k$ suspicious locations with values in $Z_k$, the indices $I_l^k$ for which the likelihood $\Lambda_\phi(Z_k; I_l^k)$ is maximum are returned as positive results. The list can be void if the likelihood for the no signal case $\Lambda(Z_k; 0)$ has the maximum value. Rather than being a classical hypothesis testing procedure, this approach is in fact a maximum likelihood estimation procedure.

In order to determine the combination of signal and noise feature candidates that is most likely, we can observe, that for a given number $l$ of signals, the maximum likelihood is obtained by the combination containing the first $l$ largest $\gamma(\mathbf{z}_i)$ factors. Therefore, if we sort the list $Z_k$ in a decreasing order so that $\gamma(\mathbf{z}_1) \geq \cdots \geq \gamma(\mathbf{z}_k)$, then, for the first $l$ indices taken as signals we obtain the maximum likelihood $\Lambda_\phi(Z_k; i_1, \ldots, i_l)$ from all possible combinations of $l$ indices. Also, we have the property

$$\Lambda(Z_k; i_1, \ldots, i_l, i_{l+1}) = \Lambda(Z_k; i_1, \ldots, i_l)(k-l)\gamma(\mathbf{z}_{i_{l+1}})\,. \tag{22}$$

for $l \geq 0$. The maximum likelihood for the $l+1$ case can be derived iteratively from the maximum likelihood for the $l$ case. Therefore, in one pass through all relevant $l$ values, the likelihoods can be computed and the case with the maximum likelihood can be identified. This procedure is very fast and avoids a comprehensive search through all $2^k$ possible combinations.

### E. Detectability evaluation using the signal specificity score

The signal specificity function defined in (14) takes nonnegative values for any $\mathbf{z}$ — here we do not require the signal distribution $f(\mathbf{z})$ to be renormalized for a restricted domain. The $\gamma(\mathbf{z})$ values, being scalars with a well defined order relationship, can be used to rank the degree of suspiciousness of the true or false signals found. In this case, we can apply the decision conventions used in the Swensson's model[5,7]. The specificity score $s = \gamma(\mathbf{z})$ can be interpreted as the hypothetical decision variable postulated by such parametric operating characteristic models. The limitations of this approach will be discussed in section IV. Let $s_{\text{max}}$ be the score of the most suspicious location found in an image. With a decision threshold set at $s_d$, if $s_{\text{max}} > s_d$, the image is declared positive (at least one signal present); otherwise it is declared negative (no signal present). If, in the case when $s_{\text{max}} > s_d$, the score $s_{\text{max}}$ corresponds to a true signal, then we also have a correct signal localization. The performance of these tests can be graphically described using relative (receiver) operating characteristic (ROC) and localization ROC (LROC). The additional test in which all features with $s > s_d$ are declared as positive is described by the free response operating characteristic (FROC)[6].

In order to study signal detectability using this decision mechanism, we need to transform the signal distribution $f$ and the false signal density $\rho$ as functions of $s$. For each $s$ value, we can define a domain $\mathcal{D}_s \equiv \{\mathbf{z}|\gamma(\mathbf{z}) \geq s\}$. For each domain $\mathcal{D}_s$, we have a fraction $\mu(s)$ of signals, and an average number $\nu(s)$ of false signals, with $\mathbf{z}$ values satisfying $\gamma(\mathbf{z}) \geq s$. In this manner, we have defined $\mu$ and $\nu$ as function of signal specificity score $s$.

The derivative $f(s) = \frac{\mathrm{d}F(s)}{\mathrm{d}s}$ represents signal distribution with $s$, where the cumulative density function is $F(s) = 1 - \mu(s)$. The density of the false signals with $s$ is given by $\rho(s) = \frac{1}{\Omega}\left|\frac{\mathrm{d}\nu(s)}{\mathrm{d}s}\right|$.

The distribution $g(s)$ of the maximum $s$ score of the false signals for a given search area $\Omega$ is given by the probability density of having a false signal with score $s$, while all other false signals have scores less than $s$, which is equivalent to having a single false signal with value greater than or equal to $s$. We have

$$g(s) = \frac{\Omega\rho(s)}{\nu(s)} P(1; \mathcal{D}_s) = \Omega\rho(s)\mathrm{e}^{-\nu(s)} \tag{23}$$

The reverse cumulative distribution, representing the probability of having at least one false signal with $\gamma(\mathbf{z}) > s$, is

$$Q(s) = \int_s^\infty g(s')\mathrm{d}s' = 1 - \mathrm{e}^{-\nu(s)}\,. \tag{24}$$

The cumulative distribution function of $g(s)$ is $G(s) = 1 - Q(s) = \mathrm{e}^{-\nu(s)}$. Since it is also defined as $G(s) = \int_0^s g(s)\mathrm{d}s = \mathrm{e}^{-\nu(s)} - \mathrm{e}^{-\nu(0)}$, we must have $\mathrm{e}^{-\nu(0)} = 0$, which is obtained for $\nu(0) = \infty$. This is a consequence of having a Poisson model; in reality, due to the finite size of the images, we can have only approximately satisfied $\mathrm{e}^{-\nu(0)} \approx 0$, where $\nu(0)$ is roughly the maximum number of signal sized features that can fit the image.

By using the equations given by Swensson[5,7], we can compute the probability $P_0(s)$ of having at least one false signal with score greater than or equal to $s$, the probability $P_1(s)$ of having a signal present image with score greater than $s$, and the probability $P_{1L}(s)$ of having the most suspicious location correctly declared as a signal (probability of a correct localization). By plotting $P_1(s)$ and $P_{1L}(s)$ against $P_0(s)$, we obtain the ROC and LROC curves, respectively. The FROC curve is obtained by plotting $\mu(s)$ against $\nu(s)$.

In practice the domain of observable variables may be restricted (due to thresholding of the components of the variable $\mathbf{z}$, or prior image processing operations such as filtering). If $s_0 > 0$ is a lower limit of the signal specificity score, then the images having only features with $\gamma(\mathbf{z}) < s_0$ (either signals or noise nodules) will be reported as having no suspicious location. In this case, for $s < s_0$, the signal distribution $f(s)$ and the most suspicious false signal distribution $g(s)$ are not known, or not defined. However, we still have the cumulative distribution functions due to their complementary definitions $F(s) = 1 - \mu(s)$ and $G(s) = 1 - Q(s)$.

When the observed $s$ values are truncated below the $s_0$ limit, we obtain a continuous ROC curve up to the point given by $P_0(s_0) = \mathrm{e}^{-\nu_0}$, $P_1(s_0) = (1 - \mu_0)\mathrm{e}^{-\nu_0}$. We still have the theoretically possible case when all images are declared as positive, which corresponds to a decision threshold set to the lowest limit $s = 0$ and is represented by the upper-left corner of the ROC graph ($P_0 =$

$1, P_1 = 1)$. The portion of the ROC curve past the point $(P_0(s_0), P_1(s_0))$ up to $(1, 1)$ can be obtained by conveniently extrapolating the $f(s)$ and $\rho(s)$ distributions for the $0 < s < s_0$ region. In this fashion extensions of the LROC and FROC curve are also obtained, consistent with the ROC curve extension.

## III. SIMULATION TESTS

In order to illustrate the theory presented above, we have considered several examples with parameterization values derived from our previous work[5] employing simulated two-dimensional positron emission tomography (PET) images. Uniform 1 cm diameter discs were considered as signals (nodules). As the measured signal metric $z$, we took the one-dimensional contrast value given by the ratio between the average image intensities within the disk of the same size as the signal, and the surrounding background. The value of $z = 1$ corresponds to the average contrast of a background region, while $z = 0$ corresponds to the limiting case of an empty 1 cm diameter region. In this examples, the signal distribution with $z$ is well modeled by a Gaussian $f(z) = \frac{1}{N_f} \exp\left[-\frac{1}{2}\left(\frac{z-\mu_f}{\sigma_f}\right)^2\right]$, where $\mu_f$ is the average contrast value, $\sigma_f$ is the standard deviation, and $N_f$ is a normalization constant. The false signal (noise nodule) distribution is also well fitted by the tail of a Gaussian, $p(z) = \frac{1}{N_p} \exp\left[-\frac{1}{2}\left(\frac{z-\mu_p}{\sigma_p}\right)^2\right]$, with $\mu_p$ and $\sigma_p$ being parameters, and $N_p$ a normalization constant. Both distributions are normalized for $z > z_0$ with the threshold $z_0 = 1.2$. As parameters, we also have the image area $\Omega = 100\text{cm}^2$, $\nu_0 = \nu(z_0)$ the average number of noise nodules in the given area (with $z > z_0$), the distribution $\phi(m)$ of the number of signals in the images containing signals, and $\eta$ the average fraction of signal present images among the total number of image samples.

We consider that the above model and parameters give a description of the true statistical behavior of the imaging system. However, depending on its training and his prior experience, an observer can have only a partial knowledge about the true values of these parameters and distribution shapes, and he may use different distributions and with different parameters in the process of evaluating the images and in decision making. For this reason, for some of the examples below, we will use two sets of parameters, one corresponding to the true physical and statistical model of the images, which is used in the Monte Carlo simulation of the false and true signals, and the other corresponding to the actual subjective knowledge of the observer about the statistical properties of the images, which is used in the image evaluation equations.

TABLE I: Model parameters used in the example presented. In the cases where different values have been used for a parameter for data generation and evaluation, the values are presented in the form $a/b$.

| Parameter | Exp. I | Exp. II | Exp. III |
|---|---|---|---|
| $\nu_0$ | 29 | 29 | 22 |
| $\mu_p$ | 1.0 | 1.0 | 1.0 |
| $\sigma_p$ | 0.35 | 0.35 | 0.26 |
| $\mu_f$ | 1.7 | 1.7 | 1.7 |
| $\sigma_f$ | 0.059 | 0.059/0.15 | 0.21 |
| $\eta$ | 0.1 | 0.1 | 0.1 |
| $\phi(1)$ | 0.33 | 0.33 | 0.33 |
| $\phi(2)$ | 0.33 | 0.33 | 0.33 |
| $\phi(3)$ | 0.33 | 0.33 | 0.33 |

### A. Example I: low contrast signals with good specificity

In order to illustrate the importance of good signal specificity, rather than directly the importance of good signal contrast, first we present a case where signals have relatively a small contrast — on average less than the average contrast of the maximum noise nodule — but confined to a relatively narrow region. The model parameters used in this case are shown in TABLE I. The parameters used for image evaluation are identical to the true values used for generating the data; thus, here we are emulating the case of a perfectly trained observer. A comparison between the signal distribution $f(z)$, the noise nodule distribution $p(z)$ and the max-scan distribution $g(z)$ (the distribution of the maximum $z$ value of the noise nodules), as well as a plot of the signal specificity function $\gamma(z)$ are shown in FIG. 1.

The simulations have been carried out by refining the procedures used in our previous works[4,5]. The sets of false signals (noise nodules) and true signals (nodules), sampled according to the models described above, have been randomly generated using the alias sampling technique[13,14]. Then each set was evaluated by using the methods described in section II D as well as by applying the Swenson's conventions for the maximum contrast $z$ and for the maximum signal specificity score $\gamma(z)$, as discussed in section II E.

For global image abnormality test evaluation we have compared five methods. The first method uses the statistic $\Lambda_\phi(Z_k)$ given by the equation (19). The second method uses the maximum

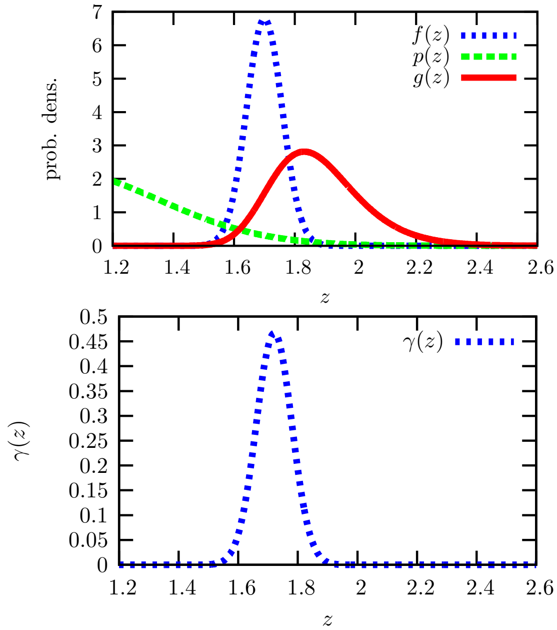


FIG. 1: The signal distribution $f(z)$, the noise nodule distribution $p(z)$, and the distribution of the maximum noise nodule distribution $g(z)$ used in Example I (top). In the bottom graphic is plotted the signal specificity function $\gamma(z)$.

individual score obtained by a suspicious feature using the $\lambda_i(Z_k)$ statistic as discussed in section II D 2. The third method uses the value $\Lambda_\phi(Z_k; I_l^k)$ of the most likely signal and noise nodule combination, as discussed in section II D 3. The fourth method uses the maximum contrast $z$ value found, and the fifth method uses the maximum signal specificity score $\gamma(z)$ found. A comparison of the ROC curves produced by all these methods is shown in FIG. 2.

A second series of comparisons are made by using the free response approach, in which all suspicious features with test scores above a given threshold are returned as positive signals. We have considered four methods. In the first, we use the $\lambda_i(Z_k)$ statistic as discussed in section II D 2. In the second method, we also used the $\lambda_i(Z_k)$ statistic, but limited to the set of features $I_l^k$ of the most likely signal and noise nodule combination, with maximum $\Lambda_\phi(Z_k; I_l^k)$, as discussed in section II D 3. The third method uses the contrast $z$ values, and the fourth method uses the signal specificity score $\gamma(z)$. A comparison of the FROC curves produced by these methods is shown in FIG. 3.

In this example, as we can see in FIG. 2, using directly the contrast $z$ leads to a very poor indicator of image abnormality, barely better than the chance line. Also, in FIG. 3, we can see

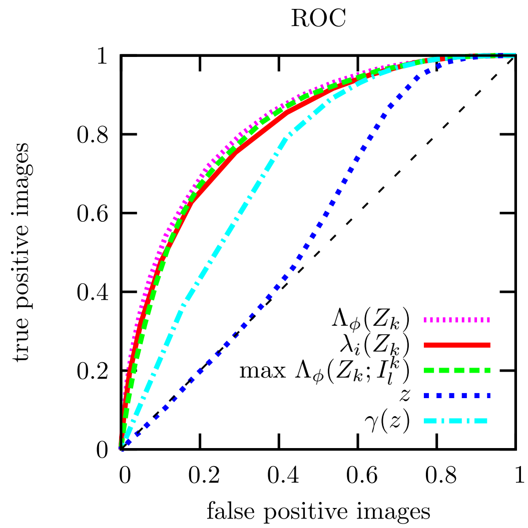


FIG. 2: Comparison of the ROC curves obtained for Example I by five different methods of assigning the image global abnormality score. The first method is based on the likelihood ratio statistic $\Lambda_\phi(Z_k)$, given by equation (19). The second method uses the maximum individual likelihood ratio $\lambda_i(Z_k)$ as discussed in section II D 2. The third method uses the value $\Lambda_\phi(Z_k; I_l^k)$ of the most likely signal and noise nodule combination, as discussed in section II D 3. The fourth method uses the maximum contrast $z$ value found, and the fifth method uses the maximum signal specificity score found $\gamma(z)$.

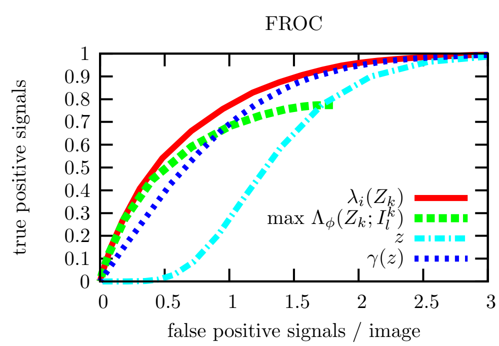


FIG. 3: Comparison of the FROC curves obtained with four different methods for Example I. In the first method, the $\lambda_i(Z_k)$ statistic is used, as discussed in section II D 2. In the second method, also the $\lambda_i(Z_k)$ statistic is used, but limited to the set of features $I_l^k$ of the most likely signal and noise nodule combination ($\max \Lambda_\phi(Z_k; I_l^k)$), as discussed in section II D 3. The third method uses the contrast $z$ values, and the fourth method uses the signal specificity score $\gamma(z)$.

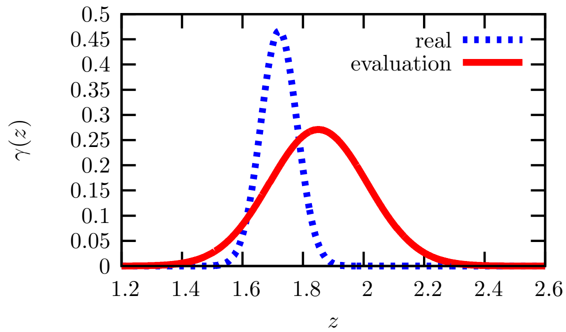


FIG. 4: Comparison between the real signal specificity function and the specificity function used for image evaluation in Example II.

that the test with contrast $z$ score, on average, returns true positive signals only after it has already returned a false positive signal. The signal specificity score represents a better indicator, but it is outperformed by the methods using the likelihood ratio applied either for the global image or to individual suspicious features. The better performance of the likelihood ratio methods is explained by the fact that they take into account the whole ensemble of suspicious features present in an image. The most likely combination approach performs well in the high specificity low sensitivity operating region, but underperforms the other methods in the low specificity high sensitivity region, where its FROC curve ends at about two false positive signals per image.

### B. Example II: small signal hidden in noise, with good real specificity, but poor specificity used in evaluation

In the second example, we consider the same parameterization as in the first example, with the difference that, for image evaluation, a less narrow signal distribution, hence a wider specificity function, is used. The model parameters used in this example are summarized in the Exp. II column of TABLE I. The real signal specificity function and the one used for image evaluation are shown for comparison in FIG. 4.

In FIG. 5 and FIG. 6, the ROC and FROC plots are shown, respectively, for the same methods as in Example I. Comparisons of the ROC curves for the tests using the global likelihood ratio statistic $\Lambda_\phi(Z_k)$ and the maximum specificity score $\gamma(z)$ are shown in FIG. 7. Comparisons of the FROC curves for the free response tests using the likelihood ratio statistic for individual suspicious locations $\lambda_i(Z_k)$ and for the test using the specificity score $\gamma(z)$ are shown in FIG. 8. In these figures, one can notice the degradation of the detectability performance for the case when the

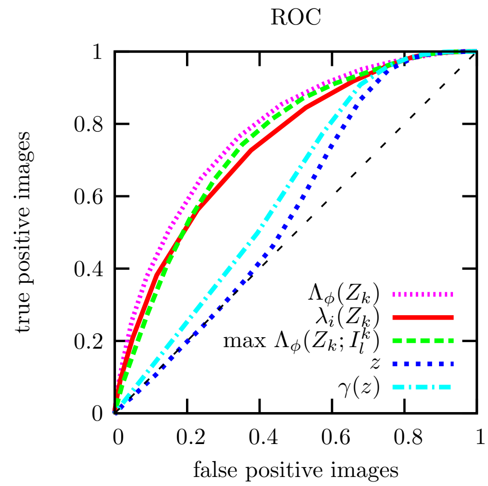


FIG. 5: Comparison of the ROC curves for Example II. The significance of the curves is the same as in FIG. 2.

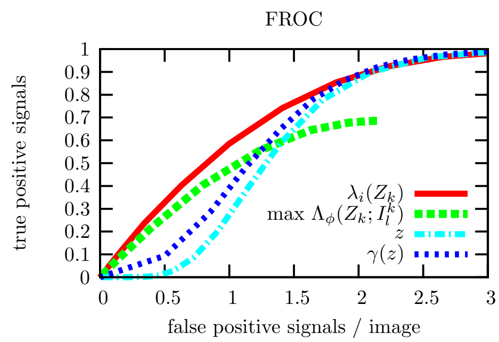


FIG. 6: Comparison of the FROC curves for Example II. The significance of the curves is the same as in FIG. 3.

observer knowledge about the signal distribution is less accurate. This degradation is particularly pronounced for the tests that apply directly the $\gamma(z)$ statistic.

### C. Example III: High signal relative to noise

In a third example, we present a more usual case when the signal contrast distribution is higher on average than the distribution of the maximum noise nodule found in an image. The model parameters used in this example are summarized in the Exp. III column of TABLE I. The signal distribution $f(z)$, the noise nodule distribution $p(z)$, and the distribution of the maximum noise nodule $g(z)$, as well as a plot of the signal specificity function are shown in FIG. 9. In this case,

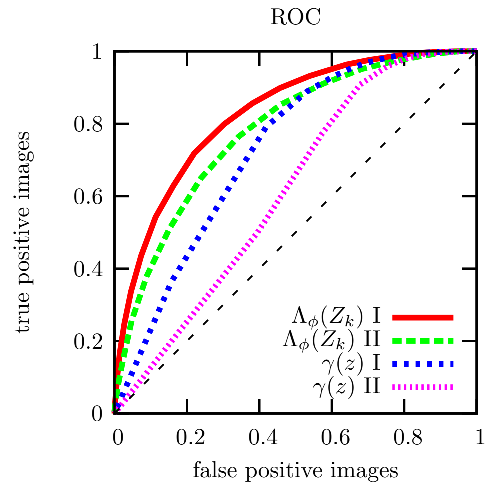


FIG. 7: Comparisons between the ROC curves in examples I and II for the tests using the global likelihood ratio statistic $\Lambda_\phi(Z_k)$ and the maximum specificity score $\gamma(z)$.

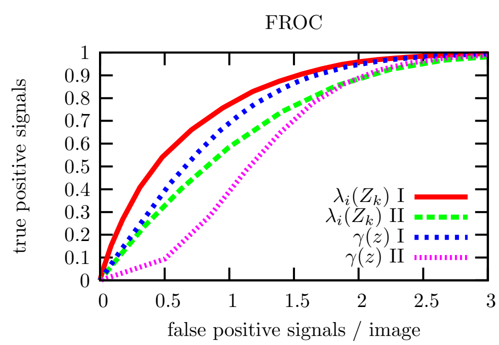


FIG. 8: Comparisons between the FROC curves in examples I and II for the free response tests using the likelihood ratio statistic for individual suspicious locations $\lambda_i(Z_k)$ and for the test using the specificity score $\gamma(z)$.

the signal specificity function $\gamma(z)$ is monotonic with the contrast $z$, therefore they act identically as indicators of location suspiciousness.

From the ROC curves shown in FIG. 10, one can see that the tests using $z$ or $\gamma(z)$ perform identically and the likelihood ratio test is only slightly better. The FROC curves comparison, shown in FIG. 11, reveals better performance of the individual signal likelihood ratio statistic $\lambda_i(Z_k)$. In FIG. 11, one can also notice that the FROC curve of the most likely diagnostic approach ends near the point where on average only one false positive signal per image is returned.

The only slightly better performance of the likelihood ratio test in FIG. 10 is explained in this

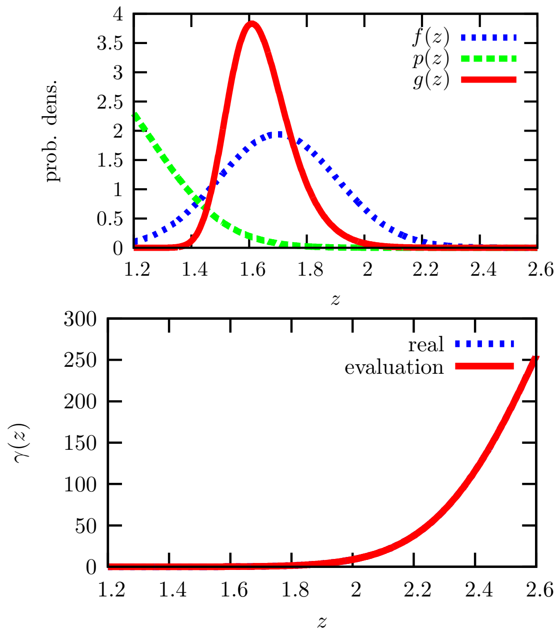


FIG. 9: Signal distribution $f(z)$, noise nodule distribution $p(z)$, and distribution of the maximum noise nodule distribution $g(z)$ used in Example III (top). In the bottom graph the signal specificity function $\gamma(z)$ is plotted.

case by the relatively reduced importance the number of signals per image has as an indicator of image abnormality compared to the individual suspicious location contrast. This can be better seen in FIG. 12 where these results are compared with the curves obtained for a modified $\phi(m)$ distribution in which almost all signal-present images contain three signals ($\phi_2(1) = 0.03$, $\phi_2(2) = 0.03$, $\phi_2(3) = 0.96$). The improvement obtained for the likelihood ratio statistic $\Lambda_\phi(Z_k)$ over the statistic $\gamma(z)$ for $\phi_2(m)$ is more pronounced than in the case of the uniform $\phi(m)$ distribution. A comparison of the FROC curves for these cases is shown in FIG. 13.

## IV. APPLICATION TO HUMAN OBSERVER MODELING

In radiology, particular attention is given to the ability of signal detection theories to model the human observer behavior. From this perspective, the detection model presented here, similarly to other recent models[10,12], can be seen as a two-step process. In the first step, the observer identifies the locations that appear as suspicious, followed by a decision step in which the confidence rankings are assigned according to a certain strategy. The first step can be seen as one involving minimal evaluation of the features characteristics, while a more detailed examination is performed

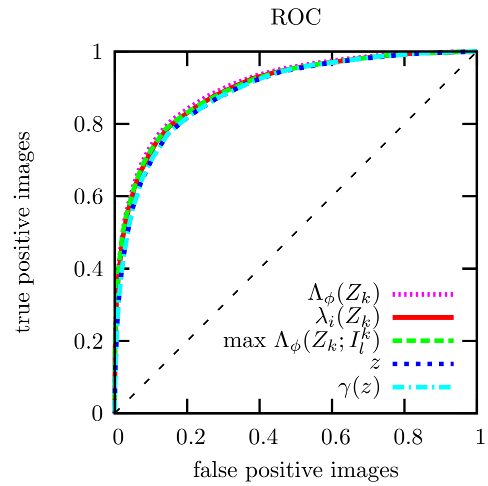


FIG. 10: Comparison of the ROC curves for the Example III. The significance of the curves is the same as in FIG. 2.

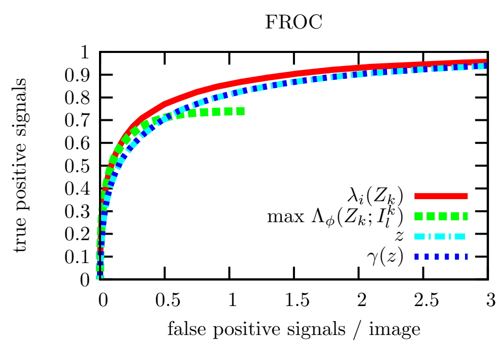


FIG. 11: Comparison of the FROC curves for the Example III. The significance of the curves is the same as in FIG. 3.

in the second step together with comparisons of the various locations and decision making. The whole process is influenced by the training, the prior experience, other relevant information, as well as the goals of the observer.

Eye movement studies[15] have shown that the human observers diagnosing mammograms unevenly focus their gaze on an image, paying more attention to a small number of localized spots, generally corresponding to suspicious locations, from which the final positive results are reported. These experiments also show that the trained radiologists are more accurate in their initial selection of the suspicious locations than lay observers. These experiments support the theories in which multiple local features are evaluated and compared, as opposed to the classical ROC model[16], which directly assumes a global image score without a specific and detailed mechanism of how

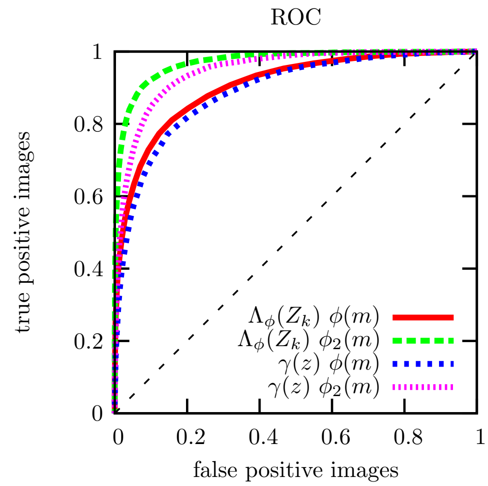


FIG. 12: Comparison of the ROC curves obtained in the Example III with the results obtained for a modified distribution of the number of signals per signal-present image $\phi_2(m)$, in which almost all signal-present images contain three signals.

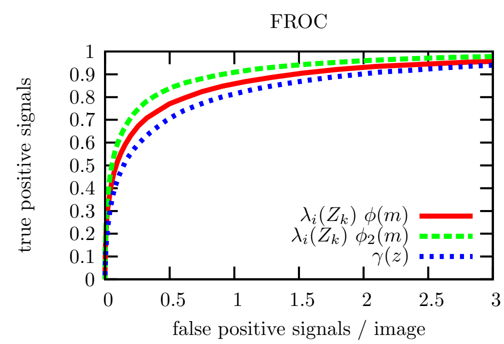


FIG. 13: Comparison of the FROC curves obtained in the Example III with the results obtained for a modified distribution of the number of signals per signal-present image $\phi_2(m)$, in which almost all signal-present images contain three signals.

this score is obtained from the multitude of suspicious features an image may have. However, these experiments give little indication about the decision mechanisms used by the image readers, and how to express quantitatively the reported confidence scores as a function of the evaluated features. On the other hand, cognitive science studies have shown that humans have a certain sense of built-in probability representation, being able to apply accurate underlying probability distributions when asked to predict the outcomes of various random variables derived from daily experiences[17]. Similarly, we can assume to be valid the hypothesis that, when asked to detect small nodules on noisy backgrounds, the human observers will also tend to adopt an optimal strategy according to their prior assumptions. The degree to which the human observer behavior relates quantitatively to the detection mechanisms presented in this paper, remains to be investigated in further studies. Until such studies are performed, we may speculate and make the following observations corroborating existing experimental observations.

An important problem in modeling the human observer is the independence and stationarity of the rankings reported. This depends on what is assumed to be their nature. If the observers assign the scores taking into account only the local characteristics of the features, then we can assimilate these marks with the variable $\mathbf{z}$ or $\gamma(\mathbf{z})$ of our model and assume that they are independent and stationary with respect to the number of signals present per image. However, if the marks are interpreted as confidence scores[10], or probabilities, they can be the result of a global evaluation of the image features, as in our likelihood ratio tests. In other words, even if the signals and the noise features occur independently, they are evaluated globally; hence their scores are not independent. The presence of various numbers of signals per image, combined also with the observer prior assumptions about the frequency of these signals, leads to violations of the stationarity assumption. The score assigned for a feature will depend on what is happening elsewhere in the image. In his LROC paper[7] Swensson shows examples of such stationarity violations and expresses reservation about the applicability of his assumptions for the multiple signals per image case. Another well documented case of stationarity violation is the satisfaction of search effect[18].

Another problem, depending on the decision mechanism used, is the behavior and the significance of the rankings assigned by the human observers in the low confidence region. If we assume that these rankings are confidence scores that cover the full range of possible values, then in a free response test at progressively lower values, we should have an increased number of reports until each region of the image is marked as suspicious. Similarly, in an image abnormality (ROC) test, the results should progressively reach a point when all images are declared positive, and in an LROC test we should always have a location marked regardless of how low its probability is perceived to be. However, in regular experiments, this behavior is not displayed. In one way, this can be explained by assuming that the observers stop making markings below a certain "reasonable" threshold. This hypothesis can be applied both with the likelihood ratio based tests, described in II D, and with the Swensson type tests, as discussed in II E. However, the problem of how this threshold is set is left open. This selective behavior of the observer can also be explained by the most likely image diagnostic mechanism, or a similar approach minimizing a cost function, without further need to define an arbitrary threshold. In the Chakraborty's search model[10], zero suspicious locations results are allowed by not constraining the noise and the signal distributions to be normalized to unity while the model postulated decision variable is let to vary form $-\infty$ to $+\infty$. This leads to truncated operating characteristic curves that the author corrects by adding Dirac's $\delta$ functions "centered at $-\infty$" to the signal and noise distributions[19]. If rigorously analyzed, this still leaves undetermined the upper-left part of the ROC plots. The approach discussed in section II E represents a more logically consistent alternative by considering truncated distributions instead.

Another problem concerning modeling of the human observer operating in realistic conditions is the parametric representation of the signal and of the noise distributions. Given the high variability of the patient images with nonuniform backgrounds and not very well specified signals, the distribution of the reported results may follow nonstandard function shapes difficult to parameterize. This problem may be complicated by the role played by the observer's prior experience and training, or the external information available, or other various constraints under which the observer operates.

The interpretation of the operating characteristic curves (ROC, LROC, FROC) depends on an understanding of the detection process used by the observers. Sometimes idealizing, in some early interpretations of the ROC plots in particular, it was assumed that an ROC curve represents a global characteristic in which different observers operate at different points according to their subjective degree of precaution or aggressiveness. This could be justified if the observers make their judgment based on a dominant physical characteristic of the image features that is monotonic with their degree of suspicion. However, as we have shown in the examples presented here, if the observable characteristics of the suspicious locations do not increase monotonically with the degree of suspicion, or these characteristics are multidimensional (size, shape, texture, etc), there is not a direct objective way in which these characteristics can be unified into a unique variable for all observers. The observer's decision involves assessing how often each feature may occur as a signal, or naturally as noise, which is a subjective evaluation and depends on the observer's prior experience and training (in our model this stands for the observer assessment of the $\gamma(\mathbf{z})$ and $\phi(m)$ distributions). Hence, each observer can have his own characteristic curve, and the results provided by a set of readers may lie in a band of operating characteristic curves and not on a unique curve, the upper side of this band being bordered by the characteristic curve corresponding to the optimal test strategy. This type of behavior has been observed experimentally[20] and has lead to less idealistic interpretations of the ROC methodologies as pragmatic means for statistical comparison of different imaging methods[21].

## V. CONCLUSIONS

The model presented in this paper is composed of a signal model and a noise model that provide mathematical descriptions of the frequency of appearance of the true signals and of the signal-like features naturally occurring in the background, respectively. When combined, they provide expressions of the likelihood functions for the whole ensemble of observed suspicious features, including all various combinations of possible signal and noise feature candidates. As a result, this formalism is able to describe several types of detection tests. We have an image abnormality test produced by taking as decision criterion the ratio between the likelihoods of the case of at least one signal present and the case of no signal present. We have an individual signal detection test that gives prescriptions regarding the classification of each individual suspicious location by using the likelihood ratio for individual features. Also we have the alternative detection mechanism in which the most likely image diagnostic is selected. This test can be seen as a particular case from a larger class of decision making methods, in which, for each possible outcome, a distinct cost is assigned, the result that minimizes the total cost being selected in the end.

The theory presented here is derived from a few simple assumptions about the signal and image background noise properties. The most important assumption is the independence of the signals and of the background noise features, which is satisfied if the signals are small and the suspicious noise features tend to occur at distances large enough at which the image autocorrelations are negligible. Although, in our examples, for convenience, we have used images with uniform background, this is not strictly necessary for the validity of this theory. The variable $\mathbf{z}$ can include the position coordinates, and $f(\mathbf{z})$ and $\rho(\mathbf{z})$, and implicitly $p(\mathbf{z})$ and $\gamma(\mathbf{z})$, can be position dependent. As $\mu$ and $\nu$ are integral values, the binomial (1) and the Poisson (3) distributions will not be affected by this change. A more detailed description of the theory with position dependent distributions we will be provided in a future paper.

Also the results presented here can be generalized by including in the suspicious feature characterization variable $\mathbf{z}$ information from external tests, or merging data from images obtained with other modalities. Moreover, as the patients' lesions are not necessarily independent, the expected number of signals and their characteristics could be made to depend on the stage of the disease. These are aspects that will be addressed in further refinements of this model.

Relating to the relevance of this model for the interpretation of human observer results, we can point out that even within the simplified framework built upon the model assumptions, the theory developed here reveals a larger variety of possible detection strategies than allowed by other human observer models that include search mechanisms[7,10,12]. Also, this theory is able to relate in a single test statistc the observer knowledge about signal appearance, the image noise patterns, and the expectations about the number of signals per image. As discussed in section IV, the measure in which the human behavior can be approximated by such statistics needs to be tested in further experiments.

**Acknowledgments**

The author expresses his thanks to Dr. Jayaram K. Udupa and Dr. Robert M. Lewitt for their suggestions and help with editing of the manuscript.

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health under grant R21-EB005434.

---

* Electronic address: popescu@mail.med.upenn.edu

1 R. J. Adler, "On excursion sets, tube formulas and maxima of random fields," *Ann. Appl. Probab.*, vol. 10, no. 1, pp. 1–74, 2000.

2 J. A. Hanley and B. J. McNeil, "The meaning and use of the area under receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, Apr. 1982.

3 D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis, and validation," *Med. Phys.*, vol. 31, no. 8, pp. 2313–2330, 2004.

4 L. M. Popescu, "Nonparametric ROC and LROC analysis," *Med. Phys.*, vol. 34, no. 5, pp. 1556–1564, 2007.

5 L. M. Popescu and R. M. Lewitt, "Small nodule detectability evaluation using a generalized scan statistic model," *Phys. Med. Biol.*, vol. 51, no. 23, pp. 6225–6244, 2006.

6 P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "Free-response approach to the measurement and characterization of radiographic-observer performance," *Journal of Applied Photographic Engineering*, vol. 4, no. 4, pp. 166–171, 1978.

7 R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.*, vol. 23, no. 10, pp. 1709–1725, 1996.

8 R. G. Swensson, "Using localization data from image interpretations to improve estimates of performance accuracy," *Medical Decision Making*, vol. 20, no. 2, pp. 170–185, 2000.

[9] R. G. Swensson, J. L. King, and D. Gur, "A constrained formulation for the receiver operating characteristic (ROC) curve based on probability summation," *Med. Phys.*, vol. 28, pp. 1597–1609, Aug. 2001.

[10] D. P. Chakraborty, "A search model and figure of merit for observer data acquired according to the free-response paradigm," *Phys. Med. Biol.*, vol. 51, pp. 3449–3462, 2006.

[11] D. P. Chakraborty and H.-J. Yoon, "Operating characteristics predicted by models for diagnostic tasks involving lesion localization," *Med. Phys.*, vol. 35, no. 2, pp. 435–445, 2008.

[12] D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, "Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model," *Med. Phys.*, vol. 29, no. 12, pp. 2861–2870, 2002.

[13] A. J. Walker, "An efficient method for generating discrete random variables with general distributions," *ACM Transactions on Mathematical Software*, vol. 3, no. 3, pp. 253–256, 1977.

[14] L. M. Popescu, "An extension of alias sampling method for parametrized probability distributions," *J. Comp. Phys.*, vol. 160, pp. 612–622, 2000.

[15] H. L. Kundel, C. F. Nodine, and E. A. Krupinski, "Searching for lung nodules: visual dwell indicates location of false-positive and false-negative decisions," *Invest. Radiol.*, vol. 4, pp. 472–478, 1989.

[16] C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[17] T. L. Griffiths and J. B. Tenenbaum, "Optimal predictions in everyday cognition," *Psychological Science*, vol. 17, no. 9, pp. 767–773, 2006.

[18] K. S. Berbaum, E. A. Franken Jr., and D. D. Dorfman et al., "Satisfaction of search in diagnostic-radiology," *Invest. Radiol.*, vol. 25, no. 2, pp. 133–140, 1990.

[19] D. P. Chakraborty, "ROC curves predicted by a model of visual search," *Phys. Med. Biol.*, vol. 51, pp. 3463–3482, 2006.

[20] C. Beam, P. M. Layde, and D. C. Sullivan, "Variability in the interpretation of screening mammograms by US radiologists," *Arch. Intern. Med.*, vol. 156, pp. 209–213, 1996.

[21] R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: A tutorial review," *Acad. Radiol.*, pp. 723–748, 2007.