How Experience and Training Influence Mammography Expertise[1]

Calvin F. Nodine, PhD, Harold L. Kundel, MD, Claudia Mello-Thoms, MSEE
Susan P. Weinstein, MD, Susan G. Orel, MD, Daniel C. Sullivan, MD, Emily F. Conant, MD

Rationale and Objectives. The authors evaluated the influence of perceptual and cognitive skills in mammography detection and interpretation by testing three groups representing different levels of mammography expertise in terms of experience, training, and talent with a mammography screening–diagnostic task.

Materials and Methods. One hundred fifty mammograms, composed of unilateral cranial-caudal and mediolateral oblique views, were displayed in pairs on a digital workstation to 19 radiology residents, three experienced mammographers, and nine mammography technologists. One-third of the mammograms showed malignant lesions; two-thirds were malignancy-free. Observers interacted with the display to indicate whether each image contained no malignant lesions or suspicious lesions indicating malignancy. Decision time was measured as the lesions were localized, classified, and rated for decision confidence.

Results. Compared with performance of experts, alternative free response operating characteristic performance for residents was significantly lower and equivalent to that of technologists. Analysis of overall performance showed that, as level of expertise decreased, false-positive results exerted a greater effect on overall decision accuracy over the time course of image perception. This defines the decision speed–accuracy relationship that characterizes mammography expertise.

Conclusion. Differences in resident performance resulted primarily from lack of perceptual-learning experience during mammography training, which limited object recognition skills and made it difficult to determine differences between malignant lesions, benign lesions, and normal image perturbations. A proposed solution is systematic mentor-guided training that links image perception to feedback about the reasons underlying decision making.

Key Words. Breast radiography; education; radiology and radiologists.

One of the outstanding characteristics of an expert in radiology is the speed and accuracy with which he or she decides whether an abnormality is present on a medical image (1–3). Acquiring expertise in radiology requires specialized training, experience, and some degree of talent. How much and what kind of training and experience has been the subject of an organized body of research that has emerged from the field of artificial intelligence (4,5). In this study, we evaluated the influence of perceptual and cognitive skills in mammography detection and interpretation by comparing the performance of experienced radiologists (mammographers), radiology residents, and mammography technologists. The study focused on the performance of the radiology residents, who were receiving training and mentor-guided experiences during mammography rotations that presumably provided a basis for mammography expertise.

It is difficult to find a yardstick to quantify the experience required to achieve expertise in mammography, but one could consider a reading on each case that results in a diagnostic report as a learning-experience trial. This measure of experience ignores immediate feedback, which is important for perceptual learning but is typically absent in clinical practice. In the context of medicine, training consists of mentored experience in which the resident reads medical images and then reviews them with the mentor.

Acad Radiol 1999; 6:575–585

[1] From the Department of Radiology, 308 Stemmler Hall, 36th & Hamilton Walk, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6086 (C.F.N., H.L.K., C.M.T., S.P.W., S.G.O., E.F.C.), and the National Cancer Institute, Rockville, Md (D.C.S.). Received March 19, 1999; revision requested May 5; revision received May 21; accepted May 21. C.F.N. supported in part by USAMRMC grant DAMD17-97-1-7103. **Address reprint requests** to C.F.N.

This training is designed to build feedback into the mentor-guided reading experience, but feedback is neither immediate nor systematic once the resident enters practice. If, for the moment, each read-and-reported case is considered an experience trial, regardless of whether it has been accompanied by feedback, expertise in mammography translates roughly into an average case reading experience equivalent of about 10,000 cases over a period of 3 years (6). This amount of experience compares favorably with estimates of the number of games a chess player plays to reach grand master status (7). The average radiology resident's case reading experience in a mammography rotation over 4 years is about 650 cases, of which only a dozen or fewer may be actual cancers. This means that extensive reading experience after residency is required to reach proficiency as a mammographer. Thus, the amount of experience that a radiology resident receives in training is literally a drop in the bucket.

Visual search is important for detecting lesions in mammograms, but in experts this search skill seems to be specifically tuned for detecting breast lesions embedded in mammograms and does not transfer to similar search tasks in which hidden words and figures are embedded in pictorial scenes (8). It may not even effectively transfer to reading radiographs of areas outside of the breast. Efficient search skills make expert mammographers fast, accurate recognizers, classifiers, and decision makers. Eye-position studies have shown that experts are faster at detecting lesions in chest or breast x-ray images than are less expert observers and that visual-gaze duration (or dwell), which is assumed to reflect visual information processing, is related to decision outcome (6,9). In general, observers dwell longest on the areas in which they report abnormalities, whether their results are true-positive or false-positive. Areas considered negative receive the shortest dwell times. False-negative decisions are the exception. In many instances, observers dwell almost as long on areas containing abnormalities that they report as negative as they do on similar areas that they report as positive, suggesting that they consider these areas to be troublesome even though they reported them as negative.

The fact that cumulative dwell time predicts misses is important in the context of the present study, because it reflects the recognition and decision making that lead up to a diagnostic outcome in much the same way that decision time reflects the gathering of information that leads up to a localization decision. However, visuospatial localization of regions of interest obtained with eye-position recording cannot be derived from decision-time data. The analysis of visual dwell and its relation to information processing leading to a decision outcome suggests that chronometric analysis of the relationship between decision times and decision outcomes may compliment visual dwell data. Experimental psychology has studied reaction time, which is closely related to decision time in the present study, because it "can help one trace the time course of information processing in the human nervous system" (10).

If the goal of mentor-guided experience during resident training is to provide the basis for expertise in mammography, then an important question is: What kind of skills are acquired? Previous research has helped to identify three general areas in which experts skills operate: (*a*) visual search, (*b*) pattern and object recognition, and (*c*) decision making. Because a key characteristic of mammography expertise is the speed-accuracy relationship in decision outcome, the present study focused on how decision making changes as a function of training and experience by comparing groups of observers with different dimensions of speed and accuracy. This comparison entails measuring decision times of observers during mammographic interpretation on a digital workstation and analyzing their decisions by comparing them against a truth table.

Three questions were explored. First, how does performance change as a function of mentor-guided reading experience? Second, how does decision outcome relate to decision time for each decision event during image perception? Finally, what is the likelihood of true versus false decision outcomes over the time course of image perception and decision making? This last question was initially addressed by Christensen et al (11), who were interested in the relationship between what they called "search time" and "perception" in the interpretation of subtle abnormalities and nonpulmonary lesions in chest radiographs. Search time was defined as how long it took to identify an abnormality. Given the possibility of multiple abnormalities per image, there could be multiple decisions per image. Each decision was timed and counted as a decision event. Maximum search time per image was 4 minutes, but most decisions took 1.84–2.68 minutes on average. To compensate for the efficiency associated with faster search times, the actual search time was adjusted by covarying it with the number of decision events within the maximum allotted search time per image. So experienced readers (faculty radiologists) made statistically significantly more decisions in less time than inexperienced readers (radiology residents). By mapping the search times of decision events against a truth table, they were able to plot the time course of true- and false-positive decision outcomes. The analysis of time-

perception data revealed that true-positive results outpaced false-positive results throughout the time course of viewing for experienced readers, whereas false-positive results overtook true-positive results during the time course of viewing for inexperienced readers.

MATERIALS AND METHODS

The mammography test set consisted of craniocaudal (CC) and mediolateral oblique (MLO) paired views from 78 unilateral mammogram cases, for a total of 156 images. The images were digitized (Lumiscan model 100 digitizer; Lumysis, Sunnyvale, Calif) by using a 100-μm spot size. The mammograms were of a single breast and were selected by two mammographers (S.G.O., D.C.S.) from a database of mammography cases taken from the archive of the Hospital of the University of Pennsylvania. These mammographers were later used in the study, but over 2 years had elapsed prior to their testing, and each mammographer contributed only about half of the mammograms to the test set. The mammograms were assembled from cases classified by mammography assessment as normal for at least 2 years, cases classified by mammography assessment as benign and proved by biopsy results to be benign, and cases classified by mammography assessment as malignant and proved by biopsy results to be malignant. The test set contained 25 cases with 15 instances of malignant masses and 14 instances of malignant calcifications shown on both views, one instance of an architectural distortion underlying a malignancy on both views of one breast, and one instance of a single malignant calcification present on only one view. It also contained 24 cases with 12 instances of benign masses and 12 instances of benign calcifications shown on both views and 26 cases considered to be normal. In addition, three practice cases were included: two showing lesions on both views and one lesion-free normal case. For all cases, the two mammographers (S.G.O., D.C.S.) selected mammograms containing subtle benign and malignant lesions. Many of the normal mammograms contained ambiguous image perturbations and thus were considered "difficult normals" by the two mammographers.

The test set was displayed on a single 19-inch, gray-scale monitor (GMA 201, Tektronix, Beaverton, Ore) interfaced to a Sun Sparc 10 computer (Sun Microsystems, Sunnyvale, Calif). At the time of testing, the brightness of the monitor was 127 cd/m^2. This brightness level is low for current state-of-the-art mammography displays and may have led to higher than normal error rates, at least for the inexperienced viewers. Each display consisted of two views of a single breast displayed in the center of the monitor at 2,048 × 2,048-pixel resolution. The gray scale was adjusted for each image by the experimenters (C.F.N., H.L.K.) to a setting that covered the gray-scale range of the breast-only portion of the image. The CC view was shown on the left half of the display screen and the MLO on the right half of the display screen. This is not a typical format for reading mammograms, but we were interested in determining how well observers with different levels of expertise could locate lesions in two views.

Three groups of observers representing different levels of mammography training and reading experience participated: staff mammographers with more than 5 years' experience as dedicated breast imagers (n = 3); 2nd-, 3rd-, and 4th-year radiology residents undergoing a mammography rotation (n = 19); and radiology technologists with 1–9 years' experience in mammographic imaging, but no reading experience (n = 9).

The procedure for testing observers was similar to the interruption technique used by Berbaum et al (12) to obtain response times during visual search. However, the observers viewed the images on a workstation. Lesion identification and decision confidence was entered by "clicking" with a mouse-driven pointer on a menu called up at the time that a lesion was localized. The time from the onset of the display until a decision was made, referred to as decision time, was automatically recorded. The observers were told that they were being tested on their ability to screen for malignancy in a two-view mammographic display of a single breast. If a malignancy was detected, they were to move the cursor to the lesion location and click on it. This action recorded the lesion location and called up a special menu from which they could classify the lesion as a mass, calcification, or architectural distortion and could rate their level of suspicion of malignancy as definitely malignant, highly suspicious for malignancy, moderately suspicious for malignancy, or low suspicion of malignancy. If the observer decided that the two views displayed were free of malignancy, he or she clicked "Return to Routine Screening" on the general menu. If the observer detected a benign lesion, he or she was instructed to treat the mammogram as lesion-free and click "Return to Routine Screening." In addition to these instructions, observers were told to localize malignant lesions on both views, if possible, and to point to the center of masses or a group of calcifications. After three practice trials with the experimenter, to familiarize themselves with the

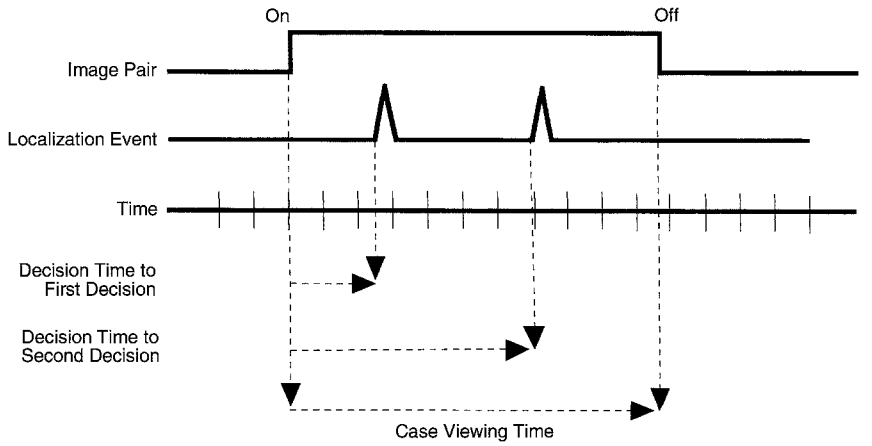

Figure 1. Diagram shows the relationship between image-display presentation and decision events signalled by the observer's clicking the location of a breast lesion on an image with the mouse. Decision time was measured from the onset of the image display to the onset of a decision event. Performance was measured for the task of reading a pair of breast images consisting of CC and mediolateral oblique MLO views. Therefore, more than one decision event was typically timed during each image-display presentation. Offset of the display occurred when the observer clicked on "Next Image."

workstation cursor operations, observers were left to view the 75-case test set on their own. Viewing time per case was unlimited. Decision times were recorded each time a lesion was localized by cursor control, but the observers were not told that their responses were being timed. Because multiple responses were made per two-view image pair, each localization event signaled the occurrence and time of a decision, indicating the presence of a true or false malignant lesion. Figure 1 shows how these events were translated into decision-time measures. For our analysis of decision times, we used the method of survival analysis to generate the cumulative time course of decision outcomes during the time course of viewing. Survival analysis has the advantage of adjusting individual decision times for decision outcomes per case by the total decision-making time required for a case. Thus, our analysis of decision times focused on the cumulative number of decision events per group over the time course of viewing. This is similar to the Christensen et al (11) analysis, which focused on the cumulative number of decision events per group over the time course of viewing 100 chest radiographs.

Analysis of cursor events for localizing, classifying, and rating lesions was accomplished by comparing the observers' decisions against a truth table. The truth table was generated from a combination of mammographic assessment by two of the authors (S.G.O., D.C.S.) and biopsy information on each case. Because all pairs of positive images contained at least two lesions, alternative free response operating characteristic (AFROC) analysis was carried out, treating the pair of positive images as the unit of analysis. This was consistent with the instructions for the task and provided evidence on how well observers with different levels of mammography expertise coordinated lesion localization in a second view, given lesion detection in the first view.

For the AFROC analysis, 30 pairs of malignant lesions were identified as appearing on 25 image pairs. These 60 lesions were counted in the malignant-positive category. The 24 image pairs containing benign lesions plus the lesion-free images (total of 50 image pairs) made up the nonmalignant category. In the AFROC analysis, we counted all correctly localized lesions within ±0.41 cm of the true location on the malignant two-view image pairs (2 standard deviations of mean accuracy of 0.28 cm for mammographers) and counted only the highest-rated false-positive results for the 50 nonmalignant image pairs (equivalent to counting false-positive images; see [13]). It should be noted that this performance criterion ignores classification information that we thought unreasonably stretched the assumptions underlying the two-alternative force choice experimental framework. Basically, the AFROC was designed to measure detection performance. However, because of the importance of the classification decision in mammography, we will provide a separate analysis of the classification data to show how this performance criterion is influenced by the level of expertise.

RESULTS

Overall Performance

Overall detection and localization of breast lesions was assessed as a function of level of expertise. We compared A1, the area under the AFROC curve, for mammographers, residents, and radiology technologists. The AFROC plots the fraction of actual target locations reported (true-positive decisions) against the fraction of images with any false-positive decisions. In our case, we plotted only the highest-rated false-positive decisions on normal or benign images. Figure 2 shows AFROC curves

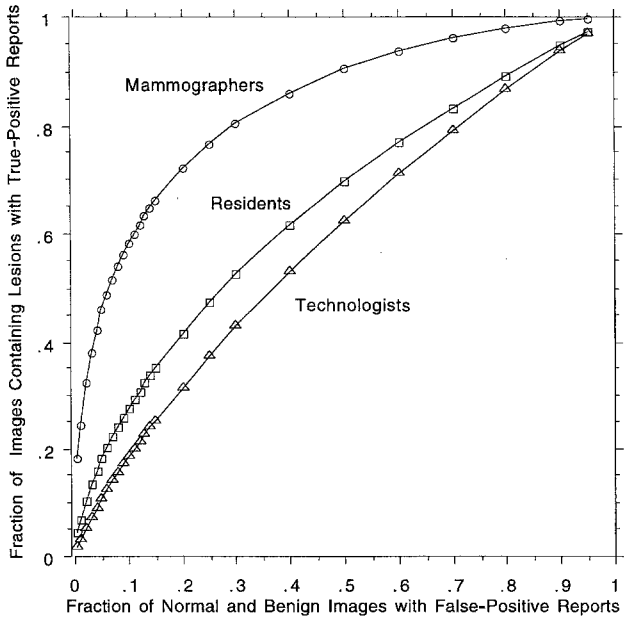

Figure 2. AFROC curves show mean decision accuracy for experienced mammographers (n = 3), radiology residents (n = 19), and mammographic technologists (n = 9). For this analysis, it was assumed that there were 60 malignant lesions on 25 image pairs (CC and MLO views) and 50 malignancy-free images. False-positive results were counted only for malignancy-free images. A computer program called ROCFIT was used to produce an ROC curve after averaging over the confidence intervals for each group of observers. (ROCFIT is part of a set of curve-fitting and estimation programs called ROCKIT, which is available at *http://www-radiology.uchicago.edu/sections* by clicking on "Kurt Rossman Laboratory" and then on "ROC Analysis.")

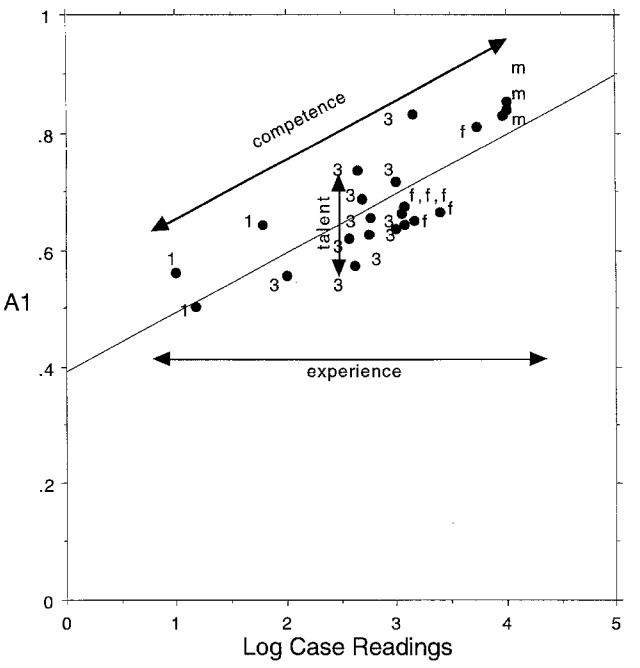

Figure 3. A regression analysis of overall performance measured as A1 as a function of $\log^{10}$ number of cases read over a 3-year period by three experienced mammographers and 19 radiology residents undergoing clinical mammography rotation. When case readings are zero, the regression line intercepts the y axis at A1 = 0.393, which is close to chance performance. With mentor-guided case reading training and experience, A1 performance increases. The numbers and letters within the figure indicate the level of training of the observers: *1* = 1st- and 2nd-year residents, *3* = 3rd- and 4th-year residents, *f* = fellows, and *m* = mammographers.

for the three groups. The average area per observer derived from analysis of variance of A1 values was 0.840 (standard deviation, 0.039) for mammographers, 0.653 (0.058) for residents, and 0.592 (0.062) for technologists. All of these values are above chance performance, which for the AFROC is 0.000. Analysis of variance of A1 values indicated, not surprisingly, that the overall performance accuracy of mammographers was statistically significantly better than that of either residents or technologists, who did not differ from one another ($P < .01$, Scheffe test). By contrasting performance for these groups, which represented different levels of training and experience, we hoped to gain insights into the nature of mammography expertise.

Relation of Case Reading Experience to Development of Mammography Expertise

To provide a clearer picture of how the three groups differ in their experience at reading mammograms, we obtained data on the number of mammographic reports generated by the residents and mammographers. The 19 radiology residents who were part of the study represented mainly 3rd-year (n = 7) and 4th-year (n = 8) residents, plus four fellows who had mammography reading experience varying from 10 to 2,465 cases over a 3-year interval. Over the same period, the three mammographers read 9,459 to 12,145 cases. The relationship between A1 and log number of cases read is shown in Figure 3 for all observers. Figure 3 shows a significant linear-regression fit of the data (R^2 = .667) with a positive slope, suggesting that reading skill, as reflected by A1 performance, increases directly with log case reading experience (F [1,22] = 44.15; $P <$ .0001). The regression line intercepts the y axis at A1 = 0.293, which implies close to chance performance with zero reading experience. A log scale was used to represent the effects of case reading experience because several investigators have suggested the relationship between practice and learning is best expressed by a power function (14). The range of case reading experience in Figure 3 was from 1.0 log case reading to 4.1 log

case readings, or from about 10 to 12,000 cases. Two residents at the beginning of mammography training with little case reading experience performed at an A1 of about 0.500. The fact that their performance is above chance at the beginning of the mammography rotation can be attributed to their talent and their subspecialty training in other areas of radiology. The training levels of the observers are indicated by the numbers or letters associated with the data points. Overall performance increases in an orderly progression with training level.

Identification of Lesions in Two Views

Our hypothesis was that one aspect of performance that might differentiate levels of expertise was how successful observers were at identifying pairs of lesions in a two-view (CC and MLO) display. This hypothesis was based on the assumption that when mammography experts detect a lesion in one view they look for confirmation in a different view. Mammographers talk about using projective geometry principles to predict from a detected lesion to a likely "plane of interest" in which to search for the corresponding "depth" lesion projection. If a detected lesion can be paired in a second view, this provides confirmation that it is a real target. To follow up on this, we analyzed malignant lesions (true-positive decisions) and benign lesions (false-positive decisions) that appeared on CC and MLO views per case by referring to the truth table. The identification of paired localizations on lesion-free areas of images (false-positive decisions) was more speculative, because these were imaginary. To account for paired localizations on lesion-free areas of images (false-positive decisions), we identified sequential decisions— from CC to MLO view or vice versa—that were classified as being malignant and of the same type (eg, mass, calcifications, or architectural distortion). Consistent with the pattern of results in the AFROC analysis, the identification of paired lesions was related to level of expertise. Proportionally more paired lesions were reported and correctly classified by the mammographers than by the residents or technologists. The proportion of correctly paired-lesions was 0.82, 0.56, and 0.50 for mammographers, residents, and technologists, respectively. Proportionally fewer lesions were seen and reported correctly in only one view by all groups, and the corresponding proportions were much lower—0.14, 0.14, and 0.12, for mammographers, residents, and technologists, respectively.

Decision Time and Decision Outcome

The regression plot in Figure 3 shows the relationship between performance and case reading experience. We hypothesized that the decision speed–accuracy relationship, which is a hallmark of expertise, should accompany this improvement in performance, so we looked at decision times as a function of decision outcome, again taking into account that observers were interpreting an image pair containing CC and MLO views and thus possibly making two or more decisions per case. To identify the sequencing of decisions per case, the paired decisions were broken down into those occurring in the CC view on the left side of the display and the MLO view on the right side of the display. For these paired decisions, decision times to the first decision were inversely related to level of expertise, with mammographers significantly faster than residents ($P < .01$, Scheffe test) and residents significantly faster than the technologists ($P < .0001$, Scheffe test). For mammographers compared with residents, 32% more of their initial responses were true-positive, and these initial responses were reported faster than those of residents. Mean decision time for the first correct decision per pair was 15.66 seconds versus 21.56 seconds (t [376] = 3.91; $P < .001$). Technologists detected fewer true-positive results and took even longer to decide (28.08 seconds). Decision time was also inversely related to level of expertise in a similar pattern for classification of localized lesions. Mammographers correctly classified 38% more lesions and did so faster than residents ($P < .05$) and technologists ($P < .001$). Mean decision time for mammographers was 16.51 seconds for classifying masses and 19.77 seconds for classifying calcifications. Both of these findings support the decision speed–accuracy relationship associated with expertise.

Finally, to provide some perspective on how true-positive results related to false-negative results, we looked at decision times when all lesions were completely missed on images containing malignant lesions. In this case, total image duration was assigned as the decision time. This result might be considered a "clean" miss in that no lesion was reported, even though a lesion was present during the entire time that the image was examined. Of 579 total false-negative decisions, 51% were clean misses. Mean decision times differed little for the clean-miss false-negative category, as they ranged from 38 to 46 seconds. However, standard deviations of the mean decision times ranged from 4.6 seconds for mammographers to 41.6 and 52.5 seconds for residents and technologists, respectively. These values indicated that the latter two groups had considerable indecision about not making a positive report after examining two views of an image containing a truly malignant lesion. The range of mean decision times for clean misses

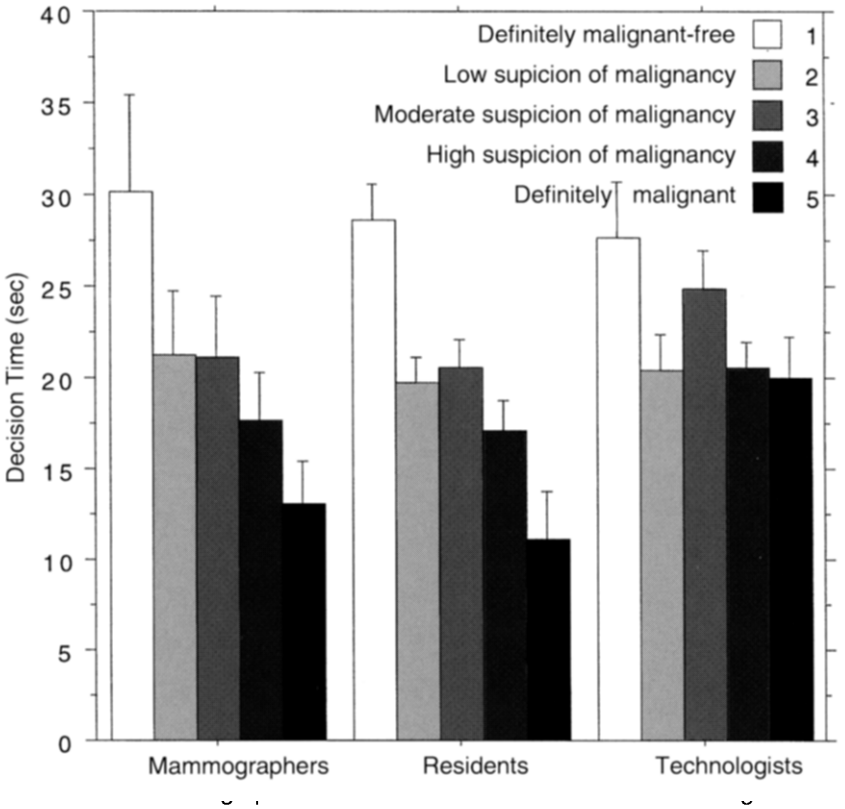

Figure 4. Decision time as a function of decision-confidence ratings for mammographers, residents, and technologists. A confidence rating of 5 indicated the lesion was definitely malignant; 4, highly suspicious for malignancy; 3, moderately suspicious for malignancy; 2, low suspicion of malignancy; and 1, definitely malignant-free.

was longer than that of any other decision outcome categories and seems to complement the finding obtained from monitoring eye position of prolonged visual dwell for false-negative decisions. Observers spent a longer time deciding to call a positive case negative. Overall, clean-miss false-negative decisions took significantly longer than true-negative decisions (t [864] = 4.22; P <.001). Of course, we cannot confirm that the true lesions were actually scrutinized from the decision time data, but the long decision times and wide variances suggest much uncertainty surrounding decision making.

Relationship of Decision Time to Use of Confidence Ratings

The similarity of the relationship of decision outcome to decision time for mammographers and residents suggests that they may be using similar underlying detection and decision strategies. One measure that reflects underlying decision strategy is how observers used the confidence ratings in making decisions. It is reasonable to assume that the more sure observers are that they have detected a lesion, the faster they are at making a decision. Figure 4 shows the relationships between decision time and use of confidence ratings for the three levels of expertise. The general pattern for the mammographers and residents was that decision times were inversely related to the confidence rating. The longest decision times were for definitely lesion-free images (rating = 1), and the shortest decision times were for definitely malignant image locations (rating = 5). This pattern suggests that both groups had a similar perceptual thresholding basis for the decision, which is an important factor in developing a decision-making strategy. The pattern for technologists showed virtually no relationship between decision time and use of confidence ratings. Only confidence 1 ratings were prolonged. No evidence showed that decision times were faster when technologists were more confident that a malignant lesion was present on an image.

Time Course of Decision Outcomes

So far, two interesting generalizations come out of the decision time analysis. First, the decision speed–accuracy relationship was found to be related to level of expertise. Figure 5 summarizes the decision speed-accuracy relationship expressed by d' (cumulative) as a function of viewing

time for mammographers, residents, and technologists. Cumulative values for true-positive and false-positive decisions to both normal and benign images on a per case basis (paired decisions) as a function of decision time were obtained from survival analysis. These values were then transformed using the formula $d' = z$ (true-positive decisions/30) – z (false-positive decisions)/50, where z can be interpreted as a deviate of the unit normal curve. This formula can be thought of as correcting the true-positive fraction by the false-positive fraction. Decision accuracy consists of detecting perturbations in images, testing them for signs of malignancy, and correctly classifying them as masses, architectural distortions, or calcifications. This complex decision requires discriminating malignant from benign lesions, and malignant from normal anatomic variants in the breast image. Decision accuracy can be expressed as A1, the area under the AFROC curve, or as d', the index of detectability derived from the true-positive fraction and the false-positive fraction at a specific decision threshold, as shown in Figure 5. Looking at performance in this way shows clear differences as a function of level of expertise.

Second, decision times were longer for false than for true decision outcomes. To consider whether these false decisions tended to occur early or late in the time course of image perception, we looked at both paired and single decisions. A paired decision is one in which the observer sequentially localized a suspected lesion (true or false) on both CC and MLO views. Figure 6 shows the mean number of paired true-positive decisions and paired false-positive decisions for normal regions of the images and benign lesions for mammographers, residents, and technologists as a function of viewing time per case. Figure 7 shows the same plot for single decisions, as contrasted with paired decisions. The most striking feature of Figure 6 is the technologists' high rate of false-positive results for normal regions in relation to the rate of their true-positive results, for paired decisions. In Figure 7, it is the high rate of false-positive results for normal regions for all groups for single decisions.

These plots show that for mammographers the rate of true-positive decisions for normal regions is faster than the rate for false-positive decisions, but false-positive decisions for normal regions continue to plague performance throughout the time course of viewing. False-positive decisions for benign lesions drop out relatively early; thus, overall performance continuously improves with decision time until about 60 seconds. Perhaps our mammographers should have considered stopping at this point, because

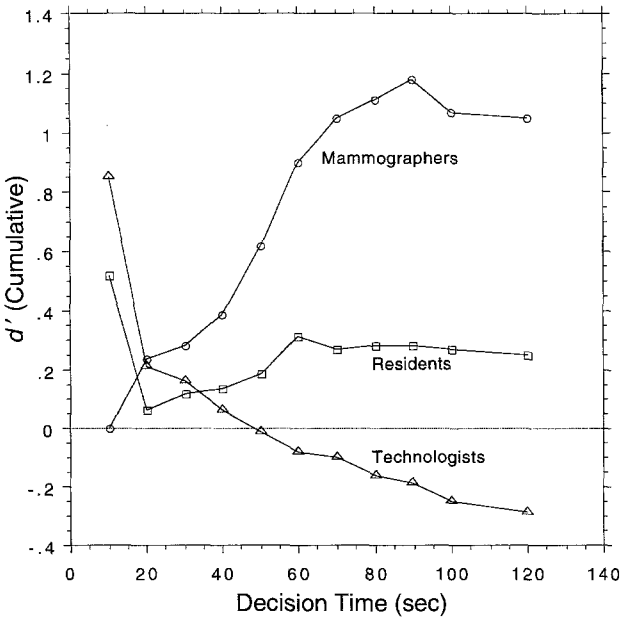

Figure 5. Speed-accuracy relationship indicated by d' as a function of decision time for mammographers, residents, and technologists. Overall performance measured by d', which is the normal deviate (z) of true-positive results minus the false-positive results, increased for mammographers and to a lesser extent for residents. Overall performance decreased below chance ($d' = 0$) for technologists, which means that false-positive results outnumbered true-positive results.

false-positive decisions for normal regions increased faster than true-positive decisions. The rate of true-positive decisions is slower for residents because of continuous competition from false-positive decisions for normal regions up to 60 seconds. As with mammographers, the false-positive decisions for benign regions peak earlier. The technologists show a decrease in overall performance over time because they continued to make more false-positive decisions for normal regions than true-positive decisions.

DISCUSSION

Understanding the Nature of Expertise

The goal of the present study was to understand better the nature of expertise in mammography. Expertise in mammography, as we have defined it here, refers to diagnostic performance skills that enable the observer to localize a breast lesion and correctly decide that it is or is not malignant on the basis of two views. Admittedly, our task was somewhat artificial in the sense that we mixed lesion detection, which is the focus of mammography screening, with diagnostic interpretation, which is the focus of diagnostic follow-up. The next step is to break the task apart and do a

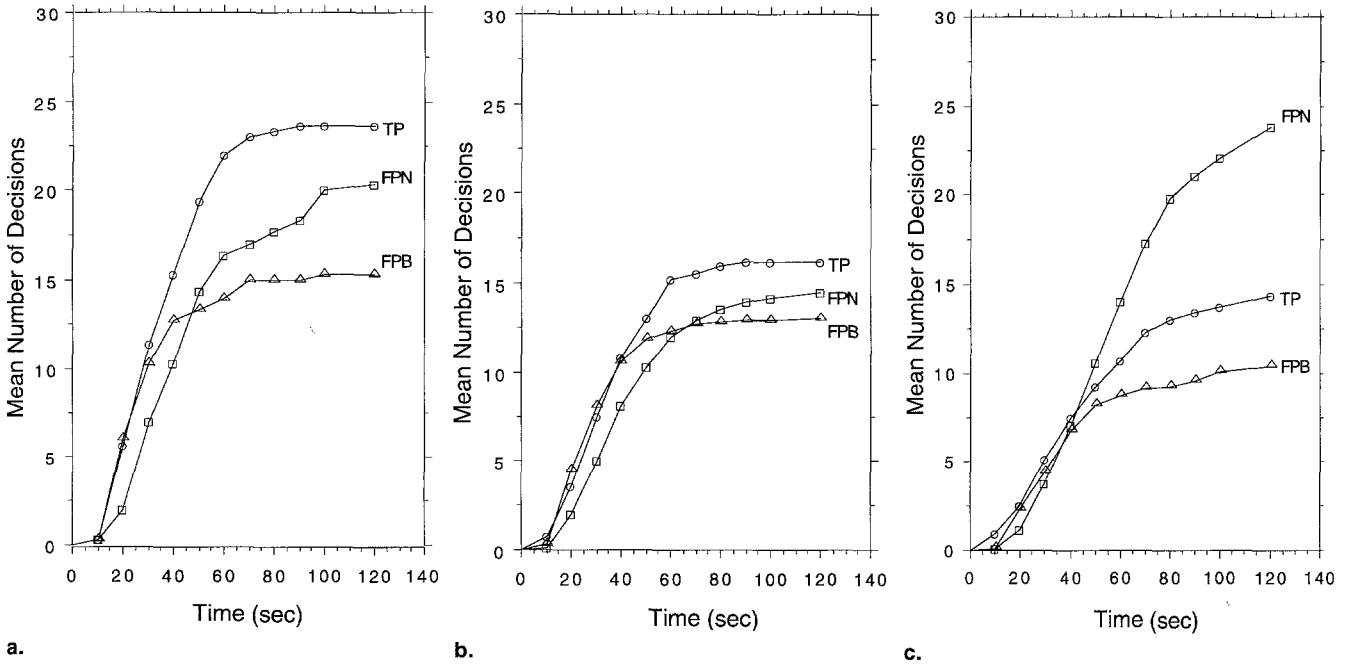

Figure 6. Cumulative mean numbers of paired decisions per case as a function of the decision time course of viewing for true-positive (*TP*) decision outcomes, false-positive decision outcomes on normal images (*FPN*), and false-positive decision outcomes on images containing benign lesions (*FPB*) for **(a)** mammographers, **(b)** residents, and **(c)** technologists. Paired decisions were measured. Of the malign cases, all but one contained lesions in both CC and MLO views. As this figure indicates, within 60 seconds, the mammographers had localized 23 (92%) of 25 paired true lesions.

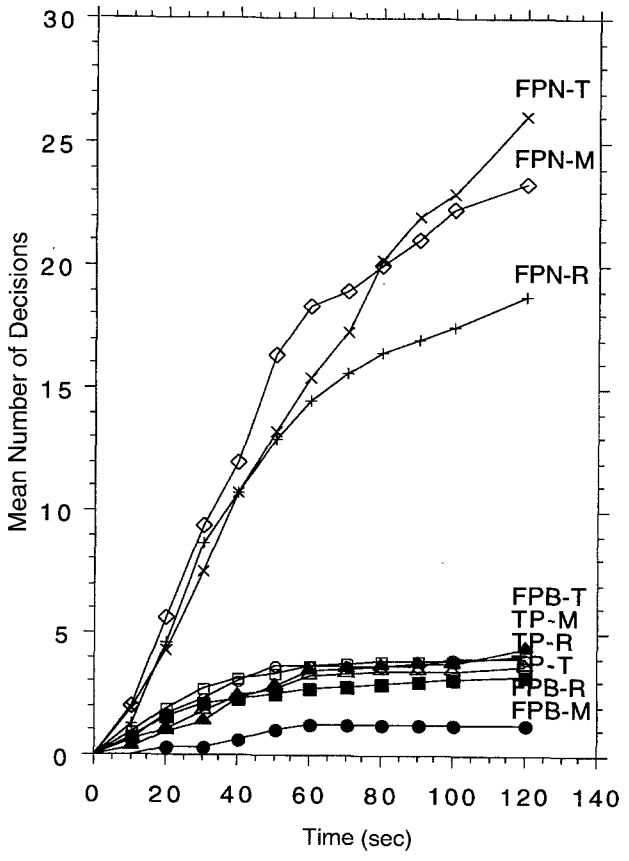

Figure 7. Cumulative mean number of single decisions as a function of the decision time course of viewing for true-positive (*TP*) decision outcomes, false-positive decision outcomes on normal images (*FPN*), and false-positive decision outcomes on images containing benign lesions (*FPB*) for mammographers (*M*), residents (*R*), and technologists (*T*).

two-part test, which will come closer to the American College of Radiology Breast Imaging Reporting and Data System format. Moreover, even though the diagnostic skills that we have studied are an essential part of mammography diagnosis, the study is limited, as only CC and MLO views were shown with no capability for prior studies, additional views, or magnification views. Additional special mammographic images, such as spot compression or magnification views, and breast ultrasound imaging, both of which are important parts of mammography expertise, were not employed in the present study. On the basis of the information these modalities provide, the mammographer may decide that the finding is normal, benign, or probably benign but recommend short-term follow-up or a biopsy.

Why Are Experts Faster and More Accurate?

Our analysis has related A1 and d', measures of overall performance, to decision time to shed light on basic

perceptual and decision-making skills. Differences in speed and accuracy between mammographers and residents seem to be related to the experience required to gain expertise, as shown in Figure 3. This suggests that experts are perceptually more sensitive in recognizing lesions than are those with less expertise because the experts have read more mammogram cases, seen more lesions, and differentiated more lesions into malignant and benign categories. In practical terms, this means that through massive amounts of experience experts became perceptually tuned to recognizing familiar common breast structures and detecting odd or novel variations in them. Three to 5 years of dedicated experience reading mammograms affects perceptual learning by exposing mammographers to a wide set of breast image configurations that represent most variations of normality and abnormality. We hypothesize that this concentrated case reading experience with mammographic images has an effect on perceptual learning by producing enhanced recognition skills akin to those attributed to chess grand masters who, according to one estimate, are capable of recognizing on the order of 50,000 different chess configurations (7). It is unclear whether enhanced object-recognition skill is the result of the development of what the artificial intelligence community refers to as "chunking" or template-retrieval structures that aid short-term and long-term memory (14) or, as we have suggested, more critically tuned visual recognition as the result of learning and refining distinctive-feature information used to recognize deviations from prototypic normal breast structures (15,16).

Supporting the argument in favor of the tuning of visual recognition, Sowden et al (16) have shown that massed practice detecting calcifications in positive-contrast mammograms (bright targets on a dark background) positively transfers to a new task in which the calcifications are displayed in negative-contrast mammograms (dark targets on a bright background). This suggests that perceptual learning improves perceptual sensitivity in the detection of low-contrast targets. Massed practice was defined as a detection trial followed immediately by feedback about the correctness of an observer's response. This improvement in perceptual sensitivity occurred even though the amount of massed practice was limited to 720 trials, followed by the transfer test. The key to improvement may be the feedback. Generalizing the results of Sowden et al (16), one cannot help but wonder if the effects of reading experience would be facilitated by computer-assisted visual feedback about decision outcomes delivered for some subset of test cases in which truth could be verified or, at least, agreement consensus reached. The purpose of systematic visual feedback is to make image perception and decision making an integral part of a perceptual-learning reading experience (6,17).

Expertise: Chest Radiology Compared with Breast Radiology

In interpreting performance differences, we have to be careful to separate studies of expertise in chest radiology from those in mammography, because chest radiology studies have emphasized the importance of input from peripheral vision in detecting pulmonary lesions. Peripheral vision is important during the search for inconspicuous pulmonary lesions because a chest radiograph contains so many anatomic landmarks (eg, heart, ribs, lungs, diaphragm). It has been suggested that these anatomic landmarks act as a map, helping peripheral guidance of search (18). Anatomic landmarks are few in the breast (eg, nipple and pectoralis muscle), and breast structures that might serve as landmarks (eg, blood vessels and ducts) are interwoven into the breast image, creating texture differences that are probably too subtle to be selected by peripheral vision during a search. As a consequence, rather than landmarks, we believe that perturbations in parenchymal structure caused by compression of the breast during imaging and desmoplastic reaction from a growing tumor provide focal points of interest during a visual search. The superimposition of parenchymal structures tends to make them visually conspicuous. Because the superimposition of parenchymal structures has the potential to mimic breast lesions, they may be detected by peripheral vision during the initial global survey, scrutinized during subsequent focal scanning, and falsely reported as true lesions. In the detection of breast lesions, it is not only important for the observer to recognize familiar features in the image but also to recognize odd or novel features, examine these in detail (as reflected by fixations and decision time), and weigh their importance in making a decision (6,19). We assume that dwell time spent fixating the lesion, like time spent examining the image prior to making a decision, represents the information processing time required to make a decision.

Decision Strategies

Our study has shown that residents develop decision-making strategies that are similar to those of experts. From a practical standpoint, this suggests that resident training in mammography is effective in providing a general framework for learning radiology image-perception skills. However, residents are not as good as experts at

identifying true breast lesions. We hypothesize that this weakness is due primarily to the lack of fine-tuned visual recognition skills. Because feedback is recognized as a critical part of the reading experience, built into the clinical mammography rotation, it is tempting to speculate that providing computer-assisted feedback training might facilitate visual recognition skills and bring resident overall performance closer to that of their mentors. Despite their limited perceptual experience, many of the radiology residents will join clinical practices and read mammograms as practicing radiologists. Does this mean that the diagnostic performance of practicing radiologists will suffer as a result? Probably, because the overall average performance of residents in the present study had an average receiver operating characteristic curve area of 0.743, which was 12% lower than the national average of 0.845 for 108 U.S. radiologists, assuming approximately the same level of case difficulty for the two test sets (20).

Finally, we have shown that decision accuracy is directly related to amount of case reading experience. At the present time, many radiology departments keep track of the number of cases read by radiologists and residents, yet no recommendations have been proposed as standards.

Our data support the need for minimum requirements in number of case readings, such as those proposed by the latest Food and Drug Administration regulations. As of April 28, 1999, this requirement was 240 case readings within the past 2 years of residency. In addition, we believe that some less abrupt transition between residency and practice (for example, double-reading experience during the 1st year of practice) would greatly improve performance standards (21).

REFERENCES

1. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, eds. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988; 311–342.
2. Kundel HL, La Follette P. Visual search patterns and experience with radiological images. Radiology 1972; 103:523–528.
3. Parasuraman R. Effects of practice on detection of abnormalities in chest x-rays. Proc Hum Factors Soc 1986; 309–311.
4. Newell A, Simon HA. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.
5. Chi MTH, Glaser R, Farr MJ. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988.
6. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996: 3: 1000–1006.
7. Chase WG, Simon HA. Perception in chess. Cognitive Psychol 1973; 4:55–81.
8. Nodine CF, Krupinski EA. Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. Acad Radiol 1998; 5: 603–612.
9. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol 1990; 25:890–896.
10. Posner MI. Chronometric explorations of mind. New York, NY: Oxford, 1986; 218.
11. Christensen EE, Murry RC, Holland K, Reynolds J, Landay MJ, Moore JG. The effect of search time on perception. Radiology 1981; 138:361–365.
12. Berbaum KS, Franken EA, Dorfman DD, et al. Time course of satisfaction of search. Invest Radiol 1991; 26:640–648.
13. Chakraborty DP, Winter LHL. Free-response methodology: alternative analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
14. Gobet F, Simon HA. Templates in chess memory: a mechanism for recalling several boards. Cognitive Psychol 1996; 31:1–40.
15. Myles-Worsley M, Johnston WA, Simons MA. The influence of expertise on x-ray image processing. J Exp Psychol Learn Mem Cogn 1988; 14:553–557.
16. Sowden P, Davies I, Roling P. Perceptual learning of the detection of features in x-ray images: a functional role for improvements in adults' visual sensitivity? J Exp Psychol Hum Percept Perform (in press).
17. Anderson JR. Cognitive psychology and its implications. 4th ed. New York, NY: Freeman, 1995.
18. Kundel HL, Nodine CF, Thickman D, Toto L. Searching for lung nodules: a comparison of human performance with random and systematic scanning models. Invest Radiol 1987; 22:417–422.
19. Ullman S. High-level vision: object recognition and visual cognition. Cambridge, Mass: MIT Press, 1996; 161.
20. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Med 1996; 156:209–213.
21. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.