Improvement in both sensitivity and specificity of readers with next generation of mammography CAD

Poster No.:	B-0206
Congress:	ECR 2013
Type:	Scientific Paper
Authors:	V. Nikitin, I. Lossev, A. Filatov, N. Bagotskaya, I. Kil; Longmont, CO/US
Keywords:	Oncology, Breast, Computer applications, Mammography, CAD, Computer Applications-Detection, diagnosis
DOI:	10.1594/ecr2013/B-0206

Purpose

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 U.S. screening population [1]. Independent or consensus double reading by two radiologists can improve cancer detection by 5-15% [2]. 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.

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. Preventative 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 US is about 84% [5]. Clearly, there is a need for improvement in both the sensitivity and specificity of mammographic screening.

This article describes clinical performance assessment of the Parascript *Accu*Detect 6.1 Computer-Aided Detection (CAD) 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 CAD. As per the study protocol, 240 mammography cases were each reviewed by 12 independent radiologists without and with CAD.

The study demonstrated that both sensitivity and specificity were increased when radiologists reviewed mammography cases with the help of the CAD.

Methods and Materials

The retrospective reader study was conducted with cases collected from 3 U.S. sites (located in Boca Raton, FL, Rochester, NY and Chicago, IL) and 2 European sites (located in Brussels, Belgium and Geneva, Switzerland). The U.S. sites had obtained

institutional review board approvals for data collection and the European sites followed the laws and regulations for retrospective collection of de-identified data in their respective countries.

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 devices (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 (each data set contained 60 cancer cases, 54 normal cases, and 6 actionable benign cases). 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. Also, the readers 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 believe it is appropriate. 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). The analysis compares readers' interpretations with the true status for each case.

In this article we analyze the CAD performance based on BI-RADS results. The receiver operating characteristic (ROC) curves built using POM will be analyzed elsewhere.

Case Collection:

All screening mammograms were collected retrospectively. A time line approach was used and all available cancer cases were collected sequentially as the screening process occurred. Normal and benign cases were collected sequentially but the time span for collecting required cases was shorter than for the cancer cases. The time period for collected cases is from January 2008 to November 2011.

The total number of cases received was 740. As a result of data review some cases were excluded. The reasons for case exclusion and number of excluded cases are given in Fig. 1 on page 5 .

Stratified sampling was applied to the remaining 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 [6].

Fig. 2 on page 5, Fig. 3 on page 6, Fig. 4 on page 7, Fig. 5 on page 8, and Fig. 6 on page 9 show distributions of study case characteristics.

Reference Standard (Truth Data):

For positive cases, we collected the Case Report Form (CRF), report of the screening exam, and any subsequent diagnostic exams. We also accrued the pathology report for each case verifying the presence of breast cancer, including a marked area boundary for a malignant lesion provided by medical imaging facilities.

For normal cases, we collected the index screening FFDM images, the associated CRF, radiology report, and the subsequent report of a negative FFDM exam performed 320-455 days following the index FFDM exam.

For cases with actionable benign findings, we collected the screening FFDM images along with the report of the screening exam, CRF, and any subsequent diagnostic exams. We also accrued the pathology report for each case 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.

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

Statistical analyses:

Recall-based sensitivity and specificity for each radiologist was computed and tabulated. 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. Also 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.

We used two different methods for estimating confidence intervals and statistical significance of the results: Bootstrapping [7] and Logistic Regression using Generalized Estimating Equation (GEE) model [8].

Bootstrapping was used to find 95% two-sided confidence intervals. There were 10,000 bootstrapped replications. 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, GEE model was used instead of usual logistic regression.

Images for this section:

Reasons for Case Exclusion and Number of Excluded Cases.

Exclusion reason	Number of Cases
Previous surgical biopsy	2
Placement of an internal breast marker	10
Palpable	4
Previous breast cancer	9
Presence of a pacemaker	9
Presence of a breast implant	3
Inadequate technical quality	3
Mismatched for-processing and for-presentation images	9
Absent views	2
Equipment other than specified	2
Other – inconsistencies in case report forms, cancer cases erroneously delivered as normal, absent follow-up reports for normal cases, case duplicates and etc.	13
Total:	66

Fig. 1: Reasons for Case Exclusion and Number of Excluded Cases.

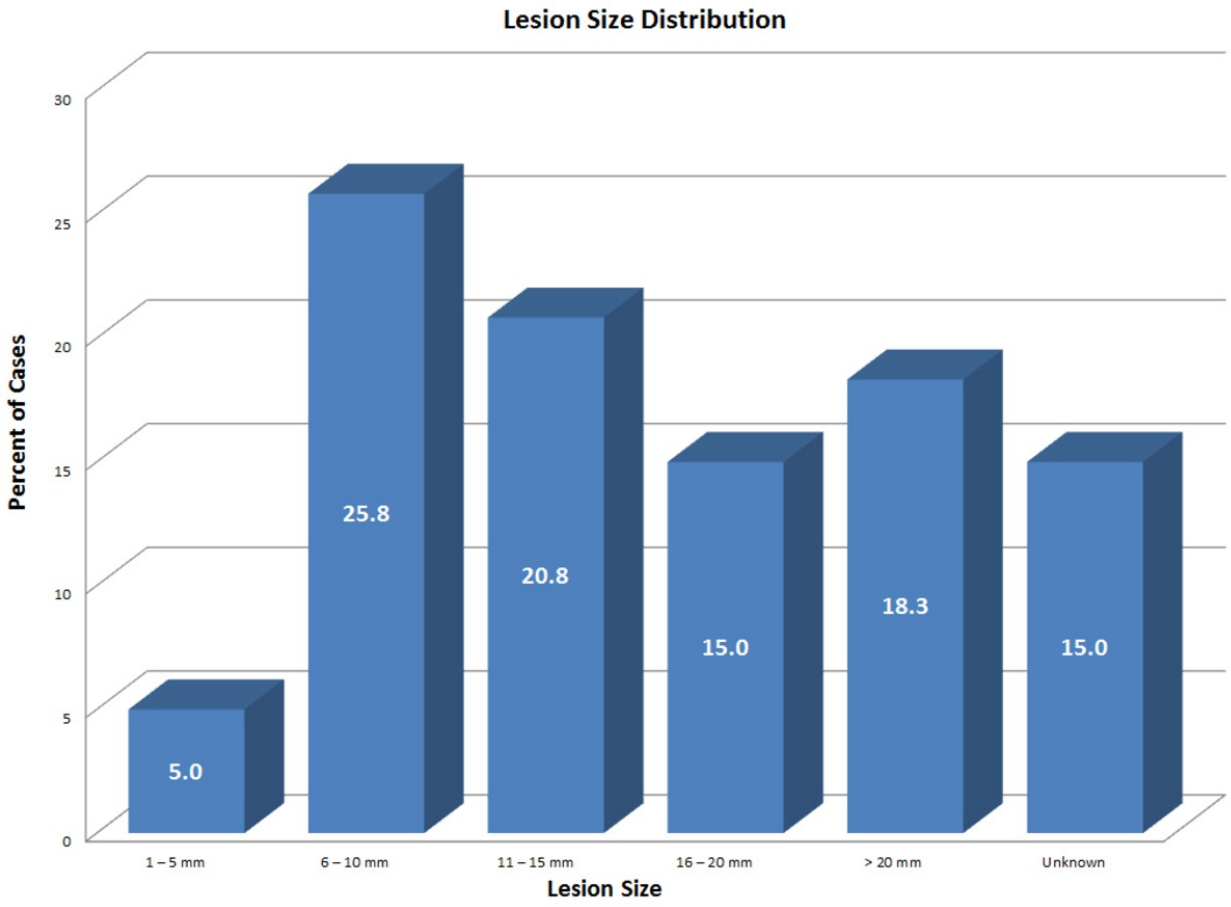

Fig. 2: Lesion Size Distribution.

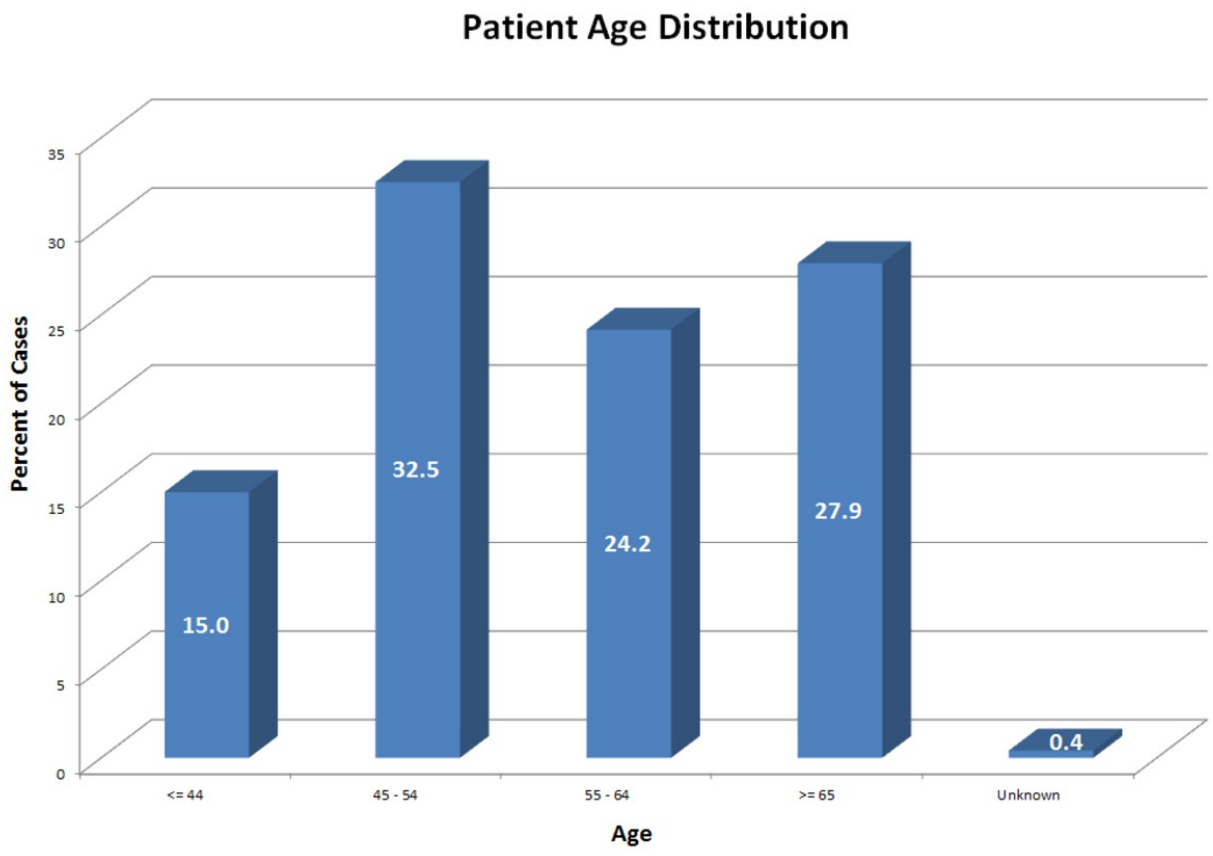

Fig. 3: Patient Age Distribution.

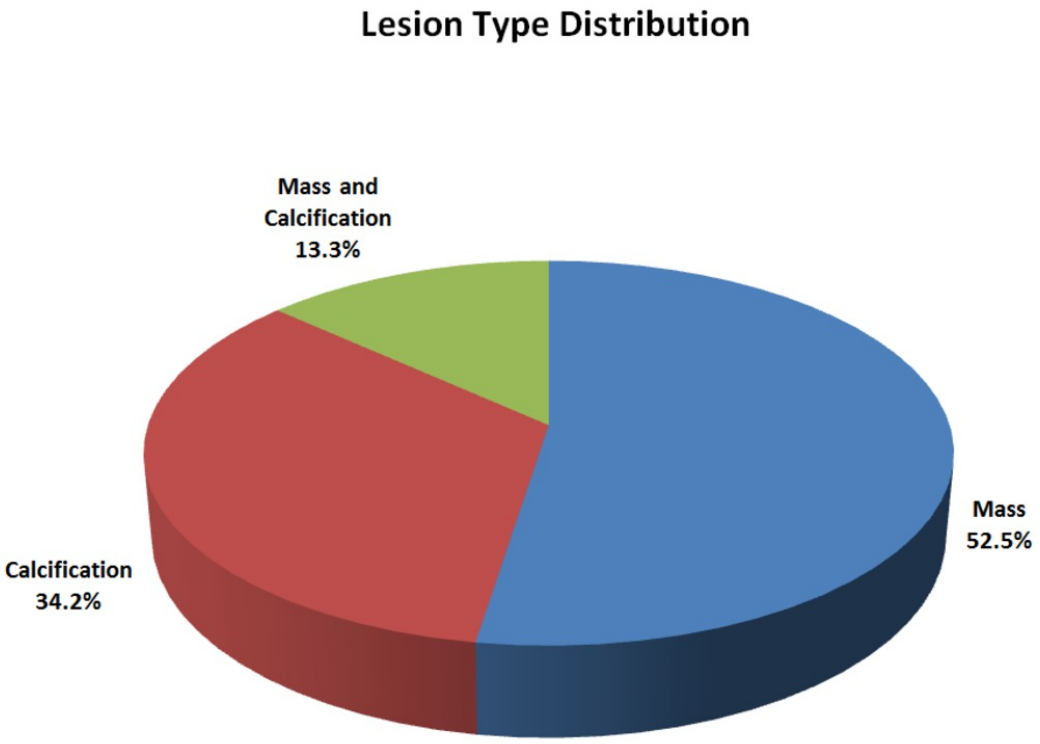

Fig. 4: Lesion Type Distribution.

Pathology Distribution.

Pathology	Number of Cases
Ductal Carcinoma in Situ (DCIS)	28
Invasive Ductal Carcinoma (IDC)	69
Lobular Carcinoma in Situ (LCIS)	1
Invasive Lobular Carcinoma (ILC)	6
Tubular Carcinoma	1
Invasive Mammary Carcinoma	1
DCIS + IDC	8
DCIS + ILC	1
IDC + Papillary Carcinoma + Metastatic Carcinoma	1
Invasive Carcinoma with Ductal and Lobular features	1
Other	3

Fig. 5: Pathology Distribution.

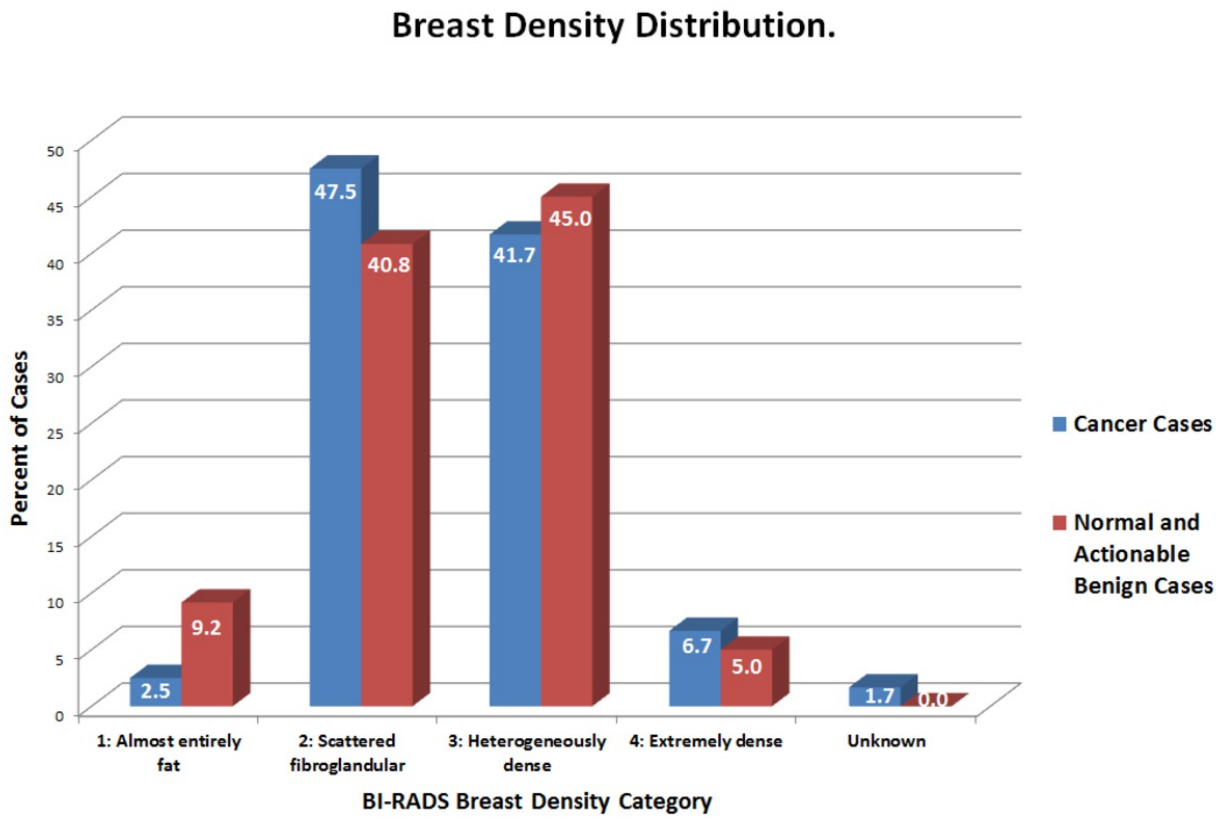

Fig. 6: Breast Density Distribution.

Results

Sensitivity and specificity for each reader without and with CAD is shown in Fig. 7 on page 11. Green color highlights improvements in readers' sensitivity and specificity with the use of CAD, while yellow color indicates the 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 with decrease in the other metric.

Without CAD, average sensitivity of readers is 91.9% and average specificity is 65.5%. Relatively high sensitivity and relatively low specificity may be attributed to the fact that readers were aware that the set of cases is enriched with cancer cases. We should also note that neither prior mammograms nor medical history were available to the readers.

Average increase in BI-RADS-based sensitivity due to assistance of CAD estimated using bootstrapping is 1.5%, 95% confidence interval (CI) = (0.3%, 2.9%). Estimation using GEE is 1.4%, CI = (0.2%, 2.7%). This difference is statistically significant ($p<0.001$ for bootstrapping and $p<0.025$ for GEE). Absolute sensitivity increase of 1.5% can be considered small but relative improvement (ratio of number of additional cancer cases detected with CAD assistance to 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 is 2.4%, CI = (1%, 3.7%) and for cancer cases where at least one calcification cluster exists it is only 0.6%, CI = (-0.1%, 1.6%).

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

Images for this section:

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%

Fig. 7: Recall-based Sensitivity and Specificity by Readers.

Conclusion

The clinical study described above has been conducted to compare reader performance for screening mammography without and with *Accu*Detect 6.1 CAD. The study has demonstrated that both sensitivity and specificity of readers have been increased when readers reviewed mammography cases with the help of the CAD. The average sensitivity increase is 1.5% (which means that 18.5% of cancer cases missed with unassisted reading have been detected with CAD-assisted reading). The average specificity increase is 4.9% which translates into the decrease of the number of recalls by 14.2%. We should note that we are not aware of any previous publications where CAD-assisted reading increased both sensitivity and specificity of cancer detection.

References

1. Rosenberg RD, Yankaskas BC, Abraham LA et al. Performance Benchmarks for Screening Mammography. Radiology*:* Volume 241: Number 1-October 2006.

2. 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.

3. R. L. Birdwell. The Preponderance of Evidence Supports Computer-aided Detection for Screening Mammography. Radiology*:* Volume 253: Number 1-October 2009.

4. "Screening for Breast Cancer: U.S. Preventive Task Force Recommendation Statement". Ann Intern Med. 2009; 151:716-726.

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. Boca Raton, Fl: Chapman & Hall, 1994.

8. Lipsitz SH, Fotzmaurice GM, Orav EJ, Laird NM. "Performance of Generalized Estimating Equations in Practical Situations", Biometrics, 50, 270-278.

Personal Information

Vadim Nikitin PhD, Parascript LLC, Longmont, USA.

vadim.nikitin@parascript.com

Ilia Lossev PhD, Parascript LLC, Longmont, USA.

ilia.lossev@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