INSTITUTE OF PHYSICS PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **51** (2006) 3449–3462 doi:10.1088/0031-9155/51/14/012

A search model and figure of merit for observer data acquired according to the free-response paradigm

D P Chakraborty

Department of Radiology, University of Pittsburgh, 3520 5th Avenue, Suite 300, Pittsburgh, PA 15261, USA

E-mail: dpc10@pitt.edu

Received 29 January 2006, in final form 2 June 2006
Published 6 July 2006
Online at stacks.iop.org/PMB/51/3449

Abstract

Search is a basic activity that is performed routinely in many different tasks. In the context of medical imaging it involves locating lesions in images under conditions of uncertainty regarding the number and locations of lesions that may be present. A search model is presented that applies to situations, as in the free-response paradigm, where on each image the number of normal regions that could be mistaken for lesions is unknown, and the number of observer generated localizations of suspicious regions (marks) is unpredictable. The search model is based on a two-stage model that has been proposed in the literature, according to which, at the first stage (the preattentive stage) the observer uses mainly peripheral vision to identify likely lesion candidates, and at the second stage the observer decides (i.e., cognitively evaluates) whether or not to report the candidates. The search model regards the unpredictable numbers of lesion and non-lesion localizations as random variables and models them via appropriate statistical distributions. The model has three parameters quantifying the lesion signal-to-noise ratio, the observer's expertise at rejecting non-lesion locations, and the observer's expertise at finding lesions. A figure-of-merit quantifying the observer's search performance is described. The search model bears a close resemblance to the initial detection and candidate analysis (IDCA) model that has been recently proposed for analysing computer aided detection (CAD) algorithms. The ability to analytically model and quantify the search process would enable more powerful assessment and optimization of performance in these activities, which could be highly significant.

Introduction

Search is a ubiquitous activity that is performed routinely in many different tasks ranging from searching for foreign objects in an airport baggage-screening display, searching for signs of cancer in a mammogram, computer aided detection algorithms that seek to detect lesions in order to assist radiologists and internet search engines capable of searching specified content

0031-9155/06/143449+14$30.00

and retrieving only relevant material. The ability to analytically model and quantify the search process would enable more powerful assessment and optimization of performance in these activities, which could be highly significant (Wolfe 2005). The topic of searching, and the associated topic of determining if an image has a target(s), has generated considerable psychophysical (Treisman and Gelade 1980, Treisman and Gormican 1988, Palmer *et al* 2000, Wolfe 1998) and medical imaging literature (Metz 1986, 1989). Past research has mostly focused on *detection*, defined as the observer's ability to correctly classify an image as target-containing or target-absent, and this ability is generally measured with the receiver operating characteristic (ROC) paradigm, which is widely used in both psychophysical (Rotello *et al* 2004) and medical imaging research (Wagner *et al* 2002).

In this paper I confine myself to visual searching of medical images and to situations where the number of normal regions that could be mistaken for lesions (distracters) is unknown. For example, in screening mammography the radiologist does not know *a priori* whether a lesion is present in an image and therefore must search the image for possible lesions. This simple statement of the radiologist's task masks the difficulty of modelling it. The number and locations of the distracters are unknown to the experimenter and indeed these are expected to vary between images and between radiologists. On some images the radiologist may find nothing to report while on others one or more regions that resemble lesions may be reported. The record of locations ('marks') found to be sufficiently suspicious to deserve reporting, and the corresponding confidence levels ('ratings') that they represent lesions, constitutes search data and equivalent information is routinely entered into the radiologist's clinical report.

The radiologist's task described above is identical to the free-response paradigm (Bunch *et al* 1978), in which the observer marks and rates suspected lesion locations. The rating is a number representing the degree of confidence that the marked location is actually a lesion, e.g., 1, 2, 3, 4 in a 4-rating free-response study. It is assumed that the number and locations of any lesions that are present (i.e., the gold standard) are available to the experimenter.

Analysis of mark-rating data requires *scoring* the data, i.e., each mark has to be classified as a lesion localization or a non-lesion localization. This is done by adopting an acceptance radius and classifying a mark that is within an acceptance radius of the centre of a lesion as a lesion localization and all other marks are classified as non-lesion localizations. In the literature the term 'detection' is occasionally used to mean either correct classification of an image or correct localization of a lesion. To avoid confusion I use the term 'detection' to mean correct classification, and the term 'lesion localization' to describe the correct identification of a particular region of the image as a lesion. The free-response receiver operating characteristic (FROC) curve is defined (Bunch *et al* 1978) as the plot, as the confidence level is varied, of lesion localization fraction relative to the total number of lesions versus the average number of non-lesion localizations per image. Analysis of free-response data in general and FROC curves in particular have been long-standing issues in medical imaging. The term 'free-response' was coined in 1961 (Egan *et al* 1961) who stated '. . . the situation described by the method of free-response is particularly difficult to analyze simply because a trial is not defined . . . a wholly satisfying technique has not yet been devised for the analysis of the (observer's) behavior in this situation' and this statement remains essentially true to this day. Several approaches to analysing FROC curves have been proposed (Bunch *et al* 1978, Chakraborty *et al* 1986, Chakraborty 1989, Chakraborty and Winter 1990, Edwards *et al* 2002, Bornefalk and Hermansson 2005) but remain controversial since they assume that the multiple decisions occurring on an image are statistically independent. A more basic issue, in my opinion, is that the lack of an analytical model for the variable numbers of mark-rating pairs, and the associated location data, limit analysis of search data, even when the data are known to be uncorrelated (e.g., in simulations).

One aim of this paper is to describe an analytical model of visual searching that accounts for the facts that the number and locations of distracters are unknown and the number of marks on an image is unpredictable. Another aim is to describe a figure of merit that allows quantification of search performance. The important issue of estimating the parameters of the search model is not addressed in this paper.

Methods

The ROC model

To provide necessary background to the search model I summarize the ROC paradigm (Metz 1986, 1989). ROC data consist of an ordinal rating for each image (e.g., 1, 2, 3, 4, 5 in a 5-rating ROC study) representing the observer's confidence that the image is abnormal. The ratings represent the binning of the observer's internal confidence level. The model used to analyse ROC data consists of two overlapping Gaussian distributions with different widths, corresponding to the normal and abnormal images. The model assumes that for each image there occurs a scalar sample $\mathbf{z}$ from the appropriate distribution. The $\mathbf{z}$-sample represents the observer's internal confidence that the image is abnormal, with higher values representing greater confidence. The model assumes that the observer adopts R ordered cutoff (or threshold) parameters ζ_i ($i = 1, 2, \ldots, R$) and the cutoff vector $\vec{\zeta}$ is defined as $\vec{\zeta} = (\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_R, \zeta_{R+1})$, where $R + 1$ is the number of ratings bins employed in the ROC study, and $\zeta_0 - \infty$ and $\zeta_{R+1} = +\infty$. The binning rule is that if $\zeta_{i-1} < \mathbf{z} < \zeta_i$ then the corresponding image is assigned to the ith bin. An algorithm for estimating the parameters of the ROC model from ratings data has been described (Dorfman and Alf 1969) and is widely used in medical imaging systems assessment.

The perceptual basis of the search model

The proposed analytical model of search is based on a descriptive model of radiological image interpretation (Kundel and Nodine 1983, 2004, Nodine and Kundel 1987). *A key aspect of this model is that the observer does not assign equal attention units to all locations in the image.* According to the radiological search model, image viewing begins with a brief *global (or preattentive) analysis* requiring a few hundred milliseconds, during which information is collected predominantly by peripheral vision and perturbations in the scene are identified. The observer then examines (i.e., *cognitively evaluates*) these regions individually using foveal vision and makes decisions whether or not to report them. These locations are termed *decision sites*. Decision sites corresponding to normal regions are termed *noise sites* and those corresponding to lesions are termed *signal sites*. The number of noise sites on an image is denoted by $\mathbf{n}$ and the corresponding number of signal sites is denoted by $\mathbf{u}$ and on a normal image $u = 0$. Both $\mathbf{n}$ and $\mathbf{u}$ are random non-negative integers that are unpredictable and usually unobservable unless one employs eye-position recordings (see below). I follow common convention to denote random variables in bold type and the corresponding realizations in normal type.

The radiological search model is based on eye-position recordings made on radiologists. By monitoring corneal reflections from an infrared light source one can measure the line-of-gaze of an observer (Duchowski 2002) and determine the locations where decisions were made. Eye-position recordings for a mammogram for two observers, an inexperienced observer (left panel) and a radiologist (right panel) are shown in figure 1. Individual fixations, defined as locations where the observer's gaze duration (dwell time) exceeded 100 ms, are indicated by

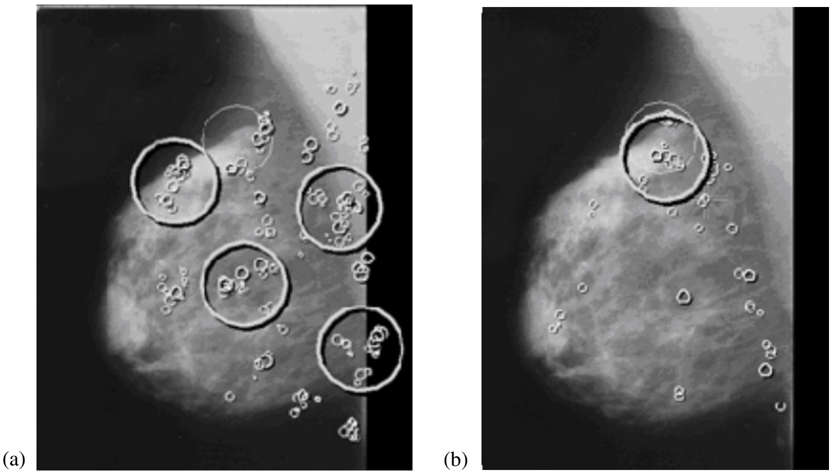

Figure 1. Eye-position recordings for a mammogram displayed on a monitor. Recording (a) is for an inexperienced observer and (b) is for a radiologist. A cancer in the image is indicated by a large low-contrast circle. Brief individual fixations (dwell time >0.1s but <1 s) are indicated by the small circles. The larger high-contrast circles (cumulative dwell time >1 s), which are regions identified by the preattentive first stage, correspond to the decision sites of the search model, i.e. these are the regions that receive cognitive evaluation at the second stage. Note that not all areas of the image receive cognitive evaluation, and the inexperienced observer has more decision sites at normal regions (4 versus 0) and fewer decision sites at lesion locations (0 versus 1) than the radiologist. In the search model this corresponds to a larger value of λ and a smaller value for ν for the inexperienced observer, implying a less efficient preattentive stage.

the small circles. Clustered fixations with a total dwell time exceeding 1 s are indicated by the large high-contrast circles. It is believed (Hillstrom 2000) that the observer makes conscious decisions (cognitive evaluations) to report or not to report only at the locations of the clustered fixations; in other words these are the decision sites of the search model. I use the term ‘locations were hit’ as shorthand for ‘locations where decisions were made’. In figure 1 the large low-contrast circle indicates a cancer. Note that the inexperienced observer has more noise sites and fewer signal sites than does the radiologist. In the example shown the numbers of noise and signal sites are $n = 4$ and $u = 0$ for the non-expert, and $n = 0$ and $u = 1$ for the radiologist.

The search model

The essence of the search model is that it regards the unpredictable numbers of lesion and non-lesion localizations as random variables. Instead of attempting the impossible task of estimating random variables one estimates the parameters of postulated distributions from which the random variables are sampled. The situation is completely analogous to ROC analysis where one does not attempt to estimate the **z**-samples. Instead one estimates the parameters of the assumed Gaussian distributions from which these are sampled. The search model is illustrated schematically in figure 2. Let $N(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. The left and right Gaussian distributions represent the probability density functions corresponding to $N(0, 1)$ and $N(\mu, 1)$ respectively, where μ is a parameter of the search model representing lesion signal-to-noise ratio when the observer knows where to look for the lesion. It characterizes the ability of the observer to extract information from a signal site during cognitive evaluation. It is influenced by external factors (e.g., complexity

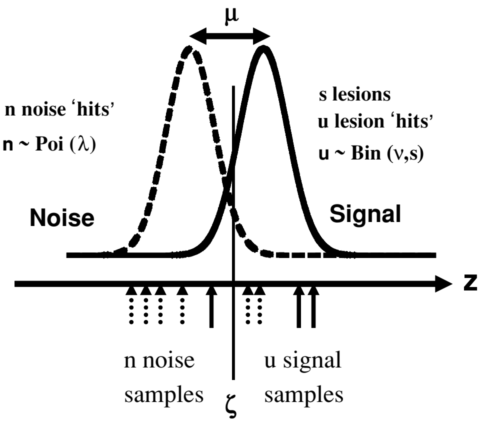

Figure 2. The search model for a single rating study. The unit normal distributions labelled 'Noise' and 'Signal' determine the confidence level samples (**z**) from noise or signal sites, respectively. Their separation μ is the lesion signal-to-noise ratio. When a z-sample exceeds ζ, the observer's threshold, the observer marks the corresponding site. The numbers of noise sites (signal sites) considered for marking are **n** (**u**) respectively. One has $\mathbf{n} \geqslant 0$ and $0 \leqslant \mathbf{u} \leqslant s$, where s is the number of lesions in the image and $\mathbf{u} = 0$ on normal images. The random variables **n** and **u** are modelled by Poisson and binomial distributions, respectively. The parameter λ denotes the mean number of noise sites per image that were considered for marking, in the preattentive stage, and ν is the corresponding probability that a signal site was considered for marking. In the example $n = 6$ (dotted up arrows), $u = 3$ (solid up arrows). Two noise site z-samples exceed the cutoff, leading to two non-lesion localizations (i.e., $f = 2$) and two signal site z-samples exceed the cutoff, leading to two lesion localizations (i.e., $t = 2$), for a total of four marks on this image. Assuming $s = 5$ (it must be at least 3) the values of λ and ν based on this one image sample are 6 and 0.6, respectively.

of the surround, lesion contrast, etc) and observer dependent factors (e.g., eyesight, expertise, etc). As an aside, it was noted earlier that the model used for analysing ROC data uses Gaussian distributions with different widths. The reason for using Gaussian distributions with the same width in the present case is parsimony, and is discussed in greater detail in a companion paper (Chakraborty 2006).

The horizontal axis in figure 2 represents z, the observer's internal confidence that a decision site represents a target. The continuous random variable **z** is modeled by $\mathbf{z} \sim N(0, 1)$ for normal images and $\mathbf{z} \sim N(\mu, 1)$ for abnormal images, where the symbol '$\sim$' is to be read as 'is sampled from'. All **z**-samples on an image are assumed to be independent. The integer **n** ($\mathbf{n} = 0, 1, \ldots$) is the number of noise sites on an image and it is modelled as a Poisson random variable (Larsen and Marx 2001) with parameter λ where $\lambda > 0$, i.e., $\mathbf{n} \sim \text{Poi}\ (\lambda)$. The parameter λ corresponds to the mean number of noise sites per image and smaller values correspond to greater preattentive search expertise at rejecting normal regions of the image from the need for cognitive evaluation. For example, the inexperienced observer in figure 1 would be characterized by a larger value of λ than the radiologist. The number of signal sites on an image can take on values $\mathbf{u} = 0, 1, \ldots, s$, where s is the total number of lesions in an abnormal image. The sampling of **u** is modelled by the binomial distribution (Larsen and Marx 2001) with trial size s and success probability ν ($0 \leqslant \nu \leqslant 1$), i.e., $\mathbf{u} \sim \text{Bin}\ (\text{s}, \nu)$. (For simplicity we assume that each abnormal image has exactly s lesions. The extension to a variable number of lesions per image (e.g., one abnormal image has $s = 1$, another has $s = 2$, etc) is indicated in appendix A.) The parameter ν is the probability that a lesion is hit during the preattentive phase, i.e., it is identified as requiring cognitive evaluation, with larger values of ν corresponding to greater preattentive phase expertise at finding lesions. For example, the inexperienced observer in figure 1 would be characterized by a smaller value of ν than the radiologist.

As with the ROC model one defines a cutoff vector $\vec{\zeta} = (\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_R, \zeta_{R+1})$ where R is the number of ratings bins employed in the free-response study, i.e., the observer is allowed to assign an integer 1 through R to each mark, with higher numbers representing greater confidence. If $\zeta_i < z < \zeta_{i+1}$ ($i = 1, 2, \ldots, R$) then the corresponding decision site is marked and rated in bin 'i', and if $z < \zeta_1$ then the decision site is not marked. It may be observed that for a given number of cutoffs the number of search data bins is 1 less than the corresponding number of data bins in a conventional ROC study. I assume that the location of the mark is at the precise centre of the decision site in question. Therefore any mark made as a consequence of a sample $z \sim N(0, 1)$ that satisfies $\zeta_i < z < \zeta_{i+1}$ will be scored as a non-lesion mark and assigned the rating 'i', and likewise any mark made as a consequence of a sample $z \sim N(\mu, 1)$ that satisfies $\zeta_i < z < \zeta_{i+1}$ will be scored as a lesion mark and assigned the rating 'i'.

In the single-rating example shown in figure 2 the number of noise sites is $n = 6$ (dotted up arrows), the number of signal sites is $u = 3$ (solid up arrows). Two noise sites exceed the cutoff leading to two non-lesion localizations, and two signal sites exceed the cutoff leading to two lesion localizations, for a total of four marks on this image. Assuming $s = 5$ (it must be at least as large as the number of signal sites) the local values of λ and ν based on this one image sample are 6 and 0.6, respectively. The number of non-lesion localizations is denoted by **f** and the corresponding number of lesion localizations is denoted by **t**. Therefore in the example shown in figure 2 one has $f = 2$ and $t = 2$. In the case of multiple ratings the quantities **f** and **t** are replaced by the vectors $\vec{f} = \{f_1, f_2, \ldots, f_R\}$ and $\vec{t} = \{t_1, t_2, \ldots, t_R\}$ respectively.

The figure of merit

The figure of merit $\theta(\mu, \lambda, \nu, s)$ is defined by assuming that the observer uses the rating of the highest rated mark ('highest rating') as the overall confidence level for the image (Swensson 1996). The calculation of $\theta(\mu, \lambda, \nu, s)$ conceptually involves the observer comparing the images in a normal–abnormal pair and attempting to select the abnormal image. The figure of merit is the fraction of correct choices in this task. Note that in these paired comparisons the location of the lesion(s) must be *unknown* to the observer, which is different from the manner in which two alternative forced choice (2AFC) studies are normally conducted (Burgess 1995). The details of the calculation of $\theta(\mu, \lambda, \nu, s)$ are deferred to appendix A.

So far I have not considered the possibility that it may not be possible to vary ν and μ independently. In fact one expects ν to approach 0 as μ approaches 0, since invisible lesions will have zero probability of being hit (strictly speaking this is true only when infinite localization precision is required, as is assumed in this work, before the observer gets credit for lesion localization). Likewise one expects ν to approach 1 as μ approaches ∞ since very high contrast lesions are certain to be hit. To reflect this dependence it is necessary to define the ν parameter in terms of another parameter, β ($\geqslant 0$), and where $\nu = 1 - \exp(-\beta\mu)$, which ensures that $0 \leqslant \nu \leqslant 1$ and that ν approaches the appropriate limits as a function of μ. The quantity β is the rate of increase of ν with μ for small μ. Without this re-parameterization, if one assumes ν to be constant and non-zero, one would have the unphysical result that $\theta(0, \lambda, \nu, s) > 0.5$ (the reason for this is basically due to the larger number of samples from abnormal images, **n** + **u**, than from normal images, **n**). The result is unphysical because with zero contrast lesions the observer's ability to distinguish between normal and abnormal images should be at the chance level. With this re-parameterization it can be shown that the figure of merit always satisfies $0.5 \leqslant \theta(\mu, \lambda, \nu, s) \leqslant 1.0$. However, since ν has a simpler physical interpretation than β, namely ν is the fraction of lesion sites that were hit, I continue to use the ν parameter to describe the model. For a given observer β may be regarded as a constant, i.e., independent of μ, so strictly speaking the basic parameters of the model are μ, λ and β.

Summary of assumptions

(1) The number of noise sites on an image follows the Poisson distribution: $\mathbf{n} \sim \text{Poi}(\lambda)$. The number of signal sites on an abnormal image follows the binomial distribution: $\mathbf{u} \sim \text{Bin(s, } \nu)$. The number of noise sites and the number of signal sites are statistically independent, so that the joint probability of **n** noise sites and **u** signal sites on an abnormal image is given by the product of the two individual probabilities.

(2) A decision variable sample **z** results at each decision site. The binned **z**-sample determines the rating assigned to the decision site. The **z**-sample from a noise site is sampled from a Gaussian distribution with zero mean and unit variance, i.e., $\mathbf{z} \sim N(0, 1)$. The **z**-sample from a signal site is sampled from a Gaussian distribution with mean μ and unit variance, i.e., $\mathbf{z} \sim N(\mu, 1)$. All **z**-samples on an image are statistically independent.

(3) A mark results when the **z**-sample exceeds the lowest cutoff. The observer marks the exact centre of the corresponding decision site.

(4) The following assumptions are needed for the figure of merit calculation. When asked to give a single summary rating to an image the observer gives the rating of the highest rated decision site. On an abnormal image this could be the rating of a noise or a lesion site. On a normal image this is necessarily the rating of a noise site. When asked to select the lesion containing image in a pair of images, one of which is normal and the other is abnormal, the observer picks the image with the highest rating—provided both images of the pair have at least one decision site. If only one of the images has at least one decision site, the observer picks that image. If none of the images has a decision site, the observer picks an image at random.

Results

Table 1 shows the dependence of the figure of merit $\theta(\mu, \lambda, \nu, s)$ on search model parameters μ, λ and ν and the (constant) number s of lesions per image. Also shown are the values of the β parameter, where $\nu = 1 - \exp(-\beta\mu)$. The figure of merit increases with μ and ν, decreases with λ and increases with s. These dependences are consistent with the physical interpretations given to the model parameters. (a) Since μ is the lesion signal-to-noise-ratio, increasing it is expected to improve performance: as the signal distribution in figure 2 shifts to the right, the chance that the highest rating on an abnormal image will exceed that on a normal image increases, i.e., $\theta(\mu, \lambda, \nu, s)$ increases. (b) Since λ is the mean number of noise sites identified by the observer, larger values lead to more noise sites and the probability that the z-sample from one of them will exceed the highest rating from the signal sites will increase, i.e., $\theta(\mu, \lambda, \nu, s)$ decreases. (c) Since ν is the probability that a lesion will be hit, as it increases the increased number of lesion hits leads to a greater probability that the z-sample from one of them will exceed the highest z-sample from the noise site, i.e., $\theta(\mu, \lambda, \nu, s)$ increases. A similar logic applies to the increase of $\theta(\mu, \lambda, \nu, s)$ with s.

Discussion

A key difference between the free response and the ROC paradigms is that in the former, one collects location data and scores the marks as non-lesion or lesion localizations according to

Table 1. This table shows the dependence of the figure of merit $\theta(\mu, \lambda, \nu, s)$ on search model parameters μ, λ and ν and the (constant) number s of lesions per image. The figure of merit increases with μ and ν, decreases with λ and increases with s. The meaning of the β parameter is explained in the text; note that $\nu = 1 - \exp(-\beta\mu)$. The figure of merit does not depend on the cutoff parameter ζ shown in figure 2.

μ	λ	ν	β	s	θ
2	1	0.9	1.151	1	0.8951
2	1	0.7	0.6020	1	0.8073
2	1	0.5	0.3466	1	0.7195
2	1	0.3	0.1783	1	0.6317
1	1	0.8	1.609	1	0.7719
2	1	0.8	0.8047	1	0.8512
3	1	0.8	0.5365	1	0.8882
4	1	0.8	0.4024	1	0.8983
3	0.5	0.7	0.4013	1	0.8445
3	1.0	0.7	0.4013	1	0.8397
3	2.0	0.7	0.4013	1	0.8316
3	4.0	0.7	0.4013	1	0.8192
3	1	0.5	0.2310	1	0.7426
3	1	0.5	0.2310	2	0.8670
3	1	0.5	0.2310	3	0.9309
3	1	0.5	0.2310	4	0.9639

their proximity to actual lesions. The location information is not collected in ROC studies. Consequently the ROC paradigm does not reward the radiologist for the ability to locate more lesions on an image while mistaking fewer non-lesion locations for lesions. It has been shown that the inclusion and analysis of the location information in the jackknife free-response receiver operating characteristic (JAFROC) method leads to improved precision in the measurement and greater statistical power in differentiating between modalities (Chakraborty and Berbaum 2004, Zheng *et al* 2005, Penedo *et al* 2005). While the JAFROC method does not assume independence of the search data, it suffers from the limitation of not using all of the available data (e.g., on a normal image it uses only the rating of the highest rated non-lesion localization and on an abnormal image it uses only the ratings of localized lesions). In order to use all the rating data one needs a search model. This was one of the motivations for this work.

The search model has two precursors in the medical imaging literature. Swensson described a model for medical imaging (Swensson 1980) that also invokes a two-stage process that has some similarities to the present work. The present work is intimately related to the 'initial detection and candidate analysis' (IDCA) approach (Edwards *et al* 2002). The term 'initial detection' refers to the first stage where the observer identifies a finite number of regions that are possible lesion candidates. The term 'candidate analysis' refers to the second stage where the observer obtains decision variable samples at the regions identified by the first stage, and marks them if they exceed the lowest cutoff. A comparison of the two models (Swensson's and IDCA) to the present work is provided in appendix B. The concept inherent in the ν parameter of the search model, that some lesions are not hit, is related to the α parameter in the contaminated binormal model (CBM) in Dorfman and Berbaum (2000). The CBM α parameter ($0 \leqslant \alpha \leqslant 1$) is the proportion of abnormal cases where the abnormalities are visible. Since CBM describes ROC data, comparisons become possible only when one considers ROC curves predicted by the two models. These are discussed in greater depth in the companion paper.

The current search model is fundamentally different from a class of models in the psychophysical literature that assume, either implicitly or explicitly, that observers search through all items (distracters + targets) in the display one-by-one, until they either find the targets or exhaust the number of items (Horowitz and Wolfe 2001, Harris *et al* 1979, Hoffman 1978). These approaches assume that the total number of distracters is known to the experimenter. In the medical imaging task the potential number of normal regions that resemble lesions (i.e., the distracters) is unknown. Indeed what constitutes a distracter depends on the expertise of the observer. In figure 1 the radiologist (panel b) did not consider any of the regions that received cognitive evaluation by the non-expert (panel a) as worthy of cognitive evaluation. In spite of this apparent lack of attention to the whole image the radiologist successfully located the lesion, whereas the non-expert did not. It may appear counter-intuitive that such lack of attention can be consistent with a good observer. Assume for the moment that neither observer marked any of the cognitively evaluation regions, i.e., the corresponding decision variable samples did not exceed the cutoff. Since both observers provide identical data (i.e., no marks) on this image, it is reasonable to ask why not reward the non-expert for paying more attention to the image? One could argue that not marking the four normal regions might outweigh the fact that the non-expert missed the lesion, and in this sense the non-expert may be better. The paradox can be resolved by the following arguments. (a) The radiologist did pay preattentive attention to the normal regions and eliminated them while not eliminating the lesion. In contrast, the non-expert failed to eliminate the normal regions during the preattentive phase and needed cognitive evaluation at the second stage to finally reject them. Moreover this observer rejected the lesion during the preattentive phase. In other words, the preattentive stage of the radiologist is more efficient. (b) This image yields $f = 0$ and $t = 0$ for both observers. However, for a subset ensemble of similar images (i.e., with the same values of n and u but random $\mathbf{z}$'s), some of the $\mathbf{z}$-samples for the non-expert will exceed the lowest cutoff and will be marked, but the non-expert will never mark the lesion. In other words this subset ensemble of images will yield $\langle f \rangle > 0$ and $\langle t \rangle = 0$ for the non-expert. By a similar argument the radiologist will yield $\langle f \rangle = 0$ and $\langle t \rangle > 0$. On both counts the search model rewards the radiologist with smaller λ and larger ν, both of which lead to larger θ.

The search model assumes that the confidence level samples occurring at decision sites on the same image are independent. To my knowledge this limitation is shared by almost all methods that have been proposed for analysing search data (Edwards *et al* 2002, Horowitz and Wolfe 2001, Eckstein *et al* 2000, Swensson 1996). An exception is the work by Swensson (1980). The Poisson assumption theoretically allows an infinite number of noise sites per image. This may not be a serious limitation when the lesion size is small compared to the image area (Edwards *et al* 2002) as in microcalcification detection but could be a limitation in other cases. The search model assumes that the observer's mark is at the precise location of the decision site. In practice the observer cannot indicate a location precisely and for clinical lesions it may not be possible to define a lesion centre that all radiologists will agree on. The search model does not address the satisfaction of search issue (Berbaum *et al* 1990).

Acknowledgments

This work was supported by a grant from the Department of Health and Human Services, National Institutes of Health, 1R01-EB005243. The author is grateful to Dr Claudia Mello-Thoms for providing figure 1 and for proofing the manuscript. The author is also grateful to

Dr Darrin Edwards for correspondence regarding the IDCA approach, and to Hong-Jun Yoon, MSEE, for implementation of the formulae.

Appendix A

Figure of merit

The unit variance Gaussian probability density function and the corresponding probability distribution function are defined by

$$\phi(z|\mu) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-(z-\mu)^2/2} \qquad \Phi(z|\mu) = \int_{-\infty}^{z} \mathrm{d}y\,\phi(y|\mu). \tag{A.1}$$

The Poisson and binomial density functions are defined by

$$\mathrm{Poi}(n|\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda} \qquad \mathrm{Bin}(u|s,\nu) = \binom{s}{u}\,\nu^u(1-\nu)^{s-u}. \tag{A.2}$$

The calculation of the figure of merit conceptually involves the observer comparing the images in a normal–abnormal pair and attempting to select the abnormal image. The figure of merit is defined as the fraction of correct choices in this task. Four cases need to be distinguished: (a) both images have at least one hit, (b) neither image has a hit, (c) only the abnormal image has a hit and (d) only the normal image has a hit. For case (b) assume that the observer picks between the images at random so that the probability of a correct choice is 0.5. For cases (c) and (d) assume that the observer picks whichever image was hit, so that the probability of a correct choice is one or zero, respectively. The final figure of merit is obtained by performing a weighted average using these probabilities.

The figure of merit for case (a), which is the most involved, is described next. Define $\mathrm{PDF}_s(z|\mu,\lambda,\nu,s)$ as the probability distribution function (PDF) of the highest rating on abnormal images each of which has at least one hit, i.e., this is the probability that on such images the highest rating does not exceed z. Define $\mathrm{pdf}(z|\lambda)$ as the probability density function (pdf) of the highest rating on normal images. These functions are related by

$$\mathrm{pdf}(z|\lambda) = \frac{\partial}{\partial z}\mathrm{PDF}_s(z|0,\lambda,0,0). \tag{A.3}$$

When both images of the pair have at least one hit, i.e., for case (a), the figure of merit $\theta_h(\mu,\lambda,\nu,s)$ is obtained by integrating $[1-\mathrm{PDF}_s(z|\mu,\lambda,\nu,s)]\,\mathrm{pdf}(z|\lambda)$ over all values of z (Swensson 1996), namely

$$\theta_h(\mu,\lambda,\nu,s) = \int_{-\infty}^{\infty} \mathrm{d}z\,\mathrm{pdf}(z|\lambda)\;[1-\mathrm{PDF}_s(z|\mu,\lambda,\nu,s)], \tag{A.4}$$

where the subscript h denotes that each image in the pair has at least one hit. The function $\mathrm{PDF}_s(z|\mu,\lambda,\nu,s)$ can be calculated as follows. First one calculates $P_{nu}(z|\mu,n,u)$, the probability that the highest rating exceeds z for abnormal images with **n** noise sites and **u** signal sites (I use appropriate subscripts to emphasize the different functions resulting from the cascaded averaging described below). By the independence assumption this is given by

$$P_{nu}(z|\mu,n,u) = 1 - [\Phi(z|0)]^n\;[\Phi(z|\mu)]^u\,. \tag{A.5}$$

Next one calculates the probability, $P_{ns}(z|\mu,\nu,n,s)$, that the highest rating on abnormal images with s lesions exceeds z. This is obtained by averaging $P_{nu}(z|\mu,n,u)$ over all allowed values of **u**. The probability of obtaining **u** samples is Bin $(u \mid s,\,\nu)$. There are two cases corresponding to $\mathbf{n} = 0$ and $\mathbf{n} > 0$:

$$\begin{aligned} P_{ns}(z|\mu, \nu, n, s) &= \sum_{u=0}^{s} \mathrm{Bin}(u|s, \nu) P_{nu}(z|\mu, n, u) \qquad n > 0 \\ P_{0s}(z|\mu, \nu, 0, s) &= \sum_{u=1}^{s} \mathrm{Bin}(u|s, \nu) P_{nu}(z|\mu, 0, u) \qquad n = 0. \end{aligned} \tag{A.6}$$

In the second equation the lower limit on $\mathbf{u}$ is unity since one is considering case (a) where both images have at least one hit (i.e., $\mathbf{n} + \mathbf{u} > 0$). The probability $P_s(z|\mu, \lambda, \nu, s)$ that the highest rating on an abnormal image with s lesions exceeds z is obtained by averaging $P_{ns}(z|\mu, \nu, n, s)$ over all values of $\mathbf{n}$. The probability of obtaining $\mathbf{n}$ samples is $\mathrm{Poi}(n|\lambda)$. Therefore

$$P_s(z|\mu, \lambda, \nu, s) = \mathrm{Poi}(0|\lambda) P_{0s}(z|\mu, \nu, 0, s) + \sum_{n=1}^{\infty} \mathrm{Poi}(n|\lambda) P_{ns}(z|\mu, \nu, n, s). \tag{A.7}$$

The desired expression for $\mathrm{PDF}_s(z|\mu, \lambda, \nu, s)$ is obtained by dividing the complement of the above expression by the average probability that an abnormal image has at least one hit, since this is the case being considered (i.e., case (a)). This normalization is needed to ensure that $\mathrm{PDF}_s(z|\mu, \lambda, \nu, s)$ is a true probability distribution function, i.e., it approaches 0 and 1 in the appropriate limits. (In the limit $z = -\infty$ all ratings exceed z and therefore the probability, $P_s(z|\mu, \lambda, \nu, s)$, that the highest rating exceeds z equals the probability that there is at least one hit, which is smaller than unity. Therefore, if one did not normalize, the PDF at $z = -\infty$ would be greater than 0.) The probability, $P_h(n, \nu, s)$, that an abnormal image with n noise sites has at least one hit is given by (δ is the Kroenecker delta function)

$$P_h(n, \nu, s) = (1 - \delta_{n,0}(1 - \nu)^s). \tag{A.8}$$

This expression can be understood as follows: for $\mathbf{n} > 0$ the delta function is zero and $P_h(n > 0, \lambda, s)$ is unity, consistent with the fact that such images are guaranteed to have at least one hit. For $\mathbf{n} = 0$ the probability that at least one lesion is hit is the complement of the probability $(1 - \nu)^s$ that none of the lesions were hit. Therefore $\mathrm{PDF}_s(z|\mu, \lambda, \nu, s)$ is given by

$$\mathrm{PDF}_s(z|\mu, \lambda, \nu, s) = \frac{[1 - P_s(z|\mu, \lambda, \nu, s)]}{\sum_{n=0}^{\infty} \mathrm{Poi}(n|\lambda) P_h(n, \nu, s)}. \tag{A.9}$$

Probabilities of the various cases

Case (a). In order for both images to have at least one hit, the normal image must have at least one hit and the abnormal image must have at least one hit. The probability that the normal image has at least one hit is $[1 - \mathrm{Poi}(0|\lambda)]$. In order for the abnormal image to have at least one hit either $\mathbf{n} > 0$, with probability $[1 - \mathrm{Poi}(0|\lambda)]$, or $\mathbf{n} = 0$ and $\mathbf{u} > 0$, with probability $\mathrm{Poi}(0|\lambda)\ (1 - \mathrm{Bin}(0, s, \nu))$. Therefore the net probability corresponding to case (a) is

$$P_a = (1 - \mathrm{Poi}(0|\lambda))\, [(1 - \mathrm{Poi}(0|\lambda)) + \mathrm{Poi}(0|\lambda)\ (1 - \mathrm{Bin}(0, s, \nu))]\,. \tag{A.10}$$

Case (b). In order for neither image to have a hit, the normal image must not have a hit, with probability $\mathrm{Poi}(0|\lambda)$, and the abnormal image must not have a hit. An abnormal image will not have a hit if (a) the number of noise sites is zero and (b) the number of signal sites is zero. The corresponding probabilities are $\mathrm{Poi}(0|\lambda)$ and $\mathrm{Bin}(0, s, \nu)$, respectively. Therefore the probability that an abnormal image does not have a hit is $\mathrm{Poi}(0|\lambda)\mathrm{Bin}(0, s, \nu)$. The probability that neither image of the pair has a hit is the product of the individual probabilities, i.e.,

$$P_b = [\mathrm{Poi}(0|\lambda)]^2\, \mathrm{Bin}(0, s, \nu). \tag{A.11}$$

Case (c). Analogous to case (a) the probability that a normal image does not have a hit and the abnormal image does is

$$P_c = \mathrm{Poi}(0|\lambda)[(1 - \mathrm{Poi}(0|\lambda)) + \mathrm{Poi}(0|\lambda)(1 - \mathrm{Bin}(0, s, \nu))]. \tag{A.12}$$

Case (d). The probability that an abnormal image does not have a hit is $\mathrm{Poi}(0|\lambda)\mathrm{Bin}(0, s, \nu)$. The probability that the normal image does have a hit is $[1 - \mathrm{Poi}(0|\lambda)]$. These results lead to the following expression for the case (d) probability:

$$P_d = [1 - \mathrm{Poi}(0|\lambda)]\,\mathrm{Poi}(0|\lambda)\mathrm{Bin}(0, s, \nu). \tag{A.13}$$

The final figure of merit is given by averaging the figure of merit values weighted by the corresponding probabilities. i.e.,

$$\theta(\mu, \lambda, \nu, s) = P_a\theta_h(\mu, \lambda, \nu, s) + 0.5P_b + P_c + 0 \times P_d. \tag{A.14}$$

For simplicity so far I have assumed that every abnormal image has a constant number (s) of lesions per image. Variable numbers of lesions can be accommodated by averaging $\theta(\mu, \lambda, \nu, s)$ over the distribution of s:

$$\theta(\mu, \lambda, \nu) = \sum_{s=1}^{\infty} h(s)\,\theta(\mu, \lambda, \nu, s), \tag{A.15}$$

where $h(s)$ is the fraction of abnormal cases with s lesions ($s = 1, 2, 3, \ldots; \sum h(s) = 1$). A Maple worksheet implementation of these results is available from the author. This was used to generate table 1.

Appendix B

Relation of the search model to Swensson's model

Swensson has described a search model for medical imaging (Swensson 1980) that has a two-stage process similar to the present search model. In his model the locations of a pool of *potential* decision sites are assumed to be known to the experimenter. This pool includes the known lesion locations. The rest are *potential* noise sites whose number I denote by N. In the described applications to observer data the pool of potential noise sites was selected by the experimenter based on regions in the images that resembled lesions. Each site in the pool is assumed to yield a pair of random decision variables ($\mathbf{X_n}$, $\mathbf{Y_n}$) or ($\mathbf{X_s}$, $\mathbf{Y_s}$) corresponding to whether they originated from non-lesion or lesion sites, respectively. The variables $\mathbf{X}$ and $\mathbf{Y}$ describe the first (preattentive) and second (cognitive evaluation) stages of the model, respectively. A cutoff parameter C determines if a particular site from the pool is selected as a candidate for cognitive evaluation, i.e., if $\mathbf{X} > C$ the site is selected. The number of noise sites/signal sites selected from the pool corresponds to $\sum\mathbf{n}/\sum\mathbf{u}$ (i.e., summed over all images) in the present model. A second cutoff parameter ζ describes the result of the cognitive evaluation, i.e., if $\mathbf{X} > C$ and $\mathbf{Y} > \zeta$ the site is marked and rated. The number of non-lesion/lesion marks corresponds to $\sum\mathbf{f}/\sum\mathbf{u}$ in the present model. Specifically, a non-lesion mark occurs if $\mathbf{X_n} > C$ and $\mathbf{Y_n} > \zeta$. Likewise, a lesion mark occurs if $\mathbf{X_s} > C$ and $\mathbf{Y_s} > \zeta$. The sampling of ($\mathbf{X_n}$, $\mathbf{Y_n}$) is assumed to be bivariate normal with means (0, 0), standard deviations (1, 1) and correlation r_n. Likewise, the sampling of (X_s, Y_s) is assumed to be bivariate normal with means (Δ_x, Δ_y), standard deviations (σ_x, σ_y) and correlation r_s. Therefore, not counting N, the model is described by seven parameters. In Swensson's model the total number of noise sites ($\sum\mathbf{n}$) is determined by N and C. The total number of non-lesion localizations is determined by $\sum\mathbf{n}$ and ζ. The number of signal sites ($\sum\mathbf{u}$) is determined by the number of lesions, Δ_x, σ_x

and C. The number of lesion localizations is determined by $\sum \mathbf{u}$, Δ_y, σ_y and ζ. At the cost of more parameters Swensson's model allows for possible correlations between **n** and **u** that are neglected in the present work. Swensson's model assumes that each site in the pool is evaluated by the observer at the first stage. This is an important distinction from the present search model which does not specify experimenter-selected sites that the observer is assumed to evaluate.

Relation of the search model to IDCA

The search model and IDCA (Edwards *et al* 2002) are closely related. The initial detection/candidate analyses stages correspond to the first stage/second stage of the search model. Both involve Poisson/binomial sampling for the noise sites/signal sites, respectively. There are minor differences. In the IDCA formalism the signal site decision variable is assumed to be sampled from $N(\mu,\sigma^2)$, i.e., the variance of the signal site decision variable is an additional parameter, whereas in the present case the sampling is from $N(\mu, 1)$. The Poisson parameter in the present case is defined for individual images, i.e., λ is the average number of noise sites per image. In IDCA it is defined over the whole image set, i.e., the IDCA Poisson parameter corresponds to λN_T in the present notation, where N_T is the total number of images. I use uppercase letters to denote random variables defined over the entire image set. For example, **n**/**u** are the number of noise sites/signal sites per image and **N**/**U** are the number of noise sites/signal sites for the whole image set (corresponding to **B** and **C** in the IDCA paper). This distinction is inconsequential as both models assume independence and therefore the variables **n** and **u** can be summed over all images without changing the statistics. More importantly, IDCA assumes that **N** and **U** are known to the experimenter. The primary intended application of IDCA is evaluation of computer aided detection (CAD) systems. In this case **N** and **U** are indeed known to the designer of the CAD algorithm. In the example quoted in the IDCA paper the total number of noise regions identified by CAD at the initial detection stage was 7165 (i.e., $N = 7165$) and the total number of lesions identified was 132 (i.e., $U = 132$). Given the total number of images (43) and the total number of lesions (171), maximum likelihood estimates of the λ and ν parameters are $\lambda = 7165/43$ and $\nu = 132/171$, respectively. In the present search model the corresponding quantities **n** and **u** are regarded as unknown. Therefore in principle λ and ν need to be estimated from the free-response data, i.e., from $\vec{f} = \{f_1, f_2, \ldots, f_R\}$ and $\vec{t} = \{t_1, t_2, \ldots, t_R\}$, a problem not addressed in this paper. This difference translates to significant differences in the calculations of statistical quantities. In the present formulation each statistic is a Poisson/binomial weighted summation over all values of **n**/**u**, subject to the restrictions that **n**/**u** cannot be smaller than the observed number of non-lesion/lesion localizations in the image. This is illustrated in appendix A: see equations (6) and (7). In the IDCA approach the summations are not performed (see equations (32) and (33) in the IDCA paper where one keeps only one term, that corresponding to the observed values of **N** and **U**). Whether this difference translates to differences in summary statistics, e.g., the figure of merit or the FROC curve, is outside the scope of this work.

References

Berbaum K S *et al* 1990 *Invest. Radiol.* **25** 133–40
Bornefalk H and Hermansson A B 2005 *Med. Phys.* **32** 412–7
Bunch P C, Hamilton J F, Sanderson G K and Simmons A H 1978 *J. Appl. Photogr. Eng.* **4** 166–71
Burgess A E 1995 *Med. Phys.* **22** 643–55
Chakraborty D P 1989 *Med. Phys.* **16** 561–8
Chakraborty D P 2006 *Phys. Med. Biol.* **51** 3463–82

Chakraborty D P and Berbaum K S 2004 *Med. Phys.* **31** 2313–30

Chakraborty D P, Breatnach E S, Yester M V, Soto B, Barnes G T and Fraser R G 1986 *Radiology* **158** 35–9

Chakraborty D P and Winter L H L 1990 *Radiology* **174** 873–81

Dorfman D D and Alf E 1969 *J. Math. Psychol.* **6** 487–96

Dorfman D D and Berbaum K S 2000 *Acad. Radiol.* **7** 427–37

Duchowski A T 2002 *Eye Tracking Methodology: Theory and Practice* (Clemson, SC: Clemson University)

Eckstein M P, Thomas J P, Palmer J and Shimozaki S S 2000 *Percept. Psychophys.* **62** 425–51

Edwards D C, Kupinski M A, Metz C E and Nishikawa R M 2002 *Med. Phys.* **29** 2861–70

Egan J P, Greenburg G Z and Schulman A I 1961 *J. Acoust. Soc. Am.* **33** 993–1007

Harris J R, Shaw M L and Bates M 1979 *Percept. Psychophys.* **26** 69–84

Hillstrom A 2000 *Percept. Psychophys.* **2** 800–17

Hoffman J E 1978 *Percept. Psychophys.* **23** 1–11

Horowitz T S and Wolfe J M 2001 *Percept. Psychophys.* **63** 272–85

Kundel H L and Nodine C F 1983 *Radiology* **146** 363–8

Kundel H L and Nodine C F 2004 *Proc. SPIE* **5372** 110–5

Larsen R J and Marx M L 2001 *An Introduction to Mathematical Statistics and Its Applications* (Upper Saddle River, NJ: Prentice-Hall)

Metz C E 1986 *Invest. Radiol.* **21** 720–33

Metz C E 1989 *Invest. Radiol.* **24** 234–45

Nodine C F and Kundel H L 1987 *Radiographics* **7** 1241–50

Palmer J, Verghese P and Pavel M 2000 *Vis. Res.* **40** 1227–68

Penedo M *et al* 2005 *Radiology* **237** 450–7

Rotello C M, Macmillan N A and Reeder J A 2004 *Psychol. Rev.* **111** 588–616

Swensson R G 1980 *Percept. Psychophys.* **27** 11–6

Swensson R G 1996 *Med. Phys.* **23** 1709–25

Treisman A and Gelade G 1980 *Cogn. Psychol.* **12** 97–136

Treisman A and Gormican S 1988 *Psychol. Rev.* **95** 15–48

Wagner R F, Beiden S V, Campbell G, Metz C E and Sacks W M 2002 *Acad. Radiol.* **9** 1264–77

Wolfe J M 1998 *Attention* ed H Pashler (London, UK: University College London Press)

Wolfe J M 2005 *Science* **308** 503–4

Zheng B, Chakraborty D P, Rockette H E, Maitz G S and Gur D 2005 *Med. Phys.* **32** 1031–4