Exploring the potential of context-sensitive CAdE in screening mammography

Georgia D. Tourassi,^{a)} Maciej A. Mazurowski, and Brian P. Harrawood

Department of Radiology, Carl E. Ravin Advanced Imaging Laboratories, Duke University Medical Center, Durham, North Carolina 27705

Elizabeth A. Krupinski

Department of Radiology, University of Arizona, Tucson, Arizona 85724

(Received 10 May 2010; revised 21 September 2010; accepted for publication 21 September 2010; published 12 October 2010)

Purpose: Conventional computer-assisted detection (CAdE) systems in screening mammography provide the same decision support to all users. The aim of this study was to investigate the potential of a context-sensitive CAdE system which provides decision support guided by each user's focus of attention during visual search and reporting patterns for a specific case.

Methods: An observer study for the detection of malignant masses in screening mammograms was conducted in which six radiologists evaluated 20 mammograms while wearing an eye-tracking device. Eye-position data and diagnostic decisions were collected for each radiologist and case they reviewed. These cases were subsequently analyzed with an in-house knowledge-based CAdE system using two different modes: Conventional mode with a globally fixed decision threshold and context-sensitive mode with a location-variable decision threshold based on the radiologists' eye dwelling data and reporting information.

Results: The CAdE system operating in conventional mode had 85.7% per-image malignant mass sensitivity at 3.15 false positives per image (FPsI). The same system operating in context-sensitive mode provided personalized decision support at 85.7%–100% sensitivity and 0.35–0.40 FPsI to all six radiologists. Furthermore, context-sensitive CAdE system could improve the radiologists' sensitivity and reduce their performance gap more effectively than conventional CAdE.

Conclusions: Context-sensitive CAdE support shows promise in delineating and reducing the radiologists' perceptual and cognitive errors in the diagnostic interpretation of screening mammograms more effectively than conventional CAdE. © 2010 American Association of Physicists in Medicine. [DOI: [10.1118/1.3501882](https://doi.org/10.1118/1.3501882)]

Key words: computer-aided detection (CAD), mass detection, screening mammography, eye-tracking

I. INTRODUCTION

Although the use of computer-assisted detection (CAdE) technology for screening mammograms has been steadily increasing across clinical practices for the past 12 years, there is ongoing controversy regarding its true clinical impact.^{1,2} The controversy stems from mixed reports on how CAdE affects the radiologist's diagnostic performance in an actual clinical setting as compared to laboratory performance. Published studies on the topic report dramatically different findings, ranging from a highly beneficial to a detrimental effect. On one hand, CAdE was strongly supported as a clinical alternative to double-reading by two recent large-scale prospective studies, one based on 321 221 mammograms³ and the other based on 31 057 mammograms.⁴ On the other hand, two earlier widely publicized studies reported no statistically significant change⁵ or even reduction on radiologists' accuracy when using CAdE.⁶ A recent editorial⁷ has delineated several factors that may explain the mixed findings, namely, differences in study design, differences in selected end points, and differences in the experience level of the partici-

pating radiologists. In any case, the disparate findings call into question the translation of CAdE from the research environment to the clinical environment.

The conventional paradigm of CAdE design is that the system operates as an *independent* second-reader to provide radiologists with visual cues and help them reduce mainly perceptual errors associated with the detection task. It is assumed that the CAdE system will be employed by dispassionate radiologists who will be able to differentiate the numerous CAdE marks into true and false positive indications. However, the single most cited complaint from radiologists about CAdE is its high false positive rate. Reportedly, radiologists often accept false CAdE cues while rejecting correct ones. Furthermore, studies suggest that radiologists demonstrate diverse behavioral patterns when considering CAdE cues and these patterns change based on factors that cannot be easily monitored or controlled.^{8–11} Impartial judgment becomes more difficult when making high-risk decisions in the face of false alarms such as invalid CAdE cues. By ignoring the user's human mindset, the conventional paradigm of CAdE design may not be the most effective way to realize the full potential of this otherwise promising technology.

TABLE I. Number of DDSM cases selected for the observer study per breast density rating and pathology.

Breast density	Cancer	Benign	Normal
1: Fatty	1	0	1
2: Fibroglandular	2	3	3
3: Heterogeneous	6	2	1
4: Dense	1	0	0

Several researchers have started to explore alternative paradigms for human-CADe integration. Specifically, there has been a lot of interest in the investigation of interactive CADe systems.^{12–14} Basically, the radiologist reviews the mammogram and selects areas to scrutinize further with the help of CADe. In this interactive capacity, CADe is designed to mainly address the cognitive rather than the perceptual error associated with the detection of breast cancer. Cognitive error is a substantial component of mammographic diagnosis error.^{15–17} Preliminary studies suggest that interactive CADe may help radiologists reduce their cognitive error without burdening them with too many false positive cues or affecting reading time,¹⁸ but further studies are needed on the topic.

In this pilot study, we investigated the potential of a new human-CADe integration paradigm in which the CADe system provides tailored decision support to complement the particular user/radiologist. We call this *context-sensitive* CADe because the CADe opinion is adapted based on the contextual information provided by the radiologist's visual search characteristics and diagnostic decisions for the specific case under review. We performed this study using an in-house CADe system we have been developing for mass detection in screening mammograms. Please note that the original concept of context-sensitive CADe was first introduced by us in Ref. 12, but the present study involves a better performing CADe system, side-by-side comparison of conventional vs context-sensitive CADe, as well as a more detailed data analysis and thorough discussions of our findings.

II. MATERIALS AND METHODS

II.A. The database

We performed an observer study using mammographic cases from the Digital Database of Screening Mammography.¹⁹ They were all cases digitized with a Lumysis scanner (50 μm per pixel, 12 bit). There were ten cancer cases, five benign cases and five normal cases. The benign and normal cases were selected randomly and covered the full range of breast densities. The malignant cases were also selected randomly but we gave higher preference to cases of higher breast density and masses of lower subtlety. Table I shows the distribution of the selected cases based on their pathology status and breast parenchymal density rating (as provided in the DDSM truth file).

The ten cancer cases contained 14 malignant masses (two cancer cases with three masses each and eight cancer cases

TABLE II. Number of malignant and benign masses per subtlety rating. The subtlety rating ranges from 1 (subtle lesion) to 5 (obvious lesion).

Mass subtlety	Malignant masses	Benign masses
1	1	0
2	1	2
3	6	1
4	2	1
5	4	1

with a single mass each). The five benign cases contained five masses (one benign mass per case). Table II shows the distribution of the masses in the data set based on their subtlety rating. The mass subtlety rating is a subjective assessment of the lesion's subtlety provided by the DDSM radiologist who reviewed the case. Mass subtlety was assessed on a five-point scale with a lower rating indicating a less obvious (i.e., more challenging) mass.

II.B. The observer study

The observer study was performed using a single view from each of the 20 mammographic cases. For the cancer and benign cases, only views of the breast that contained the mass lesions were considered in the selection. For the normal cases, a single view per case was selected randomly. The 20 images represented a balanced mix of left and right breasts as well as craniocaudal and mediolateral oblique views. The images were first preprocessed for optimized softcopy display. The quality of the processed images was approved by an experienced breast imager with more than 11 yr of experience in mammographic interpretation. This radiologist did not participate as a reader in the observer study.

Six radiologists (three experienced breast imagers and 34th year residents with at least one rotation in mammography) reviewed the 20 images while wearing an H6 head-mounted eye-head tracker, with 60 Hz sampling speed, and eye-head integration from Applied Science Laboratories (Bedford, MA). Eye tracking data were collected for each observer and case he/she reviewed. The eye-tracking data were the (x,y) coordinates of the observer's eye position sampled every 1/60 s. The criteria used for grouping the raw eye-position data into fixations, clusters, and cumulative clusters are well established.²⁰ First, the x,y coordinates were grouped sequentially according to a running mean distance calculation having an average of 0.50° (1.7 mm) radius threshold²⁰ to determine fixation points. Fixations were then grouped into clusters. A cluster is essentially a circular area that has 2.5° radius centered at the mean x,y location of the group of fixations contributing sequentially to the cluster. This radius corresponds to the useful visual field.^{20,21} Finally, cumulative clusters were calculated by combining fixation clusters generated when the radiologist refixates the same image area at any point in time. A refixation cluster contributes to a cumulative cluster if they overlap by at least 50% and a new centroid is defined by the average of the contributing clusters.

The radiologists read the mammograms as they regularly do, reporting all suspicious locations. The total dwell time per fixation cluster was collected for every image location the radiologists reported. Image locations that attracted prolonged dwelling (>1000 ms) but were not reported were also recorded. We used a 1000 ms threshold for cumulative cluster duration because a previous eye-tracking study²² showed that using this threshold identified approximately 80% of all true positive (TP) lesion reports and 65% of all false negatives.

II.C. Our CADe system

The same 20 mammograms were analyzed using our in-house, information-theoretic CADe (IT-CADe) system. Our system is summarized below but more details can be found in our published studies (e.g., Refs. 23–27). IT-CADe uses a knowledge database of proven mass and normal templates. The system is designed to operate mainly in an interactive manner, i.e., to interrogate any image location as directed by the user/radiologist. However, the system can be used in a conventional manner as well. By scrolling IT-CADe across the breast region, IT-CADe interrogates every image location and flags those that are highly suspicious to contain true masses.

IT-CADe was applied initially on the 20 test cases in a scrolling mode to measure its standalone performance. Here is a brief description of the image processing and decision steps involved. First, the breast region is segmented. At each location within the segmented breast region, a fixed size ROI is extracted around the location to serve as query to the system. The system compares the similarity between the query ROI and the knowledge templates. The similarity assessment is measured using mutual information (MI).²⁸ The MI is calculated directly from the ROI histograms using histogram binning. Then, the difference between the query's average MI similarity to the knowledge mass templates vs the knowledge normal templates is used as a decision variable to determine whether a mass is present in the query ROI or not. Pixel groupings that exceed a predetermined decision threshold are considered highly suspicious for depicting masses.

II.D. Context-sensitive IT-CADe analysis

Upon completion of the radiologists' viewing sessions the 20 test cases were reanalyzed by IT-CADe guided by the user's eye-tracking and reporting for each case. Depending on the specific user's level of attention at a particular image region, IT-CADe operated with different decision thresholds. In other words, the level of scrutiny applied by IT-CADe in a particular image region was based on IT-CADe's expected supporting role for the user and case at hand.

Specifically, for image locations that were reported by the radiologists (regardless of their dwelling time), IT-CADe offered a second opinion using a lax decision threshold, corresponding to 95% per-image malignant mass sensitivity. Enforcing the CADe system to operate conservatively for image locations that already generate a positive response from the radiologists makes intuitively sense. It ensures that

the radiologists' sensitivity would not be compromised if they choose to trust the system. Although the FDA-approved role of CADe systems discourages radiologists to disregard suspicious image findings they already perceived in the absence of related CADe marks, it is still important to assess whether context-sensitive CADe could help radiologist reduce their false positive (FP) *decision errors* while not jeopardizing their sensitivity and trust in the system.

In contrast, for image regions that attracted long dwells but were not reported by the radiologists, IT-CADe operated with a moderate decision threshold (i.e., 85% per-image malignant mass sensitivity). Providing decision support with a more aggressive decision threshold is well justified for these image regions. They were deemed normal by the radiologists after thorough inspection. Therefore, the desired role of CADe for these mammographic regions would be to cue any malignant masses the radiologists missed [i.e., false negative (FN) *decision errors*] while not overcalling too many of their true negative (TN) decisions.

Finally, for image regions that attracted short or no dwell and went unreported by the radiologists (either due to lack of search or due to lack of conspicuous findings), IT-CADe operated with a strict decision threshold to ensure that it generates an acceptably low number of false positive cues [i.e., 0.25 false positives per image (FPsI)]. For these image regions, the desired benefit would be to reduce the radiologists' *FN search and recognition errors* while producing as few as possible FP cues. Maintaining a low FPsI is important to reduce the chance that radiologists will dismiss correct CADe cues in those regions.

Based on the above operating criteria, if the centroids of the radiologists' cumulative fixation clusters were closer than 2.5 mm from an IT-CADe mark (generated while operating with the appropriate decision threshold), they were considered as locations for which IT-CADe gave a positive response.

II.E. Statistical analysis

In contrast with conventional CADe systems, context-sensitive IT-CADe is user-dependent. Therefore, its standalone performance varies across radiologists. To test the significance of the difference in IT-CADe's standalone sensitivity and FPsI between conventional and context-sensitive operating modes, we used the student's paired *t*-test to take into account reader variance in context-sensitive mode. In addition, to investigate whether the change in sensitivity was statistically significant for a given radiologist, we used the McNemar test for clustered data,²⁹ which takes into account that the mammograms contained variable number of malignant masses. To investigate whether the FPsI change was statistically significant for a given radiologist, we used the Wilcoxon signed rank test.

We also applied multireader multicase (MRMC) methodology to compare the radiologists' average performance using conventional vs context-sensitive IT-CADe under a hypothetical scenario of clinical use. For this analysis, we assumed that both the conventional and the context-sensitive

TABLE III. Radiologists' diagnostic performance without CADe.

Radiologist	Sensitivity (malignant masses)	Sensitivity (benign masses)	Recall rate (or FPsI)
1	85.7% (12/14)	100% (5/5)	0.1 (2/20)
2	78.6% (11/14)	100% (5/5)	0.1 (2/20)
3	78.6% (11/14)	80% (4/5)	0 (0/20)
4	85.7% (12/14)	100% (5/5)	0.1 (2/20)
5	100% (14/14)	100% (5/5)	0.05 (1/20)
6	85.7% (12/14)	100% (5/5)	0.05 (1/20)

IT-CADe systems operate at the same specificity. This way, we evaluated the expected change in the radiologists' sensitivity. The main advantage of MRMC analysis is that it takes into account the correlation between readings of the same case by different radiologists, as well as the correlation between readings of different cases by the same radiologist. There are several MRMC techniques proposed for analysis of ROC-type rating data.³⁰ We applied the one-shot estimation technique³¹ which has been recently extended for use with binary data such as in our case.³² To account for some of the shortcomings of this technique and confirm consistency of the observed trends, we also investigated generalized linear mixed models for the same task.³³ SAS/STAT software, Version 9.2 for Windows 368 (Cary, NC) was used for this analysis.

III. RESULTS

III.A. Radiologists' performance

Table III shows the diagnostic performance of the individual radiologists. The radiologists' average sensitivity for malignant masses was 85.7% (95% confidence interval: 79.5% and 91.9%) and their average sensitivity for benign masses was 96.7% (95% confidence interval: 90.1% and 100%). Their average recall rate was 0.067 FPsI (95% confidence interval: 0.035 and 0.099). The recall rate is the ratio of the total number of false positive image locations each radiologist reported divided by the total number of images (i.e., 20). Each radiologist reported 0–2 false positive decisions across the whole test set.

The radiologists' eye-position recordings on the 20 test cases were combined in a single data set and they were grouped according to the radiologist's dwell times and decisions for each recording. Based on dwell time, the radiologists' cumulative fixation clusters were grouped as:

(i) LONG dwells (i.e., image locations that attracted prolonged viewing, >1000 ms), (ii) SHORT dwells (i.e., image locations that attracted some viewing but <1000 ms), and (iii) NO dwells (i.e., image locations that attracted no fixations). The radiologists' decisions were grouped as: (i) REPORTED locations, i.e., TP and FP decisions made by a radiologist, and (ii) UNREPORTED locations, i.e., TN and FN decisions made by a radiologist. For this analysis, a correctly reported mass was considered a true positive decision regardless if it was a malignant or a benign mass. The same rule was applied for false negative decisions. FN errors included all masses missed by the radiologists regardless of their malignancy status. Table IV shows the detailed statistics for each group. Note all lesion-free image locations for which the participating observers did not report any findings and did not dwell were considered TN decisions of short or no dwelling.

Grouping the radiologists' decisions based on their dwell times helps categorize the various types of diagnostic decisions. Each subgroup essentially captures the diagnostic errors (i.e., perceptual and cognitive) made by the radiologists. For example, reported locations with long dwells consist of TP decisions and *FP decision errors* since this subgroup contains image locations that the radiologists decided to report after prolonged viewing and careful consideration. Reported locations with short dwells consist of TP decisions and *FP recognition errors* since they contain image locations that the radiologists recognized as highly suspicious even after limited viewing. Unreported locations with long dwells include TN decisions as well as *FN decision errors* because the radiologists decided to disregard some true masses after prolonged viewing and scrutiny. Finally, unreported locations with short or no dwells combine all *FN errors of search (no dwell) and recognition (short dwell)*, true masses that went unreported because the radiologists failed to search in the image areas where these masses were or failed to recognize their signs after limited viewing. The latter group also contains all correctly unreported normal (lesion-free) locations.

III.B. Conventional IT-CADe performance

Table V reports the IT-CADe system's standalone performance on the 20 DDSM images when operating in conventional mode using a fixed decision threshold. Corresponding sensitivity and specificity are shown for three progressively

TABLE IV. Diagnostic decisions made by the radiologists grouped according to the dwell time of the radiologist who made the corresponding decision.

		Diagnostic decision	
		Yes	No
Dwelling time	Long	67 TPs (49 malignant+18 benign) and 6 FPs	84 TNs and 6 FNs (5 malignant+1 benign)
	Short	33 TPs (23 malignant+10 benign) and 2 FPs	5 FNs (5 malignant+0 benign)
	No	0 TPs and 0 FPs	3 FNs (2 malignant+1 benign)

TABLE V. Standalone IT-CADe performance on the 20 test cases for three different operating points.

Operating point	Sensitivity (malignant masses)	Specificity (FPsI)
Lax (very conservative)	100% (14/14)	6.95
Moderate	85.7% (12/14)	3.15
Strict (highly aggressive)	21.4% (3/14)	0.30

stricter decision thresholds (i.e., lax, moderate, and strict). We call an operating point “lax” if the IT-CADe system applies a lenient decision threshold which tends to overcall image locations (i.e., thus achieving high sensitivity at the expense of low specificity). The opposite is true for a strict operating point. These thresholds were determined during the system’s development stage using DDSM/Lumisys cases independent of the observer study. Specifically, the system was tested on 60 cases (40 cancer and 20 normal) reserved exclusively for IT-CADe optimization and internal validation. Figure 1 shows the corresponding free response operating characteristic (FROC) curve when the final system is tested on those 60 test cases. The figure includes both the *per-image* and the *per-case* performance. The three operating points utilized with context-sensitive IT-CADe were selected from the per-image FROC curve and they are shown in Fig. 1.

For the lax and moderate decision thresholds, IT-CADe demonstrated 95% and 85% per-image malignant mass sensitivity, respectively (or 98% and 95% per-case malignant mass sensitivity), during the development stage. The strict decision threshold was selected to correspond to a very low FPsI rate, similar to that typically reported for commercial CADe products (i.e., 0.25 FPsI). However, at such low FPsI, the in-house IT-CADe system achieved 40% per-image and 65% per-case malignant mass sensitivity during the development stage, which is much lower than that reported by commercial CADe products.

Operating conservatively with the lax decision threshold, IT-CADe detected 100% (14/14) of the malignant masses in

FIG. 1. FROC standalone performance of the IT-CADe system used to establish the three operating points (*lax*, “*moderate*,” and “*strict*”) for context-sensitive application of the system.

TABLE VI. Detection performance of conventional vs context-sensitive IT-CADe.

Radiologist	Conventional IT-CADe		Context-sensitive IT-CADe	
	Sensitivity (%)	FPsI	Sensitivity (%)	FPsI
1			100	0.35
2			85.7	0.40
3	85.7	3.15	92.9	0.35
4			92.9	0.40
5			100	0.40
6			92.9	0.40

the 20 test cases. At the same decision threshold, IT-CADe generated an average of 6.95 FPsI marks (determined across all 20 images). As expected, the IT-CADe specificity progressively improved at the expense of sensitivity as the decision threshold became more aggressive. For the most part, the IT-CADe performance was comparable in the development and test sets in terms of both sensitivity and FPsI for all three operating thresholds, with few exceptions. At the moderate operating point, the average number of false positives per image was lower in the development set than in the test set (2.33 vs 3.15 FPsI). At the strict operating point, the IT-CADe sensitivity was notably lower in the test set than what was expected based on the development set (21.4% vs 40%).

III.C. Context-sensitive IT-CADe performance

Table VI shows the IT-CADe performance when the system operates in the proposed context-sensitive mode (using location-dependent decision thresholds in response to the radiologist’s dwelling and reporting pattern for the specific location). Note that since context-sensitive IT-CADe is driven by the user/radiologist, it provides personalized decision support. Therefore the performance of context-sensitive IT-CADe varies across the six radiologists. In contrast, the performance of conventional IT-CADe is the same across all radiologists (“one-size-fits-all” decision support). Table VI provides a side-by-side comparison of the two IT-CADe operating modes.

The table illustrates how the same CADe system can have different performance if it offers second opinions using user-adaptive thresholds compared to fixed thresholds. For five out of six radiologists, context-sensitive IT-CADe provided a second opinion with higher sensitivity than conventional IT-CADe. For one radiologist, both modes show the same sensitivity. The average sensitivity achieved by context-sensitive IT-CADe across all six radiologists was 94% (95% confidence interval: 88.4% and 99.7%). Compared to 85.7% sensitivity achieved by conventional IT-CADe (same across all radiologists), the improvement was statistically significant (student’s paired *t*-test, two-tailed *p*-value=0.0125). However, McNemar’s test did not show statistically significant difference in sensitivity using context-sensitive IT-CADe vs conventional IT-CADe individually for any of the five radi-

ologists for whom an improvement was observed (two-tailed p -values > 0.489 for all five radiologists). The lack of statistical significance can be attributed to the small number of cases.

In contrast, context-sensitive IT-CADE produced a dramatically lower number of false positive cues per image for all radiologists than conventional IT-CADE. The average FPsI rate achieved by context-sensitive IT-CADE was 0.38 FPsI (95% confidence interval: 0.34 and 0.43). This was statistically significantly lower than the 3.15 FPsI rate shown by conventional IT-CADE across all radiologists (student's paired t -test, two-tailed p -value = 0.0001). The Wilcoxon signed-ranks test showed that the reduction in FPsI was also statistically significant for each radiologist individually (two-tailed p -values < 0.0005 for all six radiologists).

III.D. Evaluation of context-sensitive IT-CADE on radiologists' data

In total, there were 100 TP (72 for malignant masses + 28 for benign masses), 14 FN (12 for malignant masses + 2 for benign masses), and eight FP decisions made by the radiologists. Context-sensitive IT-CADE confirmed 100% (72/72) of the TP malignant mass decisions and correctly disagreed with 58.3% (7/12) of their FN decisions related to malignant masses, cuing those missed mass locations. In addition, IT-CADE disagreed with 37.5% (3/8) of the radiologists' FP decisions.

Here is a more detailed description of the results, based on grouped analysis.

- (1) Reported locations with long dwells: There were 68 TP and 6 FP such recordings (see Table III). Of the 68 TP decisions, 50 corresponded to malignant masses and 18 to benign masses. IT-CADE confirmed correctly 100% (49/49) of the radiologists' TP decisions which corresponded to malignant masses and 55.6% (10/18) of the radiologists' TP decisions which corresponded to benign masses. There were also six FP decisions in total made by the radiologists (one by radiologists 1, 2, 5, and 6 and two made by radiologist 4). These six FP decisions were completely different and thus there was not a single image location that triggered a false positive response from multiple observers. IT-CADE correctly disagreed with three of them (50%), calling them normal. This result suggests that IT-CADE operating conservatively on radiologists' reported locations of long dwelling could still improve the radiologists' specificity without compromising their sensitivity.
- (2) Reported locations with short dwells: This subgroup represents positive decisions made by radiologists after brief visual consideration, namely, true masses that were likely easy to detect and challenging normal parenchymal areas with masslike characteristics. There were 33 TP recordings (23 malignant and 10 benign) and 2 FP recordings. The FP decisions were made by radiologists 1 and 2 in different mammograms. Similar to subgroup 1, the only possible benefit that a CADE system can offer with subgroup 2 is the reduction of possible FP

errors. However, radiologists may be even more reluctant to change their mind under the influence of CADE for this subgroup since it contains findings that the radiologists perceived, diagnosed, and reported fast. IT-CADE operating conservatively with the lax decision threshold confirmed 100% (23/23) and 60% (6/10) of the malignant and benign TP decisions, respectively. None of the two FP decisions was eliminated; IT-CADE agreed with the radiologists that both image locations depicted indeed a true mass.

- (3) Unreported locations with long dwells: When tested on areas with long dwells that were not reported by the radiologists (84 TNs and 6 FNs) using the moderate decision threshold, IT-CADE correctly disagreed with five FN decisions (all malignant). Note that the six FN decisions were made by three different radiologists (1, 3, and 4) and corresponded to five different masses, four malignant and one benign. However, IT-CADE also wrongly overcalled 6/84 (7.1%) TN decisions. These six additional FP decisions made by context-sensitive IT-CADE were across all 6 radiologists $\times$ 20 images (=120 readings), thus they reflect an increase of 0.05 (6/120) FPsI. This finding suggests that IT-CADE was still able to identify 100% (5/5) of FN decision errors made by the radiologists (with respect to malignant masses) with a minimal overcalling of their TN decisions.
- (4) Unreported locations with short or no dwells: There were eight such FN decisions (seven malignant + one benign) made by the radiologists. The seven malignant FN decisions corresponded to five different masses. Operating with the strict threshold IT-CADE correctly disputed 2/7 (28.6%) radiologists' FN search and recognition errors related to malignant masses while generating only five additional FP cues (0.25 FPsI) in those regions of short or no dwelling.

III.E. Impact on individual radiologist's performance

Next, we determined the expected change in the radiologists' performance under the hypothetical scenario that the radiologists would trust completely the opinions offered by the IT-CADE system. Please note that for conventional CADE support, it is assumed that each radiologist would not change his/her mind for locations they chose to report, even if conventional IT-CADE did not mark the same locations. This is consistent with the FDA-approved use of CADE technology. Since the amount of support offered by context-sensitive IT-CADE varies depending on the individual radiologist, we set the conventional IT-CADE system to operate at the same specificity level as that demonstrated on average by context-sensitive IT-CADE (i.e., 0.4 FPsI as shown in Table VI). Fixing the specificity allowed us to compare conventional vs context-sensitive CADE with respect to their impact on the radiologists' sensitivity for malignant masses. Table VII shows the expected effect for each radiologist.

In this hypothetical scenario, three out of six radiologists would experience an increase in their sensitivity with con-

TABLE VII. Change on individual radiologist's sensitivity with conventional vs context-sensitive CADe.

Radiologist	Conventional IT-CADe	Context-sensitive IT-CADe
1	85.7% → 92.9% (+6.6%)	85.7% → 100% (+14.3%)
2	78.6% → 85.7% (+7.1%)	78.6% → 85.7% (+7.1%)
3	78.6% → 78.6% (0%)	78.6% → 92.9% (+13.7%)
4	85.7% → 85.7% (0%)	85.7% → 92.9% (+6.6%)
5	100% → 100% (0%)	100% → 100% (0%)
6	85.7% → 92.9% (+6.6%)	85.7% → 92.9% (+6.6%)

ventional IT-CADe support. The radiologists' average sensitivity would increase from $85.7\% \pm 7.8\%$ to $89.3\% \pm 7.5\%$. With context-sensitive IT-CADe support, five out of six radiologists would experience an increase in their sensitivity. Their average sensitivity would increase to $94.1\% \pm 5.4\%$.

MRMC analysis of the differences in average sensitivity across the three modes showed that the estimated improvement in sensitivity with conventional IT-CADe was $3.6\% \pm 2.3\%$ (Z-test, two-tailed p-value=0.13), while the improvement with context-sensitive IT-CADe was $8.3\% \pm 2.7\%$ (Z-test, two-tailed p-value=0.002). The relative improvement with context-sensitive IT-CADe vs conventional IT-CADe was $4.8\% \pm 2.8\%$ (Z-test, two-tailed p-value=0.09) (Table VIII).

The above MRMC analysis using the one-shot estimation technique ignores the clustered nature of the binary data treating each malignant lesion as an independent case. Furthermore, the Z-test is not the most appropriate choice given the small sample size. To account for the shortcomings of the previous model, we also fitted a logistic regression model for the probability of a positive reading of a case as a function of the reading mode with random effects for radiologists and cases. At the 0.05 confidence level, there was no statistically significant difference between "no CADe" and conventional IT-CADe modes [two-tailed p-value=0.46, $F(1, 231)=0.55$] or between conventional and context-sensitive IT-CADe [two-tailed p-value=0.25, $F(1, 231)=1.34$]. Only the difference between no CADe and two-tailed p-value=0.07, $F(1, 231)=3.38$). Therefore, given the small data set, both analyses demonstrate the same consistent trends; context-sensitive IT-CADe appears to be an effective mode for providing decision support and it appears to be more promising than the conventional mode.

Last but not least, we studied how context-sensitive IT-CADe may affect the performance gap between experienced

and less-experienced radiologists compared to conventional IT-CADe. Table VIII shows the student's two-tailed unpaired *t*-test p-value of the difference between the average sensitivity of the less-experienced radiologists¹⁻³ and that of the experienced radiologists⁴⁻⁶ under each reading mode: Without CADe support, with conventional IT-CADe support, and with context-sensitive CADe support. Although the experienced radiologists have on average higher sensitivity than the less-experienced radiologists under all reading modes, the difference between the two groups is notably less pronounced with context-sensitive IT-CADe support, as suggested by the higher p-value.

IV. DISCUSSION

Over the past decade there has been great improvement in the performance of CADe systems in mammography, both with respect to sensitivity and specificity. Yet, many radiologists still consider CADe's mass sensitivity levels and false positive rates far from adequate. Consequently, researchers continue their efforts to improve the standalone performance of CADe systems (e.g., Refs. 34–37). The task remains challenging and further performance gains are expected to come in small increments.

In this study, we asked the question whether additional benefits could be derived by improving not only CADe's standalone performance but by also rethinking the existing model of human-CADe interaction. CADe systems are developed for the general population of prospective patients. The process typically involves global optimization of the CADe decision thresholds to balance clinically acceptable sensitivity and specificity levels across the target population. However, this conventional human-CADe paradigm follows the one-size-fits-all thinking.

In this study, we explored the feasibility of a personalized human-CADe integration paradigm where the CADe system provides decision support tailored to each user's perceptual and cognitive patterns for the specific case under review. Specifically, we aimed to determine whether a CADe system with standalone performance that is far from ideal could be more effective when used in the personalized mode than in the conventional mode.

Guided by an eye-tracking device and applied in a targeted manner, our in-house CADe system offered a second opinion using variable decision thresholds. Depending on the user's visual search characteristics (i.e., dwell times) and decisions for each case, the CADe system anticipated the type of error (e.g., search, recognition, decision) a radiologist could possibly make and offered a second opinion using the appropriate decision threshold. Essentially, the decision threshold controlled how vigilant the system was at different image locations.

The motivation behind our approach stemmed from how we anticipate the radiologists to respond under the influence of CADe cues. Specifically, for image locations that a radiologist reports as suspicious after prolonged viewing, our CADe system was asked to be conservative to ensure that the radiologist's sensitivity (as well as trust for the system)

TABLE VIII. Average sensitivity of the radiologists grouped based on their experience level under the three reading scenarios.

Reading mode	Less experienced radiologists (%)	Experienced radiologists (%)	p-value
Without CADe	81	90.5	0.15
With conventional IT-CADe	85.7	92.9	0.29
With context-sensitive IT-CADe	92.9	95.3	0.64

would not be compromised. Even with a conservative decision threshold that encourages overcalling, the system still identified 50% of the radiologists' FP decision errors. In contrast, for image locations that attracted prolonged viewing but finally went unreported by the radiologists, the CADE operating with a moderate decision threshold was still able to identify 100% of the FN errors with marginal impact on specificity. In this scenario, a radiologist may be willing to reconsider a CADE cue more seriously if he/she is alerted that although the system operates with lower vigilance, it still considers the location highly suspicious and worthy of further investigation. Finally, for all other image regions which did not attract much visual attention and no findings were reported by the radiologist, our IT-CADE with a very aggressive decision threshold was still capable of flagging 28.7% of missed malignant masses while generating fewer than 0.25 FPsI in those areas. In this scenario, a radiologist may be more willing to reconsider a CADE cue generated by the system when it operates with a high negative predictive value.

There were several limitations in our study. First, it was based on a small number of cases. Many study findings that did not reach statistical significance could be attributed to the small data set. Second, the observer study was based on single image viewing which is not consistent with clinical practice. However, our main goal was to explore the potential of context-sensitive CADE relative to conventional CADE. Third, a fixed threshold for total dwelling time (1000 ms) was used across all radiologists based on recommendations from previously published eye-tracking studies. Finally, our analysis did not include other parameters of the radiologists' eye-tracking patterns (e.g., number of returns, time from beginning of reading until the radiologist fixates on a reported location for the first time, etc.). Including these parameters and performing data mining to derive better user-adaptive decision rules could potentially lead to a more sophisticated and effective implementation of the context-sensitive CADE paradigm. However, given the small scale and exploratory nature of our study, we believe simpler analysis is preferable initially.

Due to the above shortcomings, our study findings should be interpreted with the necessary caution. The presence of the "laboratory effect" is well documented in the CAD literature; findings from small-scale, retrospective, laboratory CAD studies are not always consistent with findings from large-scale prospective studies in a real clinical setting.³⁸ The risk of the laboratory effect is particularly high with our own pilot study given the small data set and the single-view reading conditions. Furthermore, the potential clinical contribution was assessed based on a hypothetical scenario, which is not ideal. Therefore, a larger scale study that simulates the clinical conditions is necessary to validate our findings and fully evaluate the clinical potential of the context-sensitive CADE paradigm. Nevertheless, a laboratory pilot study is the necessary first step when studying a new paradigm to establish proof-of-concept and identify important trends. These trends can help refine the paradigm and guide the design of the larger scale studies. Our pilot study identified three inter-

esting trends with respect to the standalone and clinical performance of our in-house CADE system. First, operating in context-sensitive mode, our system provided decision support with consistently higher sensitivity and significantly higher specificity than when operating in conventional mode. Second, context-sensitive CADE could help radiologists identify and reduce possible perceptual and cognitive diagnostic errors they make more effectively than conventional CADE. Third, context-sensitive CADE appears to be more effective in bridging the performance gap among radiologists of various experience levels. Therefore, the context-sensitive CADE paradigm has the potential to maximize the benefit of existing CADE technology by providing personalized decision support that meets the individual needs of each user. We believe that these findings encourage further exploration of other human-CADE integration paradigms than the one currently in practice. Taking into account behavioral informatics issues is a promising path to better human-CADE synergy and thus more effective clinical integration of CADE technology.

ACKNOWLEDGMENTS

This work has been supported in part by NIH/NCI Grant No. R56 CA101911. The authors thank Dr. Brandon Gallas and Dr. Xin He from the FDA for the one-shot estimate MRMC analysis of the data. The authors also thank Dr. Yuliya Lokhnygina and Dr. Andrzej Kosinski from the Department of Biostatistics and Bioinformatics at Duke University Medical Center for the generalized mixed model analysis of the same data.

^{a)} Author to whom correspondence should be addressed. Electronic addresses: georgia.tourassi@duke.edu and gt@deckard.duhs.duke.edu; Telephone: (919) 684-1447; Fax: (919) 684-1491.

¹ R. L. Birdwell, "The preponderance of evidence supports computer-aided detection for screening mammography," *Radiology* **253**, 9–16 (2009).

² L. E. Philpotts, "Can computer-aided detection be detrimental to mammographic interpretation?," *Radiology* **253**, 17–22 (2009).

³ M. Gromet, "Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms," *AJR, Am. J. Roentgenol.* **190**, 854–859 (2008).

⁴ F. J. Gilbert *et al.*, "Single reading with computer-aided detection for screening mammography," *N. Engl. J. Med.* **359**, 1675–1684 (2008).

⁵ D. Gur *et al.*, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *J. Natl. Cancer Inst.* **96**, 185–190 (2004).

⁶ J. J. Fenton *et al.*, "Influence of computer-aided detection on performance of screening mammography," *N. Engl. J. Med.* **356**, 1399–1409 (2007).

⁷ R. M. Nishikawa and L. P. Pesce, "Computer-aided detection evaluation methods are not created equal," *Radiology* **251**, 634–636 (2009).

⁸ L. Strigini, A. A. Povyakalo, and E. Alberdi, "Human-machine diversity in the use of computerized advisory systems: A case study," in *Proceedings of DSN, International Conference on Dependable Systems and Networks*, 2003, pp. 249–258.

⁹ E. Alberdi *et al.*, "Use of computer-aided detection (CAD) tools in screening mammography: A multidisciplinary approach," *Br. J. Radiol.* **78**, S31–S40 (2005).

¹⁰ E. Alberdi, P. Ayton, A. A. Povyakalo, and L. Strigini, "CAD in mammography: Lesion-level vs. case-level analysis of the effects of prompts on human decisions," *Journal of Computer Assisted Radiology and Surgery* **3**, 115–122 (2008).

¹¹ E. Alberdi, L. Strigini, P. Ayton, and A. A. Povyakalo, "Why are people's decisions sometimes worse with computer support?," *Lect. Notes Comput. Sci.* **5775**, 18–31 (2009).

- ¹²G. D. Tourassi, M. A. Mazurowski, and E. A. Krupinski, "Perception-driven IT-CADe analysis for the detection of masses in screening mammography: Initial investigation," *Proc. SPIE* **7624**, 762406-1–762406-6 (2010).
- ¹³M. Samulski, A. Hupse, C. Boetes, G. den Heeten, and N. Karssemeijer, "Analysis of probed regions in an interactive CAD system for the detection of masses in mammograms," *Proc. SPIE* **2763**, 726314-1–726314-9 (2009).
- ¹⁴B. Zheng, "Computer-aided diagnosis in mammography using content-based image retrieval approaches: Current status and future perspectives," *Algorithms* **2**, 828–849 (2009).
- ¹⁵C. F. Nodine, H. L. Kundell, S. C. Lauver, and L. C. Toto, "Nature of expertise in searching mammograms for breast masses," *Acad. Radiol.* **3**, 1000–1006 (1996).
- ¹⁶C. F. Nodine *et al.*, "Blinded review of retrospectively visible unreported breast cancers: An eye-position analysis," *Radiology* **221**, 122–129 (2001).
- ¹⁷E. A. Krupinski, C. F. Nodine, and H. L. Kundell, "Enhancing recognition of lesions in radiographic images using perceptual feedback," *Opt. Eng. (Bellingham)* **37**, 813–818 (1998).
- ¹⁸N. Karssemeijer, M. Samulski, M. Kallenberg, A. Hupse, C. Boetes, and G. Den Heeten, "Effectiveness of an interactive CAD system for mammographic mass detection," presented at the Proceedings of the 94th Scientific Assembly and Annual Meeting, McCormick Place, Chicago, IL, 2008.
- ¹⁹M. Heath *et al.*, "Current status of the digital database for screening mammography," in *Digital Mammography* (Kluwer Academic, Dordrecht, 1998).
- ²⁰H. L. Kundell and P. S. LaFollette, "Visual search patterns and experience with radiological images," *Radiology* **103**, 523–528 (1972).
- ²¹N. H. Mackworth, "Stimulus density limits the useful field of view," in *Eye Movements and Psychological Processes*, edited by R. A. Monty and J. W. Senders (Erlbaum, Hillsdale, 1976), pp. 307–322.
- ²²E. A. Krupinski and R. M. Nishikawa, "Comparison of eye position versus computer identified microcalcification clusters on mammograms," *Med. Phys.* **24**, 17–23 (1997).
- ²³G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, and C. E. Floyd, Jr., "A computer-assisted detection of mammographic masses: A template matching scheme based on mutual information," *Med. Phys.* **30**, 2123–2139 (2003).
- ²⁴G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, Jr., "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Med. Phys.* **34**, 140–150 (2007).
- ²⁵G. D. Tourassi, B. Harrawood, S. Singh, and J. Y. Lo, "Information-theoretic CAD system in mammography: Entropy-based indexing for computational efficiency and robust performance," *Med. Phys.* **34**, 3193–3204 (2007).
- ²⁶M. A. Mazurowski, J. Zurada, and G. D. Tourassi, "Selection of examples in case-based CAD systems," *Phys. Med. Biol.* **53**, 6079–6096 (2008).
- ²⁷G. D. Tourassi, R. Ike III, S. Singh, and B. Harrawood, "Evaluating the effect of image preprocessing on an information-theoretic CAD system in mammography," *Acad. Radiol.* **15**(5), 626–634 (2008).
- ²⁸T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
- ²⁹N. A. Obuchowski, "On comparison of correlation of correlated proportion for clustered data," *Stat. Med.* **17**, 1495–1507 (1998).
- ³⁰N. A. Obuchowski, "New methodological tools for multiple-reader ROC studies," *Radiology* **243**, 10–12 (2007).
- ³¹B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**, 353–362 (2006).
- ³²B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multi-reader multi-case variance analysis for binary data," *J. Opt. Soc. Am. A* **24**(12), B70–B80 (2007).
- ³³C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models* (Wiley, New York, 2001).
- ³⁴B. Zheng *et al.*, "Multiview-based computer-aided detection scheme for breast masses," *Med. Phys.* **33**, 3135–3143 (2006).
- ³⁵S. van Engeland and N. Karssemeijer, "Combining two mammographic projections in a computer aided mass detection method," *Med. Phys.* **34**, 898–905 (2007).
- ³⁶Y. Wu, J. Wei, L. M. Hadjiiski, B. Sahiner, C. Zhou, J. Ge, J. Shi, Y. Zhang, and H.-P. Chan, "Bilateral analysis based false-positive reduction for computer-aided mass detection," *Med. Phys.* **34**, 3334–3344 (2007).
- ³⁷J. Wei, H.-P. Chan, B. Sahiner, C. Zhou, L. M. Hadjiiski, M. A. Roubidoux, and M. A. Helvie, "Computer-aided detection of breast masses on mammograms: Dual system approach with two-view analysis," *Med. Phys.* **36**, 4451–4460 (2009).
- ³⁸D. Gur *et al.*, "The 'laboratory' effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations," *Radiology* **249**, 47–53 (2008).