J Digit Imaging (2011) 24:772–777
DOI 10.1007/s10278-011-9389-7

False Positive Marks on Unsuspicious Screening Mammography with Computer-Aided Detection

Mary C. Mahoney · Karthikeyan Meganathan

Published online: 6 May 2011

Abstract The contribution of computer-aided detection (CAD) systems as an interpretive aid in screening mammography can be hampered by a high rate of false positive detections. Specificity, false positive rate, and ease of dismissing false positive marks from two CAD systems are retrospectively evaluated. One hundred screening mammographic studies with a BI-RADS assessment code of 1 or 2 and at least 2-year normal mammographic follow-up were retrospectively reviewed using two CAD systems. Breast density, CAD marks, and radiologist's ease of dismissing false positive marks were recorded. Specificities from the two CAD versions considering all marks were 23% and 15% (p value=0.07); mass marks, 35% and 17% (p value<0.01); and calcification marks 62% and 75% (p value=0.01). The two CAD versions did not differ regarding mean and median marks per case for all marks (2.3, 2.0 and 2.3, 2.0, p value=0.65) or mass marks (1.6, 1.0 and 1.8, 2.0, p value=0.15), but differed for calcification marks (0.8, 0 and 0.5, 0, p value<0.01). Slightly higher specificity and fewer marks per case observed in dense breasts did not reach statistical significance. The reviewing radiologist classified most marks from both CAD systems (84% and 88%) as very easy/easy to dismiss. The two CAD versions had small differences in specificity and false positive marks. Differences, although not statistically significant, in specificities and false positive rates between dense and non-dense breasts warrant further research. Most false positive marks are easily dismissed and should not affect clinical performance.

Keywords Screening mammograms · Computer-aided detection (CAD) · False positive marks · Specificity · Breast density

M. C. Mahoney (✉)
Barrett Cancer Center, University of Cincinnati Medical Center,
234 Goodman Street, Mail Location 772,
Cincinnati, OH 45267, USA
e-mail: mahonemc@healthall.com

K. Meganathan
Department of Public Health Sciences, University of Cincinnati,
PO Box 670840, Cincinnati, OH 45267-0840, USA
e-mail: meganak@ucmail.uc.edu

Background

Computer-aided detection (CAD) systems are being widely used as an interpretive aid in screening mammography. Approval from the U.S. Food and Drug Administration and insurance companies' willingness to reimburse CAD services highlight the acceptance of CAD as an interpretive aid. Several retrospective and prospective studies that have evaluated the diagnostic accuracy and contribution of CAD as an interpretive aid resulted in mixed conclusions [1, 2]. Most studies which indicate a 5–20% increase in cancer detection rates often reported a comparable increase in recall rates [3–19]. Some studies have explored radiologists' experience in evaluating mammograms and training with CAD as factors affecting the efficacy of CAD as an interpretive aid. Slightly increased benefits from CAD have been observed with experienced radiologists compared to novice radiologists and residents [20]. Luo et al. [21] reported improvements in a radiologist's performance after undergoing a 4-week training involving a hypermedia instructional program in CAD-aided mammography interpretation.

Early CAD studies have reported a mean of 3–5 false positive CAD marks per four-image mammogram [3–5, 22–25], defining a false positive as any CAD mark present

in a non-cancerous area. More recent CAD studies, some of which involved digital mammograms, have reported a mean of two false positive CAD marks per four-image mammogram [26–30]. In addition to retrospectively assessing the specificity and rate of false positive marks in two versions of CAD in screening mammography among unsuspicious cases in a university clinic setting, the purpose of this article is to assess the reviewing radiologist's ease of dismissing CAD false positives. The effects of breast density on specificity and rate of false positive marks are also assessed.

Materials and Methods

The computer records of a dedicated breast center were retrospectively reviewed for patients who underwent a screening mammogram in the year 2000 with a final assessment of Breast Imaging Reporting and Data System (BI-RADS) 1 or 2 and had subsequent normal follow-up studies (BI-RADS 1 or 2) in years 2001 and 2002 at the same center. The screening mammographic studies of 100 randomly chosen patients were digitized, using a spot size setting of 43.5 μm and optical density linear range of 0.03 to 3.80 OD, and evaluated with a CAD system (SecondLook v5.0, CADx Systems, Inc., Beavercreek, Ohio). This CAD system, with only one operating point, was set to a case-based sensitivity of 89.3% at the rate of 2.0 false positive marks per four-image case. A dedicated breast radiologist, with 10 years of breast imaging experience, reviewed all mammographic films.

Age of the patient, number of films per study, breast density, number of CAD marks (masses and calcifications), and reviewing radiologist's ease of dismissing the CAD marks were recorded. Breast density was assessed subjectively by the mammographer according to BI-RADS as predominately fatty, scattered fibroglandular densities, heterogeneously dense, and extremely dense. Patients with entirely fatty or scattered fibroglandular breasts are considered to have "non-dense" breasts (n=61 patients), and patients with heterogeneously dense or extremely dense breasts are considered to have "dense" breasts (n=39 patients). Ease of dismissing the CAD marks was subjectively rated by the reviewing radiologist, based on the time required to dismiss the mark, as very easy, easy, average, hard, and very hard. Dismissal of very easy or easy marks such as vascular calcifications required only a few seconds whereas dismissal of hard or very hard marks required as many as 3 min of evaluation and review of images. The same 100 screening mammographic studies were then digitized and evaluated on a subsequent version, SecondLook v7.2, and the reviewing radiologist recorded similar data. This CAD system, with high, medium, and low sensitivity operating points, was set to the mid-level case-based sensitivity of 94% at the rate of 2.0 false positive marks per four-image case.

All patients with CAD marks were considered false positives for cancer detection. All marks identified by both CAD versions in these 100 unsuspicious cases were considered false positive marks. McNemar's test was used to compare paired proportions of categorical variables. Chi-square test was used to compare unpaired proportions of categorical variables. Wilcoxon signed-rank test was used to compare paired mean and median values of continuous variables. Mann–Whitney test was used to compare unpaired mean and median values of continuous variables. Two-tailed p values <0.05 were considered statistically significant.

Results

The 100 patients whose screening mammograms were analyzed had a mean age of 57.5 years (SD=10.3; range, 32–84). The distribution of their breast densities is 8% entirely fatty, 53% scattered fibroglandular, 28% heterogeneously dense, and 11% extremely dense. The mean number of films in the mammography study per patient was 4.5 (SD=0.9; range, 4–9).

Table 1 summarizes the specificity rates of CAD v5.0 and v7.2. Considering all marks (masses and calcifications) by CAD, v5.0 marked 77 cases and v7.2 marked 85 cases (specificities of 23% and 15%, p value=0.07) with a common of 71 cases marked by both versions. Considering only masses marked by CAD, v5.0 marked 65 cases and v7.2 marked 83 cases (specificities of 35% and 17%, p value<0.01) with a common of 61 cases marked by both versions. Considering only calcifications marked by CAD, v5.0 marked 38 cases and v7.2 marked 25 cases (specificities of 62% and 75%, p value=0.01) with a common of 21 cases marked by both versions. It is interesting to note that although both versions did not have statistically different specificities when considering all marks, v5.0 had a significantly higher specificity than v7.2 when considering only masses, while v7.2 had a significantly higher specificity when considering only calcifications.

Table 1 Specificities in CAD versions 5.0 and 7.2

	V5.0 (n=100)	V7.2 (n=100)	McNemar's test p value
All marks	23.0%	15.0%	0.07
Masses	35.0%	17.0%	<0.01
Calcifications	62.0%	75.0%	0.01

Table 2 summarizes specificity rates observed in groups of patients with non-dense (n=61) and dense (n=39) breasts. The specificities of both CAD versions, when considering all marks, marks on masses and marks on calcifications, were slightly lower, but not statistically significantly different, in patients with non-dense breasts than in patients with dense breasts.

Table 3 summarizes the false positive rates of v5.0 and v7.2. Among the 100 studies, a total of 234 marks (158 masses and 76 calcifications) were identified by v5.0, and a total of 226 marks (179 masses and 47 calcifications) were identified by v7.2. However, v5.0 and v7.2 varied considerably in their markings and had only 76 common marks (52 masses and 24 calcifications) identified in both versions. As shown in Table 3, v5.0 and v7.2 did not differ significantly in the mean and median numbers of all marks per case (2.3, 2.0 vs. 2.3, 2.0, p value=0.65) nor the mean and median numbers of masses per case (1.6, 1.0 vs. 1.8, 2.0, p value=0.15) but differed significantly in the mean and median numbers of calcifications per case (0.8, 0.0 vs. 0.5, 0.0, p value<0.01).

Table 4 summarizes false positive rates observed in groups of patients with non-dense (n=61) and dense (n=39) breasts. When considering all marks, marks on masses and marks on calcifications, the rates of false positive marks from both CAD versions were slightly higher, but not statistically significantly different, in patients with non-dense breasts than in patients with dense breasts. This is in accordance with the slightly lower specificities observed in patients with non-dense breasts than those with dense breasts.

The reviewing radiologist subjectively rated the ease of dismissal of CAD marks based on the time required to dismiss the mark. Of the 234 marks from v5.0, 197 marks (84%) were classified as very easy or easy to dismiss and 37 (16%) were classified as average, hard, or very hard to dismiss. Of the 226 marks from v7.2, 198 marks (88%) were classified as very easy or easy to dismiss and 28 (12%) were classified as average, hard, or very hard to dismiss. Among the 76 marks identified in both versions, 57 (75%) were classified as very easy or easy to dismiss in v5.0, while 62 (81.6%) were classified as very easy or easy to dismiss in v7.2. It is important to clarify that these 76 marks were the same CAD marks, but the reviewer assessed the marks from v7.2 several years later than the assessment for v5.0. Thus, despite radiologist's knowledge that all CAD marks were false positives, the nonsignificant difference in the ratings of common marks between v5.0 and v7.2 is a reflection of the consistency of this assessment. Among marks that were identified in one or the other version but not both, 140 (88.6%) of 158 marks from v5.0 were classified as very easy or easy to dismiss, while 136 (90.1%) of 150 marks from v7.2 were classified as very easy or easy to dismiss.

Table 2 Specificities in patients with non-dense and dense breasts

	Non-dense breast (n=61)	Dense breast (n=39)	Chi-square test p value
V5.0			
All marks	18.0%	30.8%	0.14
Masses	31.2%	41.0%	0.31
Calcifications	57.4%	69.2%	0.23
V7.2			
All marks	9.8%	23.1%	0.07
Masses	13.1%	23.1%	0.20
Calcifications	70.5%	82.1%	0.19

Discussion

The false positive rate of CAD requires radiologists to dismiss high numbers of CAD marks compared to the numbers of cancers detected in a screening population [1]. With a false positive rate of approximately two marks per four-image mammogram, a typical screening population of 1,000 women would generate 2,000 false positive marks to be dismissed while detecting approximately five cancers based on a mixed incidence/prevalence of five cancers per 1,000 screening mammograms. With each of the cancers detected in two views (craniocaudal and mediolateral oblique), there would be 200 marks to dismiss for each view of a detected cancer. Although these are large numbers of marks to dismiss, our study is the first that we are aware of to provide context regarding ease of dismissing marks. Our study showed that 12% (v7.2) or 16% (v5.0) of these marks were average, hard, or very hard to dismiss. With recall rates from screening mammography approaching these rates of average-to-very-hard marks to dismiss, rather than a 200:1 ratio of distracting/relevant marks based on all false positives compared to views of detected cancers, CAD could be considered to approximately have a 1:1 ratio of distracting/relevant marks based on average-to-very-hard marks to dismiss compared to views of lesions warranting recall.

Zheng et al. [31] have shown the impact of CAD sensitivity and false positive rate on radiologist performance with the four combinations of a CAD system with sensitivities of 90% or 50% and false positive rates of two or eight marks per four-image mammogram case (multiplying their per image false positive rates by 4). Their results showed that a CAD system with 90% sensitivity and two false positives per case improved radiologist performance, a CAD system with 90% sensitivity and eight false

Table 3 False positive rates in CAD versions 5.0 and 7.2

	V5.0 (n=100)		V7.2 (n=100)		Wilcoxon signed-rank test p value
	Mean	Median (interquartile range)	Mean	Median (interquartile range)	
All marks	2.3	2.0 (1.0, 3.0)	2.3	2.0 (1.0, 3.0)	0.65
Marks on masses	1.6	1.0 (0.0, 2.0)	1.8	2.0 (1.0, 3.0)	0.15
Marks on calcifications	0.8	0.0 (0.0, 1.0)	0.5	0.0 (0.0, 0.5)	<0.01

positives per case or a CAD system with 50% sensitivity and two false positives per case had no significant impact on radiologist performance, while a CAD system with 50% sensitivity and eight false positives per case was detrimental to radiologist performance. Since CAD sensitivity is about 90% or higher [22–27], Zheng's results are another indicator that distracting/relevant marks ratio of the CAD systems reported in our study and other studies with current CAD that have about two false positives per case [25–29] should improve radiologist performance [3–19].

Although both versions of the CAD system in our study did not have statistically different specificities when considering all marks, v5.0 had higher specificity in terms of only masses than v7.2, and v7.2 had higher specificity in terms of only calcifications than v5.0. Although CAD specificity is not reported in the literature as commonly as false positive rate, four studies assessed specificity. Two studies were based on mostly or all [25, 29] two-image cases; thus, comparison with our results is difficult. Two other studies used four-image cases, but only reported specificity of CAD for calcifications. Brem et al. [22] reported 63% and 58% specificities for calcifications in non-dense and dense breasts, and Yang et al. [26] reported 67% and 69% specificities for calcifications in non-dense and dense breasts; in both studies, there was no statistical difference in specificity for calcifications in non-dense and dense breasts. Our study had similar results with 57.4%, 70.5% and 69.2%, 82.1% specificities for v5.0 and v7.2 in non-dense and dense breasts, respectively. Interestingly, both versions have slightly lower, but not statistically significant, specificities in terms of all marks, masses only, and calcifications only in non-dense breasts than in dense breasts.

False positive rate is commonly reported in the CAD literature, with earlier studies reporting means of 3–5 false positives per four-image case [3–5, 22–25] and more recent studies reporting means of 2–3 false positives per four-image case [26–30]. Recent studies estimate a range of 0.45 to 0.55 false positive marks per image from digital mammograms [26, 30], indicating that the false positive rate is comparable between film screen and digital mammograms with a mean of 2–3 marks per four-image mammogram. With v7.2, The et al. [28] reported a mean false positive rate of 2.3 false positives per four-image case with digital mammography, which is the same as our mean of 2.3 false positives per four-image case with screen-film mammography.

In our study, v5.0 and v7.2 did not differ significantly in mean and median numbers of all marks and marks only on masses but differed significantly in mean and median number of marks only on calcifications. The most relevant comparisons to our results come from studies with the same versions of the CAD system we studied. With v5.0, Malich et al. [29] and Brem et al. [23] reported mean false positive rates for all marks of 2.5 and 2.4 per four-image case (multiplying Malich's rates by 2 since they used two-image unilateral cases and multiplying Brem's rates by 4 since they reported per image rates), which compare well with our mean

Table 4 False positive rates in patients with non-dense and dense breasts

	Non-dense breast (n=61)		Dense breast (n=39)		Mann–Whitney test p value
	Mean	Median (interquartile range)	Mean	Median (interquartile range)	
V 5.0					
All marks	2.6	2.0 (1.0, 4.0)	2.0	2.0 (0.0, 3.0)	0.23
Marks on masses	1.7	1.0 (0.0, 2.0)	1.4	1.0 (0.0, 2.0)	0.33
Marks on calcifications	0.9	0.0 (0.0, 1.0)	0.6	0.0 (0.0, 1.0)	0.21
V 7.2					
All marks	2.6	2.0 (1.0, 4.0)	1.8	1.0 (1.0, 3.0)	0.05
Marks on masses	2.0	2.0 (1.0, 3.0)	1.5	1.0 (1.0, 3.0)	0.12
Marks on calcifications	0.6	0.0 (0.0, 1.0)	0.3	0.0 (0.0, 0.0)	0.14

of 2.3 false positives per four-image case. The Malich and Brem studies yielded mean false positive rates for masses and calcifications of 1.7, 0.8 and 1.6, 0.8, which also compare well with our mean results of 1.6, 0.8 with v5.0.

Interestingly, both CAD versions in our study have slightly higher, but not statistically significant, mean and median numbers of all marks, marks only on masses, and marks only on calcifications in non-dense breasts than in dense breasts. Brem did not stratify false positive rate by breast density, but Malich did, and although they showed a trend towards lower false positive rates with lower breast densities, the only statistically significant result they reported was a lower mean false positive rate for masses with entirely fatty breasts compared to the other breast density categories. Fewer CAD marks and higher specificity in dense breasts are likely the result of difficulty in identification or differentiation of patterns from dense tissue material.

Version 5.0 and 7.2 did not differ significantly in the proportion of marks classified as very easy or easy to dismiss—neither among the 76 marks that were identified in both versions nor among marks that were identified in one or the other version but not both. Therefore, it appears that although the false positive rate remains at a mean of 2.3 false positive marks per case, the majority of these marks are quickly evaluated by the radiologist and easily dismissed.

Although the total number of marks did not change between v5.0 and v7.2, the fact that only about one third of the marks were common is interesting. Although reproducibility may be a factor due to redigitization of the same screen-film mammograms several years later, this result may largely be a reflection of the significant changes in the CAD algorithms from v5.0 to v7.2. Although CAD sensitivity is not reported in this study, the manufacturer's strategy from v5.0 to v7.2 was to increase CAD sensitivity while maintaining the CAD false positive rate. Since detection of masses is fundamentally more challenging for CAD than detection of calcifications, the false positive rate for calcifications was intentionally reduced to allow for an increase in the false positive rate for masses without a change in the overall false positive rate. This strategy could potentially facilitate increases in CAD sensitivity for masses while maintaining CAD sensitivity for calcifications (Hoffmeister JW, personal communication).

The retrospective nature of this study resulted in some limitations. First, all mammograms evaluated by CAD were screen film. Although there is an increasing trend toward digital mammography, this study still reflects the majority of practices. Second, the assessment of ease of dismissing CAD marks was made by one radiologist, who knew the marks were all false positive. However, the consistency of comparison between the two versions, several years apart, is an important result of this study.

Conclusion

We report specificity and false positive rate for two versions of a CAD system that are consistent with prior literature and provide more detailed stratifications of results by breast density. As far as we are aware, we are the first to show that most false positive CAD marks are easily dismissed by the interpreting radiologist and should not interfere with clinical performance. The slightly better specificities observed with CAD in dense breasts warrant further research in this area.

References

1. Birdwell RL: The preponderance of evidence supports computer-aided detection for screening mammography. Radiology 253:9–16, 2009
2. Philpotts LE: Can computer-aided detection be detrimental to mammographic interpretation? Radiology 253:17–22, 2009
3. Brem RF, Baum J, Lechner M, et al: Improvement in sensitivity of screening mammography with computer-aided detection: a multi-institutional trial. AJR 181:687–693, 2003
4. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al: Potential contribution of computer-aided detection of the sensitivity of screening mammography. Radiology 215:554–562, 2000
5. Freer TW, Ulissey MJ: Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 220:781–786, 2001
6. Birdwell RL, Bandodkar P, Ikeda DM: Computer-aided detection with screening mammography in a university hospital setting. Radiology 236:451–457, 2005
7. Cupples TE, Cunningham JE, Reynolds JC: Impact of computer-aided detection in a regional screening mammography program. AJR 185:944–950, 2005
8. Morton MJ, Whaley DH, Brandt KR, et al: Screening mammograms: interpretation with computer-aided detection—prospective evaluation. Radiology 239:375–383, 2006
9. Dean JC, Ilvento CC: Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. AJR 187:20–28, 2006
10. Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ: Prospective assessment of computer-aided detection in interpretation of screening mammography. AJR 187:1483–1491, 2006
11. Destounis SV, DiNitto P, Logan-Young W, et al: Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. Radiology 232:578–584, 2004
12. Gur D, Sumkin JH, Rockette HE, et al: Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. JNCI 96 (3):185–190, 2004
13. Fenton JJ, Taplin SH, Carney PA, et al: Influence of computer-aided detection on performance of screening mammography. NEJM 356:1399–1409, 2007
14. Gromet M: Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR 190:854–859, 2008
15. Khoo LAL, Taylor P, Given-Wilson RM: Computer-aided detection in the United Kingdom National Breast Screening Programme: prospective study. Radiology 237:444–449, 2005

16. Gilbert FJ, Astley SM, McGee MA, et al: Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. Radiology 241:47–53, 2006
17. Gilbert FJ, Astley SM, Gillan MGC, et al: Single reading with computer-aided detection for screening mammography. NEJM 359:1675–1684, 2008
18. Brancato B, Houssami N, Francesca D, et al: Does computer-aided detection (CAD) contribute to the performance of digital mammography in a self-referred population? Breast Cancer Res Treat 111:373–376, 2008
19. Feig SA, Sickles EA, Evans WP, Linver MN: Re: Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system [letter]. J Natl Cancer Inst 96:1260–1261, 2004. author reply 1261
20. Hukkinen K, Vehmas T, Pamilo M, Kivisaari L: Effect of computer-aided detection on mammographic performance: experimental study on readers with different levels of experience. Acta Radiol 47(3):257–263, 2006
21. Luo P, Qian W, Romilly P: CAD-aided mammogram training. Acad Radiol 12(8):1039–48, 2005
22. Brem RF, Hoffmeister JW, Rapelyea JA, et al: Impact of breast density on computer aided detection for breast cancer. AJR 184:439–444, 2005
23. Brem RF, Rapelyea JA, Zisman G, Hoffmeister JW, Desimio MP: Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology. Cancer 104:931–935, 2005
24. Malich A, Sauner D, Marx C, et al: Influence of breast lesion size and histologic findings on tumor detection rate of a computer-aided detection system. Radiology 228:851–856, 2003
25. Ciatto S, Ambrogetti D, Bonardi R, et al: Comparison of two commercial systems for computer-assisted detection (CAD) as an aid to interpreting screening mammograms. La Radiol Med 107:480–488, 2004
26. Yang SK, Moon WK, Cho N, et al: Screening mammography-detected cancers: sensitivity of a computer-aided detection system applied to full-field digital mammograms. Radiology 244:104–111, 2007
27. Skaane P, Kshirsagar A, Stapleton S, et al: Effect of computer-aided detection on independent double reading of paired screen-film and full-field digital screening mammograms. AJR 188:377–384, 2007
28. The JS, Schilling KJ, Hoffmeister JW, et al: Detection of breast cancer with full-field digital mammography and computer-aided detection. AJR 192:337–340, 2009
29. Malich A, Fischer DR, Facius M, et al: Effect of breast density on computer aided detection. J Digit Imaging 18:227–233, 2005
30. Leon S, Brateman L, Honeyman-Buck J, Marshall J: Comparison of two commercial CAD systems for digital mammography. J Digit Imaging 22(4):421–423, 2009
31. Zheng B, Ganott MA, Britton CA, et al: Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology 221:633–640, 2001