Published in final edited form as:

Acad Radiol. 2012 December ; 19(12): 1474–1483. doi:10.1016/j.acra.2012.09.002.

Application of threshold-bias independent analysis to eye-tracking and FROC data

Dev P. Chakraborty, Ph.D.⁽¹⁾, Hong-Jun Yoon, Ph.D.⁽¹⁾, and Claudia Mello-Thoms, M.S., Ph.D.⁽²⁾

Dev P. Chakraborty: dpc10@pitt.edu; Claudia Mello-Thoms: mellothomsc@upmc.edu

⁽¹⁾Department of Radiology, University of Pittsburgh, Presbyterian South Tower, Room 4771, 200 Lothrop Street, Pittsburgh, PA 15213, 412-605-1553 (p), 412-605-1554 (f), 412-605-1553 (phone), 412-605-1554 (fax)

⁽²⁾University of Pittsburgh School of Medicine, Department of Biomedical Informatics and Department of Radiology, The Offices at Baum, 5th floor, Room 516, 5607 Baum Blvd, Pittsburgh, PA 15206-3701, Phone: (412) 648–9314

Abstract

Rationale and Objectives—Studies of medical image interpretation have focused on either assessing radiologists' performance using, for example, the receiver operating characteristic (ROC) paradigm, or assessing the interpretive process by analyzing eye-tracking (ET) data. Analysis of ET data has not benefited from threshold-bias independent figures-of-merit (FOMs) analogous to the area under the ROC curve. The aim was to demonstrate the feasibility of such FOMs and to measure the agreement between figures-of-merit derived from free-response ROC (FROC) and ET data.

Methods—Eight expert breast radiologists interpreted a case set of 120 two-view mammograms while eye-position data and FROC data were continuously collected during the interpretation interval. Regions that attract prolonged (>800ms) visual attention were considered to be virtual marks, and ratings based on the dwell and approach-rate (inverse of time-to-hit) were assigned to them. The virtual ratings were used to define threshold-bias independent FOMs in a manner analogous to the area under the trapezoidal alternative FROC (AFROC) curve (0 = worst, 1 = best). Agreement at the case level (0.5 = chance, 1 = perfect) was measured using the jackknife and 95% confidence intervals (CI) for the FOMs and agreement were estimated using the bootstrap.

Results—The AFROC mark-ratings FOM was largest 0.734, CI = (0.65, 0.81) followed by the dwell 0.460 (0.34, 0.59) and then by the approach-rate FOM 0.336 (0.25, 0.46). The differences between the FROC mark-ratings FOM and the perceptual FOMs were significant ($p < 0.05$). All pairwise agreements were significantly better than chance: ratings vs. dwell 0.707 (0.63, 0.88), dwell vs. approach-rate 0.703 (0.60, 0.79) and rating vs. approach-rate 0.606 (0.53, 0.68). The ratings vs. approach-rate agreement was significantly smaller than the dwell vs. approach-rate agreement ($p = 0.008$).

Conclusions—Leveraging current methods developed for analyzing observer performance data could complement current ways of analyzing ET data and lead to new insights.

Introduction

Observer performance measurements in radiology involve collecting ratings, usually one per image as in the receiver operating characteristic (ROC) paradigm (1, 2), or one per perceived suspicious region as in the free-response receiver operating characteristic (FROC) paradigm (3). The rating, which could be an integer or continuous variable, represents a decision about how confident the observer is in the presence of abnormality in the image (for ROC) or at the location of the perceived suspicious region (for FROC). The analysis of observer performance data, to which Professor Charles E. Metz has made seminal contributions, uses a figure-of-merit (FOM), such as the area under the ROC curve (AUC), which rewards/penalizes correct/incorrect decisions in a manner commensurate with their ratings (4). For example, in the ROC paradigm high rated abnormal images and low rated normal images are rewarded more than intermediate rated images; and low rated abnormal images and high rated normal images are penalized more than intermediate rated images. Likewise, a FROC FOM (5) rewards high rated lesions and unmarked normal images while penalizing high rated marks on normal images. In either paradigm the FOM depends on reader skill and on the difficulty of the cases but is independent of threshold-bias inherent in an observer's usage of the rating scale.

Threshold-bias can be thought of as how conservative or liberal the observer is in using the rating scale: a conservative observer tends to be more reluctant to give high ratings while the liberal observer is less reluctant. Two observers may use the ratings scale quite differently yet in ROC analysis they could yield identical AUCs. ROC analysis eliminates threshold-bias by measuring the *differential* ability of the ratings to correctly classify diseased and non-diseased images. Threshold-bias independence is an important advantage of ROC/FROC area-based FOMs over sensitivity-specificity analysis (6). As an example consider the study (7) conducted to evaluate the effectiveness of screening mammography by estimating the variability in radiologists' ability to detect breast cancer. Fifty accredited mammography centers were randomly sampled from across the United States. One hundred eight (108) radiologists from these centers gave blinded interpretation to the same set of 79 randomly selected screening mammograms. Ground truth for these women had been established either by biopsy or by 2-year follow-up. The observed range of sensitivity (TPF) was at least 40% and the range of FPF was at least 45%. The study shows that a large part of the variability in sensitivity and specificity is due to the radiologists' variable thresholds for reporting disease. Once this is accounted for the variability of the AUC measure, representing the intrinsic variability in diagnostic abilities of the mammographers is only 11%.

While the ROC paradigm uses information about *decisions* made on images, and the FROC paradigm uses information about decisions made on perceived suspicious regions, neither paradigm collects information about the *process* used by the observer to arrive at these decisions. Such information may be collected using eye-tracking (ET) instrumentation that allows near real-time monitoring of the line of gaze of the observer. Algorithms are available that reduce the raw data to a few locations per image that attracted visual attention and yield parameters which describe the magnitude of the attention, such as time-to-hit the location of the lesion and dwell-time fixating abnormalities and other locations. ET methods have proven very useful in understanding medical image perception and central to discoveries about why interpretative failures occur. As one example, use of ET methodology led to the discovery that 70% of unreported breast cancers in mammograms (8), and 70% of unreported lung nodules in chest radiographs are in fact visually inspected by radiologists, often for as long as correctly reported lesions (9). These findings effectively refute a long held belief that a cancer was missed because the radiologist never looked at it.

However, analyses of eye-tracking data — principally those pertaining to visual attention parameters such as time-to-hit and dwell-time — do not exploit the potential of a figure-of-merit that eliminates threshold bias. Instead the focus has been on the actual values of dwell and time-to-hit, which can be thought of as the counterparts to the ratings that are collected in ROC/FROC studies (more on this below). In this sense current ET data analysis is roughly comparable to sensitivity-specificity analysis in ROC research. Our purpose was to leverage some of the methodology from ROC/FROC analysis to rectify this situation by introducing perceptual FOMs that are threshold-bias independent.

Previous studies involving radiologists reading mammograms have used a two-stage design dictated by the experimenter (10): in the first *search* stage the radiologists are instructed to visually search the image until they decide whether or not there are reportable lesion(s) present during which period, which typically lasts 5 – 45 seconds (good radiologists tend to take less time to decide), their eye-positions are recorded. Once they complete the search stage, eye-position recording is terminated, and a second *reporting* stage starts, where radiologists use a mouse-controlled cursor to mark the locations of the detected lesions. One issue with this design is that during the reporting phase a radiologist may discover something new and proceed to investigate this finding, but because eye-position recording has been terminated, that information is not captured. In the data collection methodology used for this study searching and reporting occur simultaneously with eye-position collection. This paradigm more closely resembles clinical practice, and potentially allows one to follow the perceptual and interpretative process entailed in case reading from beginning to end without interruptions.

The approach followed in this work builds on an analogy between the FROC and the ET paradigms. In the FROC paradigm one records mark-rating pairs, where a mark is the location of a perceived suspicious region and the rating is the confidence that the identified region actually harbors a lesion. In the ET paradigm the apparatus records the locations of regions that attracted visual attention. In the proposed approach to analysis of eye-tracking data the cumulated gaze time at each such location (dwell) is interpreted as an ET-derived rating. The only essential difference between the two paradigms is that FROC collects *physical* mark-rating pairs while ET provides *virtual* mark-rating pairs.

In the following we describe the methodology in which FROC and ET data were simultaneously collected, the analysis of the data, definitions of the FOMs, and the method used to measure the agreement between the FROC and ET derived FOMs.

Materials and Methods

The study received ethical approval by the Institutional Review Board (IRB # PRO09040434). Eye-position and FROC data were simultaneously collected for breast radiologists as they interpreted digital mammograms.

Selection of Observers, cases and ground truth

Eight Mammography Quality Standards Act (MQSA) certified breast radiologists participated in this study. Patients were imaged using a Selenia full-field digital mammography system (Hologic Inc, Bedford, MA) as part of the regular screening practice in our institution. The case set consisted of 120 two-view (CC, cranial-caudal, and MLO, medial-lateral oblique) digital mammogram cases of which 59 contained a solitary biopsy verified cancer and 61 were lesion-free and had been stable for 2 years. For each patient we only included either their left or right breast in this study, not both. For the cancer cases the true lesion location was established by having an expert MQSA-certified breast radiologist, who otherwise did not participate in the study, mark the location of the center of the cancer

on each abnormal case using pathology reports and additional imaging as guides. These expert localizations served as the gold standard for the purposes of classifying marked regions as lesion-localizations (LLs) or non-lesion localizations (NLs).

FROC and Eye-Tracking Data Collection

The data collection is shown schematically in Fig. 1. The radiologists sat, on average, 60 cm from a workstation containing two calibrated medical-grade 5 Megapixel flat-panel portrait-mode displays (model C5i, Planar Systems Inc, Beaverton, OR), with a resolution of 2048×2560 pixels, typical brightness of 146 ftL and 3061 unique shades of gray. A head-mounted eye-position tracking system (ASL Model H6, Applied Sciences Laboratory, Bedford, MA) was worn that used an infrared beam to calculate line-of-gaze by monitoring the pupil and the first (and strongest) corneal reflection. A magnetic head tracker was used to monitor head position, and this allows the radiologists to freely move their head from side to side as well as towards the displays, up to 20 cm, at which point they were outside the range of the head tracker. The eye-tracker integrates eye-position and head position to calculate the intersection of the line of gaze and the display plane. The system has an accuracy (measured as the difference between true eye-position and computed eye-position) of less than 1° of visual angle, and it covers a visual range of 50° horizontally and 40° vertically.

Study Protocol

The radiologists were instructed to mark the locations of malignant lesions only; they were specifically instructed to not report definitely benign findings. The radiologists interpreted the 120 two-view digital mammograms usually in 2 sessions, each lasting approximately 1 hour. Data collection was preceded by calibration of the eye-tracker using a 9-point (3×3) grid displayed on both monitors. The calibration was performed on the left hand side monitor and checked using the 4 corners and the center points in both monitors. After calibration, the radiologists used a mouse-controlled cursor to click on a 'Next' button in the display. Two views of the same breast then appeared (one view per monitor) and the eye-tracker received a signal to begin recording the raw data. Upon detecting a suspicious finding that warranted reporting, the radiologist guided the mouse-controlled cursor to the center of the lesion and clicked on it upon which a small pop-up menu appeared where they selected their confidence level using a 1 to 5 point scale (5 = highest confidence). Locations thus identified were overlaid on the image with a small circle containing the radiologist's confidence level. Overlaid information could be toggled on/off, and the marks and ratings could be modified or deleted until the radiologist was satisfied with the interpretation. Once satisfied that all reportable regions had been appropriately marked and rated the radiologist clicked on a button marked 'Next Case' upon which signal the eye-tracker terminated the recording for that image and the data was committed to hard-disk; no further changes were permitted after this. If no cancers were deemed present on the case, the radiologist could click on the 'Next Case' button without making any marks. After every 5 cases the calibration of the eye-tracker was re-checked by having the radiologists fixate the 4 corner points and the center point of the 9-point grid. If necessary the eye tracker was recalibrated. On average, for each reading session the eye tracker only had to be recalibrated once or twice.

The following information was automatically captured by the computer:

- i. The (x,y) location of marks made by the radiologists. Each mark was compared to the locations of the actual lesion and classified as FROC lesion localization (FROC LL) if it fell within 2.5° (termed the proximity criterion) of visual angle (roughly 200 pixels). Otherwise it was classified as FROC non-lesion localization (FROC NL).

- ii. The confidence level rating for each mark.
- iii. Time-stamped, raw eye-position data collected during the entire time that the radiologists were examining the cases. This data, which was acquired at 60 frames per second, included flags to indicate when image manipulation activities (such as marking, rating or window/level adjustments) and blinks occurred. The flagged data frames were excluded from the analysis.

Eye-position Data Clustering

The raw-data was analyzed using ASL's EyeNAL analysis program which converts the raw-data for an image to fixations $f_1, f_2, \dots, f_n$ and associated dwell times $d_1, d_2, \dots, d_n$. Each fixation represents the grouping of at least 3 temporally sequential raw eye-position points within 0.5° of visual angle of each other and totaling at least 100ms of gaze time.

Customized software performed a preliminary clustering of the fixations $f_1, f_2, \dots, f_n$ to a smaller set of *small-clusters* $s_1, s_2, \dots, s_N$ ($N < n$) using the following algorithm (11). If at least three consecutive fixations ($f_i, f_{i+1}, f_{i+2}, \dots$) fell within a circle of radius 2.5° , their locations were averaged and assigned to the first *small-cluster* s_1 and the corresponding cumulated dwell $d_{s_1}^s$ was calculated. The first fixation that violated the spatial proximity criteria was used to start a second small-cluster s_2 , which was only completed (and recorded) if at least 3 consecutive fixations contributed to it. Note that subsequent small-clusters could be spatially proximal to the first small-cluster s_1 , but they are assigned to different small-clusters, e.g., s_3 , because the fixations comprising them were not temporally sequential to those comprising s_1 . Next, *big clusters* $b_1, b_2, \dots, b_L$ ($L < N$) were generated by identifying small-clusters such as $s_1, s_3, \dots$ that were spatially within 2.5° of each other. The

corresponding cumulative dwells are denoted $d_{b_1}^b, d_{b_2}^b, \dots, d_{b_L}^b$. Only those big-clusters with associated dwell > 800 ms were used in the final analysis. The threshold of 800 ms was chosen because it has been suggested as being the minimum processing time required for detection, identification and resolution of any perceived findings (12). Figure 2 depicts an example of the search strategy of a radiologist as he/she inspected a given case. In this figure we show the locations of fixation clusters, aka, small clusters (a), as well as the locations of big clusters (b), and finally that of big clusters whose dwell is > 800 ms (c).

The ET perceptual marks were separated into two groups:

- i. *ET Lesion Localizations (ET LLs)*: these were big-clusters that occurred within 2.5° of visual angle from the center of a malignant lesion;
- ii. *ET Non-Lesion Localizations (ET NLs)*: all other big-clusters.

Measures of Visual Attention

At each big cluster location the following eye-position quantities were calculated:

- i. *Dwell time (D)*: this was defined as the cumulative gaze in seconds of all fixations that comprised the big-cluster with total dwell exceeding 800 ms.
- ii. *Approach-rate (A)*: this was defined as the reciprocal (s^{-1}) of shortest time-to-hit a big-cluster with total dwell exceeding 800 ms, i.e., approach the center of the big-cluster to within 2.5° . The reciprocal is taken to maintain a common directionality. In most cases greater perceptual attention is expected to be accompanied by greater approach-rate a and larger values of dwell d (the exception to this occurs for very large lesions, which “pop-out” from the surrounding background but do not need much cognitive processing to be resolved – in this case, dwell is not expected to be long).

Dwell time has been linked to the amount of cognitive processing at a given location, and a dwell threshold has been proposed to separate the different types of errors of omission (False Negative outcomes) (13). Approach-rate can be thought of as a perceptual measure of how much a perceived area “pops-out” from the background, and it has been shown to be significantly related to the likelihood that a given breast cancer will be reported by radiologists (14), with greater approach-rates being related to correct decisions (15).

Generalized ratings, figures-of-merit, agreement and confidence intervals

The eye tracking paradigm is conceptually similar to the FROC paradigm in the sense that both yield decisions at locations found by the observer. In effect, the big-clusters are regarded as *virtual marks*. In the FROC paradigm the observer consciously marks regions that are considered sufficiently suspicious for presence of a lesion, and the degree of suspicion is recorded by the rating $r (= 1, 2, \dots, 5)$. Analogously, eye-tracking yields the locations of regions that attracted visual attention long enough to allow a conscious decision to be made at the location (the big-clusters), and for each region there is a dwell time and an approach-rate. Dwell and approach-rate can be regarded as generalized ratings. Just as a figure-of-merit can be defined from FROC mark-rating data, likewise figures-of-merit can be defined from the eye-tracking virtual marks and generalized ratings. Details are in Appendix 1, Eqn. 1, where three figures-of-merit are defined, $\theta_{R,j}$, $\theta_{D,j}$ and $\theta_{A,j}$, where R stands for ratings, D for dwell and A for approach-rate and j is the reader index. These are analogous to the area under the alternative FROC (AFROC) curve and give equal importance to all cases and NL marks on diseased-cases are not used. Each figure-of-merit ranges from 0 to unity, unlike the area under the area under the ROC curve, which ranges from 0.5 to unity.

A jackknife-based method for measuring individual case-level agreement between any pair of figures-of-merit is described in Appendix 2. Defined there are Γ_{RD} , Γ_{DA} , and Γ_{RA} , which measure agreement between ratings and dwell, dwell and approach-rate and ratings and approach-rate, respectively. Each agreement measure ranges from 0.5 (chance level agreement) to 1 (perfect agreement). A bootstrap-based method for obtaining confidence intervals for figures-of-merit and agreements is also described in Appendix 2. The two-sided Wilcoxon signed rank test was used to measure the significance of differences between matched pairs of variables, one pair per reader, such as numbers of marks, ratings, figures-of-merit and agreements.

Results

For each reader and the reader-average (last row), Table 1 lists the average numbers of physical (FROC) marks per image and the average numbers of virtual (ET) marks per image. They are listed separately for normal images (NOR) and abnormal images (ABN) and further split between NL marks and LL marks. On normal images only NL marks are possible but on abnormal images both NL and LL marks are possible. The average number of physical or virtual NL marks per image is independent of whether the image is normal or abnormal, e.g., $0.902 \sim 1.104$ and $3.709 \sim 3.667$ and the difference was not significant at the 5% level (two-sided Wilcoxon signed-rank test). While the numbers in the NL columns can exceed unity, the numbers in the LL columns never exceed unity. This is because the number of LLs cannot exceed the number of lesions and the dataset contained only 1 lesion per breast while the numbers of NLs are not similarly constrained. The average the number of physical NL marks is only about 24% of the number of virtual NL marks (e.g., $0.902/3.709 = 0.243$). The average number of physical LL marks is slightly (but not significantly) smaller than the average number of virtual eye-tracking marks (e.g., compare 0.716 and 0.735). In other words almost all perceived lesions are marked which, combined with the finding the number of marked NLs is much smaller than the number of perceived

NLs, is consistent with a higher sampling distribution for the LLs as compared to NLs. The behaviors evident in Table 1 can be understood using the search-model, to be discussed later.

Table 2 lists, using the same layout as in Table 1, the average generalized ratings for marked regions, i.e., rating, dwell, and approach-rate. Since they are on different scales, the ratings, dwells and approach-rates cannot be meaningfully compared to each other. For example, the average physical NL rating was 1.776, but this value is strongly influenced by the 1–5 rating scale that was used. If a 1–100 point scale had been used, the average physical NL rating would be much larger. Likewise, the dwell and approach-rate scales could be altered by a factor of thousand by using milliseconds instead of seconds as the unit of time. However, comparisons within a particular generalized rating are meaningful. For each reader the NL generalized ratings on normal images was smaller than the corresponding LL generalized ratings, as expected if the generalized ratings are to have any predictive ability in differentiating lesions from normal images. The average physical NL rating on abnormal cases was higher than that on normal cases (compare 2.644 with 1.776) and the same was true for dwell (compare 3.286 with 2.842) but the opposite behavior was observed for approach-rate (compare 0.690 with 0.796). The difference was significant for ratings ($p = 0.008$) but not for dwell or approach-rate. Although the difference was not significant for approach-rate, the direction is consistent with the perceptually-based hypothesis that true lesions have a greater “attractive” effect on visual attention than other regions.

Table 3 lists the figures-of-merit θ values for rating R , dwell D and approach-rate A . The last row lists the reader averages and 95% bootstrap confidence intervals. Unlike Table 2, the figures-of-merit in Table 3 are directly comparable to each other – they are invariant to the scale factor effects that make direct comparisons of two different generalized ratings meaningless. In fact the figures-of-merit are invariant to any monotonic increasing transformation of the generalized ratings (16). For individual readers the θ defined over all images was largest for rating, followed by dwell and least for approach-rate ($\theta_R > \theta_D > \theta_A$). The Wilcoxon signed-rank test was significant ($p = 0.008$) for all comparisons except dwell vs. approach-rate.

Table 4 lists for each reader the agreement measures Γ and the last row lists the reader-average (AVG) agreement and the 95% bootstrap confidence bounds. Since all lower bounds exceeded 0.5, all agreements were significant at the 5% level. The ratings vs. approach-rate agreement values were significantly smaller than the corresponding dwell vs. approach-rate values (paired t-test, $p = 0.008$).

The values in Tables 1 can be understood using the theoretical framework provided by the search-model (17, 18). The number of perceived regions per normal image is 3.7, the λ -parameter of the search model. An estimate of the fraction of actual lesions that are perceived is 0.735, the ν -parameter of the search model. Since only about 24% of the number of virtual NL marks are actually marked, and since the search-model assumes a unit normal distribution for the NL confidence levels, an estimate of the lowest threshold would be $\zeta_1 = -\Phi^{-1}(0.243) = 0.697$, where Φ^{-1} is the inverse of the normal cumulative distribution function (e.g., $-\Phi^{-1}(0.0275) = 1.96$). Since almost all perceived lesions are marked and since the search model assumes a unit variance normal distribution centered at μ for the LL ratings, a rough estimate of the separation between the two Gaussian distributions characterizing the ratings distributions of NLs and LLs would be two standard deviations above the lowest threshold, i.e., 2.697, the μ -parameter of the search model. Using these estimates the search-model predicted value of the AFROC figure-of-merit is equal to 0.68, which is within the confidence interval indicated in Table 3. The website

www.devchakraborty.com has software for calculating the figure-of-merit for specified values of search model parameters.

Discussion

The experimental methodology used in this study resembled, as much as possible, the radiologists' clinical task, and differed essentially from that typically employed in past ET studies. Traditionally, eye-tracking recording is started at image onset and continues while the reader examines the image – no reader interactions with the display are permitted during this interval. When the reader indicates that a decision has been made, the eye-tracking recording is terminated and the reader renders a decision, e.g., case is “normal” or “abnormal”, sometimes followed by a confidence level in the decision. This is then followed by the marking of the location(s) of any perceived lesion(s) (10). A drawback to this design is that the reader may find other relevant regions during the reporting of the lesions perceived during the search phase, but this data is lost because eye-position collection has been terminated for the case. In the current methodology data collection begins at image onset and continues until the reader has indicated that he/she is ready to move on to the next case, and data entry occurs concurrently with the eye-tracking recording, as the reader is free to interact with the display (window-level adjustments are allowed) and mark and rate zero or multiple (1 or more) suspicious regions. Modifications, including deletion, to marks and ratings are permitted. The primary advantage of this methodology is that it permits simultaneous capture of FROC physical marking data and ET “virtual” marking data throughout case reading, with no investigator initiated interruptions. In our opinion this makes for a richer data set and we have by no means exhausted its analysis. In spite of the different experimental methodology, the values reported here are consistent with what previous studies have shown. For example, for LLs in ABN cases, time-to-hit was on average 0.84 sec, which agrees with Ref. (14). The dataset and an R-program implementing the analysis are available on the referenced website.

Ultimately the role of the radiologist is to separate normal from abnormal patients. Because process leads to performance, one expects agreement between the FROC and ET measures. Approach-rate had weaker agreement with ratings than dwell, Table 4. This is reasonable because in normal cases, radiologists still fixate somewhere, which results in an approach-rate to a normal area, which if not marked pulls up the FROC figure-of-merit, but pulls down the approach-rate figure-of-merit. In terms of dwell, radiologists do not dwell long on normal cases, while they do dwell longer on locations where they report the presence of a lesion: if that report is correct it pulls up both figures-of-merit. In the authors' opinion, the agreement would have been stronger had it not been for the relative spatial inaccuracy of eye-tracker identified suspicious regions (roughly 2.5 degrees or 200 pixels) compared to those marked by the radiologists (tens of pixels).

Although the average number of physical or virtual NL marks per image is independent of whether the image is normal or abnormal, Table 1, NLs on abnormal images were rated significantly higher than those on normal images, Table 2, suggesting that they are sampled from a distribution centered higher than 0, but not as high as μ . This finding is consistent with the hypothesis that the presence of the lesion creates a disturbance in the radiologist's perception, even if the lesion is not seen. The radiologist senses “there is something here out of the ordinary” and as a result the suspicious regions tend to be rated higher than suspicious regions on normal cases. This behavior is not predicted by the search-model, and shows one way that ET data could be used to extend the search-model to more realistic clinical data.

The fact that truth was ascertained by only one mammographer could be viewed as a potential weakness of the study. However, this reader had access to all additional materials

for the patients, such as additional imaging (spot compressions, ultrasound) and the pathology reports where the biopsies were carried out, so the degree of uncertainty regarding truth is felt to be quite small.

Because for each case the radiologists viewed two views of a breast two figures-of-merit can be defined - *view-based* and *case-based* - that differ depending on how the two views are handled in the scoring (19). In the view-based figure-of-merit the two views are treated as a single “big” image, and the total number of lesions is the sum of the number of lesions reported in the truth table in each view. Thus, although the dataset contained only images with one lesion per affected breast, in the view-based figure-of-merit the total number of lesions per case could equal 2. All lesion localization ratings are used in view-based scoring, with the possibility that the same physical lesion is rated twice, once in each view. In the case-based figure-of-merit the two views are regarded as a quasi-three-dimensional representation of the breast, and two marks on different views that correspond to the same physical lesion are regarded as a single mark, and the higher of the two ratings is assigned to it. All results in this paper are for case-based figures-of-merit. Furthermore, it is possible to define figures-of-merit over all cases, as done in this paper, or over abnormal cases only (the LL ratings are compared to NLs on abnormal cases). Extended tables showing these results are available on request from the authors.

Conclusions

Leveraging current methods developed for analyzing observer performance data could complement current ways of analyzing ET data and lead to new insights.

Acknowledgments

This work was supported by grants NIH/NIBIB R01 EB008688 and AHRQ K01 HS018365.

References

1. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*. 1978; 8(4):283–98. [PubMed: 112681]
2. Metz CE. ROC Methodology in Radiologic Imaging. *Investigative Radiology*. 1986; 21(9):720–33. [PubMed: 3095258]
3. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr Eng*. 1978; 4(4):166–71.
4. Chakraborty DP. New Developments in Observer Performance Methodology in Medical Imaging. *Semin Nucl Med*. 2011; 41(6):401–18. [PubMed: 21978444]
5. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. *Med Phys*. 2004; 31(8):2313–30. [PubMed: 15377098]
6. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine*. 1987; 6:411–23. [PubMed: 3114858]
7. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine*. 1996; 156(2):209–13. [PubMed: 8546556]
8. Kundel HL, Nodine CF, Krupinski EA. Computer-Displayed Eye Position as a Visual Aid to Pulmonary Nodule Interpretation. *Investigative Radiology*. 1990; 25:890–6. [PubMed: 2394571]
9. Krupinski, EA.; Nodine, CF. Gaze duration predicts the locations of missed lesions in mammography. In: Gale, AG.; Astley, SM.; Dance, DR.; Bowyer, K., editors. *Digital Mammography: Proceedings of the 2nd International Workshop on Digital Mammography*. Elsevier Science B.V; 1994. p. 399–403.

10. Nodine C, Mello-Thoms C, Kundel H, Weinstein S. Time course of perception and decision making during mammographic interpretation. *AJR*. 2002; 179:917–23. [PubMed: 12239037]
11. Nodine CF. Recording and analyzing eye-position data using a microcomputer workstation. *Behaviour research methods, instruments, & computers*. 1992; 24(3):475–85.
12. Hillstrom A. Repetition effects in visual search. *Percept Psychophys*. 2000; 2:800–17. [PubMed: 10883586]
13. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol*. 1978; 13:175–81. [PubMed: 711391]
14. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. *Academic Radiology*. 2008; 15(7):881–6. [PubMed: 18572124]
15. Mello-Thoms C, Hardesty LA, Sumkin JH, Ganott M, Hakim C, Britton CA, et al. Effects of lesion conspicuity on visual search in mammogram reading. *Acad Radiol*. 2005; 12:830–40. [PubMed: 16039537]
16. Hanley JA. The Robustness of the “Binormal” Assumptions Used in Fitting ROC Curves. *Med Decis Making*. 1988; 8(3):197–203. [PubMed: 3398748]
17. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol*. 2006; 51:3463–82. [PubMed: 16825743]
18. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol*. 2006; 51:3449–62. [PubMed: 16825742]
19. Yoon HJ, Zheng B, Sahiner B, Chakraborty DP. Evaluating computer-aided detection algorithms. *Medical Physics*. 2007; 34(6):2024–38. [PubMed: 17654906]
20. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*. 1997; 4(1):49–58. [PubMed: 9040870]
21. Chakraborty DP, Hagood TM, Ryan J, Marom EM, Evanoff M, McEntee MF, et al. Quantifying the clinical relevance of a laboratory observer performance paradigm. *Br J Radiol*. 2012; 57:2873–904.
22. Chakraborty DP. Measuring agreement between ratings interpretations and binary clinical interpretations of images: a simulation study of methods for quantifying the clinical relevance of an observer performance paradigm. *Phys Med Biol*. 2012; 57:2873–904. [PubMed: 22516804]

Appendix 1

Figures-of-merit derived from FROC and eye-tracking data

In what follows the uppercase character X refers to a *generalized rating*, where X can be R for ratings, D for dwell or A for the approach-rate. The corresponding lowercase character x denotes the *realized* generalized rating, where x can be r , d or a , corresponding to realized values of R , D or A .

Notation

Readers are indexed by j where $j = 1, 2, \dots, J$ and J is the total number of readers. Cases are indexed by k_t where t indicates the disease-status at the case (patient) level, with $t = 1$ for disease-free cases and $t = 2$ for diseased cases; k_1 ranges from 1 to N_N for disease-free cases and k_2 ranges from 1 to N_A for diseased cases. Marks are indexed by l_{sS} where s indicates the truth at the site (location) level, with $s = 1$ for a non-lesion localization and $s = 2$ for a lesion localization; $l_1 = 1, 2, \dots$, indexes marks of type $s = 1$ and $l_2 = 1, 2, \dots, N_{k_2}$, indexes marks of type $s = 2$, where N_{k_2} is the total number of lesions visible to the truth-panel on image k_2 . Since the dataset contained one true lesion per diseased case, and case-based scoring was used $N_{k_2} = 1$; $x_{jk_t l_{sS}}$ denotes the realized generalized rating of mark l_{sS} for reader j and case k_t . If the case has no marks of type $s = 1$, the corresponding x is assigned the value - 2000 as are the x 's of unmarked lesions.

Figures-of-merit

In case-based scoring the two views are regarded as a quasi-three-dimensional representation of the breast, and two marks on different views that correspond to the same physical lesion are regarded as a single mark and the higher of the two ratings is assigned to it. The figure-of-merit θ is defined by

$$\theta_{x,j} = \frac{1}{N_N N_A} \sum_{k_1=1}^{N_N} \sum_{k_2=1}^{N_A} \psi \left(\max_{l_1} (x_{jk_1 l_1 1}), \max_{l_2} (x_{jk_2 l_2 2}) \right) \quad (1)$$

The subscript X corresponds to the generalized rating. The kernel function ψ is defined as unity if the second argument exceeds the first, 0.5 if the two are equal and 0 otherwise; for reader j , $\max_{l_1} (x_{jk_1 l_1 1})$ is the maximum over the set of all non-lesion localizations of the generalized ratings on disease-free case $k_1 1$, or -2000 if the set is empty; $\max_{l_2} (x_{jk_2 l_2 2})$ is the maximum lesion-localization rating over the two views on diseased case $k_2 2$, or -2000 if the lesion is not marked in both views. Examination of Eqn. (1) reveals that if all lesions are marked and none of the normal cases is marked, then the figure-of-merit is unity (because the kernel function ψ yields unity for every term in the summations). In the other extreme, if none of the lesions is marked, and every normal image has at least one mark, the figure-of-merit is zero (because the kernel function yields zero for every term in the summations). These are extreme cases, and in general the figure-of-merit will range from zero (worst performance) to unity (perfect performance), unlike the area under the ROC curve, which ranges from 0.5 to unity.

Appendix 2

Measuring agreement between pairs of figures-of-merit

The jackknife pseudo-value θ_k^* corresponding to a figure-of-merit θ is defined by (we temporarily drop other indices)

$$\theta_k^* = K\theta - (K-1)\theta_{(k)} \quad (2)$$

Here $K = N_N + N_A$ is the total number of cases, θ is the figure-of-merit using all cases, $\theta_{(k)}$ is the figure-of-merit when case k is removed (jackknifed) from the analysis and k runs from 1 to K .

Hanley and Hajian-Tilaki (20) showed the pseudo-value θ_k^* can be regarded as the contribution of case k to the summary statistic θ . This interpretation of the pseudo-values has been extended (21, 22) where it is shown that the pseudo-value yields a measure of the *correctness* of the decision on a case.

The fluctuation of the pseudo-values $\Delta\theta_k^*$ is defined by (the dot represents the average over cases)

$$\Delta\theta_k^* = \theta_k^* - \theta_{\bullet}^* \quad (3)$$

One expects positive fluctuations to be associated with images on which high confidence correct decisions were made (be they FROC or eye-tracking derived) and negative fluctuations with cases on which high confidence incorrect decisions were made. The

agreement measure $\Gamma_{XX', ALL}$ between generalized ratings X and X' (where $X \neq X'$) is defined as follows:

$$\Gamma_{XX'} = \frac{\sum_{j=1}^J \sum_{k=1}^K \psi\left(0, \Delta\theta_{X,jk}^* \Delta\theta_{X',jk}^*\right)}{JK} \quad (4)$$

The summation over j and division by J results in an average over all readers, and likewise the summation over K and division by K results in an average over cases. The subscript XX' denotes an agreement defined between generalized ratings X and X' . If positive fluctuations of $\Delta\theta_{X,ALL,k}^*$ are accompanied by positive fluctuations of $\Delta\theta_{X',ALL,k}^*$ and negative with negative, then agreement will be perfect because the second argument of the kernel function will always exceed the first, leading to $\Gamma_{XX', ALL} = 1$. Chance level agreement would correspond to $\Gamma_{XX', ALL} = 0.5$, since about half the time the fluctuations will be in opposite directions. Three agreement measures can be defined between FROC and eye-tracking figures-of-merit: Γ_{RD} , Γ_{DA} and Γ_{RA} . One can also define agreement for an individual reader – one simply omits the reader averaging step. For example, the agreement measure corresponding to Eqn. (4) for reader j is defined by

$$\Gamma_{XX',j} = \frac{\sum_{k=1}^K \psi\left(0, \Delta\theta_{X,jk}^* \Delta\theta_{X',jk}^*\right)}{K} \quad (5)$$

Confidence intervals for FOMs and agreement

We randomly bootstrapped the data 200 times (i.e., re-sampled with replacement) readers and cases (i.e., both were treated as random factors). For each bootstrapped dataset the calculations described above were performed to estimate the figures-of-merit and agreement indices, for example $\Gamma_{XX',b}$, where b is the bootstrap index, $b = 1, 2, \dots, B = 200$. From the resulting B -dimensional array the lower and upper cutoff values, such that 95% of the observed values were contained within that range, were determined – this is the 95% confidence interval for $\Gamma_{XX'}$. The confidence intervals were not overly sensitive to increasing the number of bootstrap to $B = 500$.

Fig. 1.

Schematic of the data collection and processing to obtain real and virtual marks: the radiologists interpreted the images using a two-monitor workstation. Concurrently, and for the duration of the interpretation, an ASL eye-position tracking system determined the line-of-gaze. The ASL fixation and clustering algorithms are described in the text. The proximity criterion, defined as 2.5° of visual angle, is the maximum distance between a lesion center and a mark for the mark to be considered a LL (correct localization). Non-lesion localizations are all other marks. ASL = Applied Sciences Laboratory; NL = non-lesion localization; LL = lesion localization.

Fig. 2. Example of (a) small-clusters in yellow, (b) big-clusters, prior to threshold, in green and (c) big-clusters, after applying the 800 ms threshold, in blue. The green 'diamond' symbol marks the location where search started, whereas the red 'cross' symbol indicates the locations marked by the radiologist as containing a malignant lesion. The small yellow dots

mark the raw eye-position data. The red circles mark the true locations of the lesions in the two views.

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

This summarizes the average numbers of physical FROC NL and LL marks per image and virtual eye-tracking (ET) NL and LL marks for individual readers split between normal (NOR) and abnormal (ABN) cases. On normal images only NL marks are possible but on abnormal images both NL and LL marks are possible. The final row lists the averages (AVG) of the corresponding columns.

Reader	Physical FROC marks			Virtual ET marks		
	NL		LL	NL		LL
	NOR	ABN		NOR	ABN	
1	0.607	0.305	0.831	3.557	2.695	0.864
2	1.098	1.034	0.695	5.852	4.898	0.695
3	0.574	0.881	0.627	4.738	5.475	0.797
4	0.541	0.797	0.678	3.475	2.831	0.831
5	2.311	3.237	0.814	6.230	7.000	0.831
6	0.770	1.017	0.780	2.328	2.814	0.627
7	1.082	1.017	0.678	2.820	2.831	0.593
8	0.230	0.542	0.627	0.672	0.797	0.644
AVG	0.902	1.104	0.716	3.709	3.667	0.735

Table 2

This summarizes average ratings, dwells and approach-rates per image for the individual readers, split between normal (NOR) and abnormal (ABN) cases, and between non-lesion localizations (NL) and lesion localizations (LL). final row lists the averages (AVG) of the corresponding columns.

Reader	Rating R			Dwell D (s)			Approach-rate A (s ⁻¹)		
	NOR	ABN	LL	NL	ABN	LL	NOR	ABN	LL
1	1.649	2.833	2.918	2.131	2.109	4.701	0.684	0.725	0.907
2	1.851	2.164	3.073	3.524	3.481	5.334	0.488	0.525	1.186
3	1.657	2.212	3.324	2.338	2.463	4.210	0.581	0.414	1.106
4	1.242	2.085	3.275	2.435	3.230	6.699	0.675	0.470	1.270
5	1.291	2.199	3.104	3.464	3.799	8.251	0.541	0.394	1.054
6	3.277	3.933	4.130	2.617	2.863	5.626	0.697	0.581	0.903
7	1.742	3.100	3.600	3.499	3.331	3.638	0.780	0.894	0.835
8	1.500	2.625	3.459	2.464	3.569	6.193	0.894	0.870	1.664
AVG	1.776	2.644	3.361	2.842	3.286	6.276	0.796	0.690	1.185

Table 3

FOM values: the last row lists the average over all readers and the bootstrap 95% confidence interval (CI). The ratings FOM was largest followed by the dwell and then by the approach-rate. The differences between the FROC mark-ratings FOM and the perceptual FOMs were significant ($p < 0.05$).

Reader	Rating	Dwell	Approach-rate
	θ_R	θ_D	θ_A
1	0.806	0.577	0.263
2	0.632	0.221	0.220
3	0.723	0.433	0.303
4	0.753	0.551	0.408
5	0.703	0.445	0.249
6	0.795	0.498	0.364
7	0.692	0.280	0.244
8	0.764	0.678	0.635
AVG (95% CI)	0.734 (0.65, 0.81)	0.460 (0.34, 0.59)	0.336 (0.25, 0.46)

Table 4

Agreements values: the last row lists the average over all readers and the bootstrap 95% confidence interval (CI). All pairwise agreements were significantly better than chance. The ratings vs. approach-rate agreement was significantly smaller than the dwell vs. approach-rate agreement ($p = 0.008$).

Reader	Rating vs. dwell	Dwell vs. approach-rate	Rating vs. approach-rate
	Γ_{RD}	Γ_{DA}	Γ_{RA}
1	0.808	0.583	0.592
2	0.550	0.683	0.533
3	0.683	0.675	0.592
4	0.817	0.692	0.608
5	0.617	0.558	0.542
6	0.708	0.733	0.608
7	0.675	0.717	0.575
8	0.800	0.983	0.800
AVG (95% CI)	0.707 (0.63, 0.88)	0.703 (0.60, 0.79)	0.606 (0.53, 0.68)