by Dr V. Nikitin, Dr I. Lossev, Dr A. Filatov, N. Bagotskaya & I. Kil

Improvement in sensitivity and specificity of readers using the next generation of mammography CAD

This article describes an independent assessment (carried out by U.S. based CRO and data collection sites in the United States and Europe) of the clinical performance of a Computer-Aided Detection (CAD) software system comparing screening mammography with and without CAD. The results show that both sensitivity and specificity were increased when radiologists reviewed mammography cases with the help of CAD.

The interpretation of mammograms is fraught with challenges. In the asymptomatic screening population, a cancer or true positive finding is rare, found only once every few hundred cases (approx. 4-5/1000) [1], while the large number and wide variety of normal or physiologic changes in a population prompts additional imaging (or call-backs) in up to 10% of the typical screening populations [1]. Independent or consensus double reading by two radiologists as generally practiced in Europe can improve cancer detection by 5-15% but is expensive [2]. Computer-Aided Detection (CAD) has become a critical tool, since it has the potential to reduce the need for double reading while increasing the cancer detection rate. While early reports of CAD performance demonstrated improved cancer detection rates both in prospective and retrospective studies, the high number of false positive CAD marks caused an increase in recall rate ranged between 9%-18%[3]. The recent U.S. Preventive Services Task Force recommendations indicated that despite significant cancer detection rates, the high false positive rate (both false positive call-backs and false positive biopsies) prompted by screening women age 40-49 tipped the risk-benefit ratio towards not recommending routine screening of younger women [4]. Meanwhile, Breast Cancer Surveillance Consortium data have shown that the average sensitivity of mammography in the US is about 84% [5]. Clearly, there is a need for improvement in both the sensitivity and specificity of mammographic screening.

The retrospective reader study described in this article was carried out in three US and two European sites. The study involved 12 radiologists — both general and specialized — reading 240 cases. The six general radiologists typically read fewer than 3,000 cases per year whereas the specialized radiologists read between 3,000 and 10,000 cases per year. The readers independently interpreted 240 Full field digital mammography (FFDM) screening cases (120 cases with cancer, 108 normal cases, and 12 cases with actionable benign findings). The images were obtained either on a Philips MicroDose L30 or a GE Senographe Essential FFDM system. The CAD software used was the Parascript AccuDetect 6.1 Computer-Aided Detection (CAD) system

Dr V. Nikitin, Dr Lossev, Dr A. Filatov, N. Bagotskaya & I. Kil
are at Parascript, Longmont, CO USA
email: vadim.nikitin@parascript.com

METHODS

The following is a short description of the principal features of the methodology used. Full details are available from the authors, as are full details of the characteristics of the cases examined in the study.

The readers used the same single CAD operating point, initially interpreting the cases unassisted by CAD and noting their findings. For every reader, the unassisted interpretation was "locked" before the reader turned on the CAD marks. After the readers turned on the CAD marks to produce the CAD-assisted interpretation, they could add or adjust the findings noted on the unassisted interpretation. For both unassisted and CAD-assisted interpretations, the readers noted the location of any suspected cancer and their level of confidence of whether the finding represented cancer on a 1-100 probability of malignancy (POM) scale. The readers also assigned an overall case-based POM score, which reflected their level of confidence of whether the case was cancerous. The readers also assigned a BI-RADS category 0, 1 or 2 to each case. The readers could change the initial recall decision after CAD-assisted read if they believed it was appropriate. Sensitivity and specificity calculations for unassisted and CAD-assisted interpretations were based on the BI-RADS category assignment: (BIRADS 0) versus (BI-RADS 1 or 2). The data analysis compared reader interpretations with the true status for each case. This article describes the CAD performance based on BI-RADS results.

REFERENCE STANDARD (TRUTH DATA)

For positive cases, Case Report Forms (CRF) were collected as well as reports of the screening exam, and any subsequent diagnostic exams. Pathology reports were also collected for each case, in order to verify the presence of breast cancer, including a marked area boundary for a malignant lesion provided by medical imaging facilities. For normal cases, the index screening FFDM images, the associated CRF, and radiology report were collected together with the subsequent report of a negative FFDM exam performed 320-455 days following the index FFDM exam. For cases with actionable benign findings, the screening FFDM images were collected along with the report of the screening exam, the CRF, and any subsequent diagnostic exams. The pathology report was collected for each case verifying the biopsy of a benign lesion and/or the report

Recall-based Sensitivity and Specificity by Readers				
Reader	Sensitivity		Specificity	
	No CAD	CAD assisted	No CAD	CAD assisted
1	95.8%	97.5%	59.2%	56.7%
2	86.7%	90.0%	80.0%	82.5%
3	89.2%	91.7%	73.3%	73.3%
4	96.7%	98.3%	46.7%	50.0%
5	92.5%	91.7%	70.0%	70.8%
6	95.8%	97.5%	67.5%	66.7%
7	89.2%	93.3%	63.3%	72.5%
8	89.2%	90.0%	63.3%	69.2%
9	94.2%	92.5%	49.2%	66.7%
10	92.5%	92.5%	76.7%	78.3%
11	95.0%	93.3%	53.3%	76.7%
12	85.8%	92.5%	83.3%	80.8%

TABLE 1. Sensitivities and specificities for each reader without and with CAD. The green color depicts improvements in readers' sensitivity and specificity with the use of CAD, while the yellow color indicates a decrease in either sensitivity or specificity with the use of CAD.

of a mammogram performed 320-455 days following the screening mammogram indicating no change in the finding.

STATISTICAL ANALYSES:

To compute the specificity and sensitivity for each reader the readers' BIRADS scores were compared with the truth data.

Recall-based sensitivity and specificity were computed and tabulated for each radiologist. Confidence intervals were computed for average recall-based sensitivity for unassisted reading, average recall-based sensitivity for CAD assisted reading and for average sensitivity difference between CAD assisted and unassisted reading. In addition, confidence intervals were computed for average recall-based specificity for unassisted reading, average recall-based specificity for CAD -assisted reading and for average specificity difference between CAD-assisted and unassisted reading. Two different methods were used for estimating confidence intervals and statistical significance of the results. These were Bootstrapping [7] and Logistic Regression using the Generalized Estimating Equation (GEE) model [8]. Bootstrapping was used to find 95% two-sided confidence intervals; there were 10,000 bootstrapped replications. A bias-corrected and accelerated method [7] was used for computation of confidence intervals. Also, confidence intervals were estimated with logistic regression model. Since results of different radiologists are correlated, the GEE model was used instead of the usual logistic regression.

RESULTS

Sensitivities and specificities for each reader without and with CAD are shown in Table 1. The green color depicts improvements in the readers' sensitivity and specificity with the use of CAD, while the yellow color indicates a decrease in either sensitivity or specificity with the use of CAD. It can be seen that four readers achieved improvements in both sensitivity and specificity with the use of CAD, two readers achieved improvements in either sensitivity or specificity while maintaining the same performance on the other metric, and six readers achieved improvement in one metric (three in sensitivity, and three in specificity) accompanied by a decrease in the other metric. Without CAD, the average sensitivity of readers was 91.9% and the average specificity was 65.5%. The relatively high sensitivity but relatively low specificity may be attributed to the fact that the readers were aware that the set of cases was enriched with cancer cases. It should also be noted that neither prior mammograms nor medical history were available to the readers.

The average increase in BI-RADS-based sensitivity due to the assistance of CAD estimated using bootstrapping was 1.5%, 95% confidence interval (CI) = (0.3%, 2.9%) whereas the estimation using GEE was 1.4%, CI = (0.2%, 2.7%). This difference is statistically significant ($p<0.001$ for bootstrapping and $p<0.025$ for GEE).

The absolute increase in sensitivity of 1.5% might be considered small but the relative improvement (i.e. the ratio of the number of additional cancer cases detected with CAD assistance compared to the number of cancer cases missed with unassisted reading) is 18.5% by bootstrapping estimation and 18.9% by GEE estimation.

The main increase in sensitivity is due to improvement in soft tissue density detection. For cancer cases where at least one soft density exists the average improvement was 2.4%, CI = (1%, 3.7%) and for cancer cases where at least one calcification cluster exists it was only 0.6%, CI = (-0.1%, 1.6%).

The average increase in BI-RADS-based specificity due to CAD assistance (estimated using bootstrapping method) was 4.9%, CI = (2.9%, 6.9%) and 5%, CI = (3.1%, 6.9%) when estimated using GEE. This difference is statistically significant ($p<0.001$ for the bootstrapping method and $p<0.025$ for GEE). The average number of recalls was decreased by 14.2% if estimated by bootstrapping and by 14.9% if estimated by GEE.

CONCLUSION

The clinical study described above was conducted to compare reader performance for screening mammography with and without AccuDetect 6.1 CAD. The study demonstrated that both sensitivity and specificity of the readers were increased when they reviewed mammography cases with the help of the CAD. The average increase in sensitivity was 1.5% (which means that 18.5% of cancer cases missed with unassisted reading were detected with CAD-assisted reading). The average increase in specificity was 4.9%, which means a 14.2% decrease in the number of recalls.

The authors are unaware of any other previous publications where CAD-assisted reading increased both the sensitivity and specificity of cancer detection.

REFERENCES

1. Rosenberg RD *et al.* Radiology 2006; 241: 55
2. Harvey S C *et al.* AJR 2003; 180: 1461.
3. Birdwell RL Radiology 2009; 253: 9
4. "Screening for Breast Cancer: U.S. Preventive Task Force Recommendation Statement". Ann Intern Med. 2009; 151: 716.
5. Breast Cancer Surveillance Consortium Web site - http://breastscreening.cancer.gov/ data/benchmarks/screening/2009/tableSensSpec.htm
6. Breast Cancer Surveillance Consortium Web site - http://breastscreening.cancer.gov/ data/variables/2011/freq_tables_pct.html
7. Efron B & Tibshirani RJ. "An Introduction to the Bootstrap". Chapman & Hall, Boca Raton, Fl: 1994. P14
8. Lipsitz SH *et al.* Biometrics 1994; 50: 270.