Improvement in ROC curves of readers with next generation of mammography CAD

Poster No.:	C-2315
Congress:	ECR 2014
Type:	Scientific Exhibit
Authors:	V. Nikitin, A. Filatov, N. Bagotskaya, I. Kil, I. Lossev, N. Losseva; Longmont, CO/US
Keywords:	Breast, Mammography, CAD
DOI:	10.1594/ecr2014/C-2315

Aims and objectives

Radiologists differ substantially in their interpretation of mammograms, and wide variability in the analysis of screening mammograms has been reported. Independent or consensus double reading by two radiologists can improve cancer detection by 5-15% [1]. In the U.S., single reading is more common because double-reading is time-consuming and expensive. Computer-Aided Detection (CAD) has become a critical tool, since it can reduce the need for double-reading while increasing the cancer detection rate.

This article describes clinical performance assessment of the Parascript *Accu*Detect® 6.1 Computer Aided Detection software performed in order to support Premarket Application for the U.S. Federal Drug Administration. The clinical study was conducted by an independent contract research organization. The study objective was to compare reader performance for screening mammography without and with *Accu*Detect CAD. As per the study protocol, 240 mammography cases were each reviewed by 12 independent radiologists without and with CAD. The area under the receiver operating characteristic (ROC) curve was chosen as the main criterion for comparison between CAD-assisted and non-assisted reading. It will be further referenced as AUC.

The clinical study has found that performance of radiologists in interpreting FFDM images, as measured by the average AUC for patient-based and for lesion-based ROC curves, has improved with assistance of CAD.

Methods and materials

The retrospective reader study was conducted with cases collected from three U.S. sites located in Boca Raton, FL, Rochester, NY and Chicago, IL and two European sites located in Brussels, Belgium and Geneva, Switzerland.

The study was conducted with 12 radiologists reading 240 cases. The group of radiologists included general and specialized radiologists. Six general radiologists read less than 3,000 cases per year and six specialized radiologists read between 3,000 and 10,000 cases per year. The readers independently interpreted 240 screening FFDM cases acquired on Philips MicroDose L30 and GE Senographe Essential FFDM systems. The set of cases consisted of 120 cases with cancer, 108 normal cases, and 12 cases with actionable benign findings. The GE Senographe Essential data set contained the same number and types of cases as the Philips MicroDose L30 data set. Eeach data set contained 60 cancer cases, 54 normal cases, and 6 actionable benign cases.

The readers initially interpreted the cases unassisted by CAD and noted 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 were instructed to use the entire POM scale with the following descriptions of quartile POM values provided as guidance.

- 1 to 25 - probably normal
- 26 to 50 - possibly normal
- 51 to 75 - possibly abnormal
- 76 to 100 - probably abnormal

For each lesion a reader marked its type (calcification or soft tissue density).

Also, the readers assigned BI-RADS category 0, 1 or 2 to each case. The readers were instructed that they could change the initial recall decision after CAD-assisted read if they believe it is appropriate. The analysis compared readers' interpretations with the true status for each case.

Sensitivity and specificity calculations for unassisted and CAD-assisted interpretations were based on the BI-RADS category assignment: (BI-RADS 0) versus (BI-RADS 1 or 2). *Accu*Detect performance based on sensitivity and specificity was analyzed in [2], where it was shown that CAD helped radiologists to increase both sensitivity and specificity.

This article analyzes CAD performance based on POM scores.

Case Collection

All screening mammograms were collected retrospectively. Process of data collection and exclusion criteria are described in [2]. The total number of GE cases received was 305. The total number of Philips cases received was 435. Stratified sampling was applied to the sets of cancer, normal and actionable benign cases to come as close as possible to those distributions of breast density, lesion type and size, cancer morphology and some demographic parameters that are characteristic for the U.S. screening population according to Breast Cancer Surveillance Consortium web site [3].

Reference Standard (Truth Data)

The following data has been collected for cases enrolled in the study:

- Positive cases: the screening FFDM images, the Case Report Form (CRF), radiology report of the screening exam and any subsequent diagnostic exams, and the pathology report verifying the presence of breast cancer, including a marked area boundary for a malignant lesion provided by medical imaging facilities.
- Normal cases: the screening FFDM images, the associated CRF, radiology report, and the subsequent report of a negative FFDM exam performed 320-455 days following the screening FFDM exam.
- Cases with actionable benign findings: the screening FFDM images, CRF, radiology report of the screening exam and any subsequent diagnostic exams, and also the pathology report verifying the biopsy of a benign lesion and/or the report of a mammogram performed 320-455 days following the screening mammogram indicating no change in the finding.

Statistical analysis

The area under ROC curve (AUC) [4] was chosen as the main criterion for comparison between CAD-assisted and non-assisted reading. ROC curves based on overall case-based POM rating were tabulated for each reader for both CAD-assisted and unassisted reading. Estimation and comparison were done using version 2.32 build 2 of DBM-MRCM software [5]. AUCs for these curves were estimated with both non-parametric and parametric methods. This software, in particular, computes p-value of null hypothesis that the average AUC of CAD-assisted reading is equal to the average AUC of non-assisted reading.

Additionally, localization ROC (LROC) curves [6] have been calculated based on POM scores for each lesion. Comparison of LROC curves was done using JAFROC software, version 4.1 [7], which finds statistical significance of difference in average AUCs. In [7], LROC curve is referred to as AFROC curve, and area under LROC curve is referred to as JAFROC Figure of Merit. The method is the same as in DBM software [5], AUCs are computed using Wilcoxon-Trapezoidal non-parametric method. LROC curves and comparison of AUCs for them give possibility to estimate performance of radiologists in localization of lesions [6].

Results

AUC for individual readers

The first method of AUC computation is Wilcoxon-trapezoidal (non-parametric) [5]. The results are shown in Fig. 1 on page 5.

The second method of AUC computation is PROPROC model (parametric) [5]. The results are shown in Fig. 2 on page 6.

Average area under patient-based ROC curves with and without CAD

Fig. 3 on page 7 shows average AUCs for patient-based ROC curves based on case POMs. AUCs are calculated by MMRC-DBM software [5].

Fig. 3 on page 7 shows that both algorithms of AUC computation confirm that average AUC increases when CAD is used, and that the increase is statistically significant ($p<=0.001$).

Average area under lesion-based ROC curves with and without CAD

Lesion-based ROC (LROC) curves have been calculated based on POM values assigned to lesions. These curves can be calculated not only for all cancer lesions but also separately for calcifications and soft tissue densities.

AUCs of L-ROC curves are computed by Wilcoxon-Trapezoidal method used in JAFROC software [7]. Fig. 4 on page 8 shows that CAD-assisted reading increases average AUC for LROC curves for all cancers as well as for calcifications and for soft tissue densities separately. It means that in addition to improvement in cancer detection (Fig. 3 on page 7), CAD also improves cancer localization. Fig. 4 on page 8 also shows that increase in AUC is statistically significant for all cancers as well as for calcifications and for soft tissue densities.

It is interesting to note that average increase in AUC is larger for soft tissue densities compared to AUC increase for calcifications.

Images for this section:

Table 1. Wilcoxon-Trapezoidal AUC by Readers

Reader	No CAD	CAD assisted
1	0.895	0.91
2	0.874	0.907
3	0.913	0.934
4	0.897	0.908
5	0.926	0.939
6	0.911	0.923
7	0.898	0.929
8	0.871	0.919
9	0.794	0.86
10	0.933	0.943
11	0.922	0.943
12	0.916	0.941

Fig. 1: Wilcoxon-Trapezoidal AUC by Readers.

Table 2. PROPROC AUC by Readers

Reader	No CAD	CAD assisted
1	0.922	0.925
2	0.917	0.932
3	0.919	0.934
4	0.906	0.918
5	0.93	0.942
6	0.923	0.932
7	0.91	0.936
8	0.888	0.924
9	0.853	0.87
10	0.94	0.947
11	0.931	0.95
12	0.93	0.95

Fig. 2: PROPROC AUC by Readers.

Table 3. Comparison of average areas under patient-based ROC curves

Model	No CAD (CI)	CAD assisted (CI)	Difference (CI)	P-value
Wilcoxon-Trapezoidal	0.896 (0.862, 0.930)	0.921 (0.893, 0.949)	0.025 (0.013, 0.038)	0.0003
PROPROC	0.914 (0.885,0.943)	0.930 (0.903, 0.957)	0.016 (0.006, 0.025)	0.001

Difference = CAD assisted – No CAD

CI = 95% Confidence Interval

Fig. 3: Comparison of average areas under patient-based ROC curves.

Table 4. Comparison of average areas under lesion-based L-ROC curves

Lesion Type	No CAD (CI)	CAD assisted (CI)	Difference (CI)	P-value
All	0.817 (0.773, 0.861)	0.838 (0.798, 0.879)	0.021 (0.011, 0.031)	0.0001
Calcifications	0.897 (0.858, 0.937)	0.910 (0.873, 0.947)	0.013 (0.003, 0.022)	0.0107
Soft tissue densities	0.825 (0.776, 0.874)	0.848 (0.803, 0.893)	0.023 (0.012, 0.035)	0.0002

Difference = CAD assisted – No CAD

CI = 95% Confidence Interval

Fig. 4: Comparison of average areas under lesion-based L-ROC curves.

Conclusion

Clinical study [2] demonstrated that Parascript *Accu*Detect® Computer Aided Detection software improved cancer detection in such a way that both sensitivity and specificity of readers increased when readers reviewed mammography cases with the help of *Accu*Detect® CAD.

In this article it has been shown that the CAD also statistically significantly improves localization of lesions, which is true for both calcifications and soft tissue densities. Additionally, it has been shown that another measure of cancer detection, area under patient-based ROC curve, is statistically significantly improved in CAD-assisted readings.

Personal information

Vadim Nikitin PhD, Parascript LLC, Longmont, USA.
vadim.nikitin@parascript.com

Alexander Filatov PhD, Parascript LLC, Longmont, USA.
alexander.filatov@parascript.com

Natalia Bagotskaya, Parascript LLC, Longmont, USA.

natasha.bagotskaya@parascript.com

Igor Kil, Parascript LLC, Longmont, USA.

igor.kil@parascript.com

Ilia Lossev PhD, Parascript LLC, Longmont, USA.
ilia.lossev@parascript.com

Nela Losseva, Parascript LLC, Longmont, USA.

nela.losseva@parascript.com

References

1. S. C. Harvey, B. Geller, R. G. Oppenheimer et al. Increase in Cancer Detection and Recall Rates with Independent Double Interpretation of Screening Mammography. AJR May 2003 vol. 180 no. 5, 1461-1467.

2. V. Nikitin, I. Lossev, A. Filatov, N. Bagotskaya, I. Kil. Improvement in both sensitivity and specificity of readers with next generation of mammography CAD. ECR2013, presentation B-0206, 2013.

3. Breast Cancer Surveillance Consortium Web site - http://breastscreening.cancer.gov/data/variables/2011/freq_tables_pct.html

4. T. Fawcett. An introduction to ROC analysis. Pattern Recognition Letters, vol. 27, pp. 861-874, 2006.

5. D. D. Dorfman, K. S. Berbaum, C. E. Metz. Receiver operating charasteristic rating analysis: generalization to the population of readers and patients with the jacknife method. Investigative Radiology, pp. 723-731, 1992.

6. P. Khurd, G. Gindi. Decisions strategies that maximize the area under the ROC curve. IEEEE Trans. Med. Imaging, pp. 1626-1636, 2005.

7. "www.devchakraborty.com," [Online].