

# DL Precise™ Improves Mammographer Performance in Simulated User Testing

## Abstract

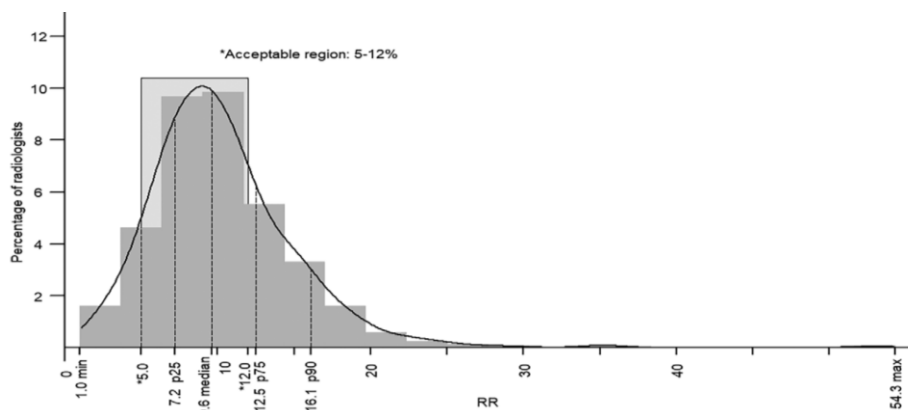
**A metropolitan New York hospital improved its performance by 12% – relative to ACR guidelines – during a 2024 study in which DL Precise™ was evaluated in a simulated use testing in its breast cancer screening program.** In the testing, six radiologists considered 17 patient screening mammograms (retrospective, single-blinded) with and without DL Precise™. With the DL Precise™ intervention, the cohort made a net 12 assessment changes that moved the cohort closer to the National Performance Benchmark<sup>1</sup>. As part of the discussion, this paper discusses two example cases that illustrate the dynamics of that improvement. Contrary to post hoc quality tools that report unassisted performance, DL Precise™ provides in-workflow image enhancement that may favorably affect recall rate performance at the time of reading. A study with a larger cohort of readers and cases will be used to confirm these initial results.

## Introduction

Healthcare quality in general, and Mammography quality in particular, are hot topics (170k hits for mammography quality on scholar.google.com). The American College of Radiology (ACR) and the US Congress through the Mammography Quality Standards Act (MQSA) have established quality standards<sup>1</sup>. These standards have provided fertile soil for a micro-industry of “quality improvement” tools, databases, and analytical software<sup>2</sup>. **But few quality tools provide in-workflow information that could impact the interpretation of mammograms**, affecting quality before it gets to the retrospective report. This pilot study sought to evaluate one such tool for its ability to change the statistical performance of a cohort of readers.

## Existing Literature

In *Radiologist Characteristics Associated with Interpretive Performance of Screening Mammography: A National Mammography Database (NMD) Study*<sup>3</sup> Lee et al compared the acceptable range of recall rates with national actual recall rates. Figure 1 is extracted from that report. **37% of readers fell outside of the acceptable range**, representing likely over-utilization of follow-up studies or likely under sensitivity to the presence of cancer.



**Figure 1** - Recall Rates from Lee et al. Dark gray bars are from national actuals; light gray box is the “Acceptable Range”

<sup>1</sup> See Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283(1):49–58.

<sup>2</sup> <https://www.qualtrics.com/> is one illustrative example among 1000s of such packages.

<sup>3</sup> *Radiology* 2021; 300:518–528 • <https://doi.org/10.1148/radiol.2021204379>

## Methods

Six radiologists were recruited from a high-volume, non-research hospital in metropolitan New York. Five of the six radiologists had enough tenure at the institution to have established individual recall rates that could be used for comparison. Their historical recall rates are shown in Figure 2.



**Figure 2** - Historical Recall Rates for radiologists participating in the pilot study. Note that Reader 4 did not have enough history at this institution to have an established recall rate for comparison.

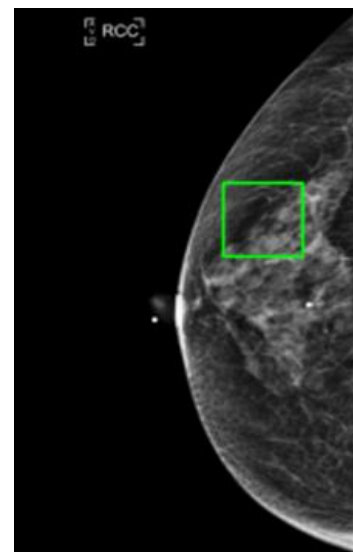
The Radiologists were asked to consider 17 anonymized screening mammograms selected by the researchers from another institution's database. These 17 cases were selected randomly from 116k cases. The selection criteria included:

- 1) A single pathology result (i.e. no cases with multiple pathology findings)
- 2) A region of interest curated by radiologists unaffiliated with the institution in this study (see Figure 3)
- 3) A cancer incidence rate in the selected sample that is close to the incidence rate of (low single digit percent) of the population of the United States.

Only Patient 15 had a biopsy proven cancer (1/17 = 5.8%), and the readers were blinded to this incidence rate.

The readers used a mammography viewer (3Palm; Carmel California), a Barco Uniti Monitor (Kortrijk, Belgium), and DeepLook Medical's (New Haven, CT) DL Precise™ software. They were asked to provide an initial BI-RADS (see Table 1) category assessment for each of the mammograms.

<b>Table 1:</b> BI-RADS Assessment Categorization System <sup>4</sup> . In this study, only categories 0, 1, and 2 were considered for screening mammograms.	
Category 0	Incomplete test, requiring additional imaging or comparison to previous mammograms
Category 1	Negative result, with no significant abnormalities
Category 2:	Benign finding, such as a mass, calcification, or lymph node



**Figure 3** - Sample Region of interest (ROI) from the test set

<sup>4</sup> <https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Reference-Card.pdf>.

The Recall Rates for each radiologist were recorded. A Category 0 assessment will trigger follow-up imaging or biopsy. As such it was used as a proxy for a Recall.

Following a 7–10-day washout period the same radiologists read the same 17 mammograms, this time using DL Precise™ to colorize the images and their recall rates were recorded a second time in the same manner. The researchers used Excel to compare the assessments from the two reads for the purpose of this paper's analysis.

## Results

Table 2 compares the changes in BI-RADS assessment before and after the introduction of DL Precise™. In each square is a figure that represents the number of readers who selected a pre- and post-assessment change (or no change). A green highlight represents reduced utilization of follow-up imaging. A red highlight shows an increased use of follow-up imaging. The blue square indicates increased utilization but is a special case considered in the Discussion section below.

Change	Patient Number																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0>0	1	3			2		1	1			4	6	4	6	5		6
0>1																1	
0>2		3			3		1	1		1			1				
1>0											1						
1>1			1	1													
1>2			1			1			2	1							1
2>0	3						1	2	1			1		1		1	
2>1						1											
2>2	2		4	5	1	4	3	2	3	4							4
?											1						

**Table 2:** Changes in BI-RADS assessments with/without DL Precise

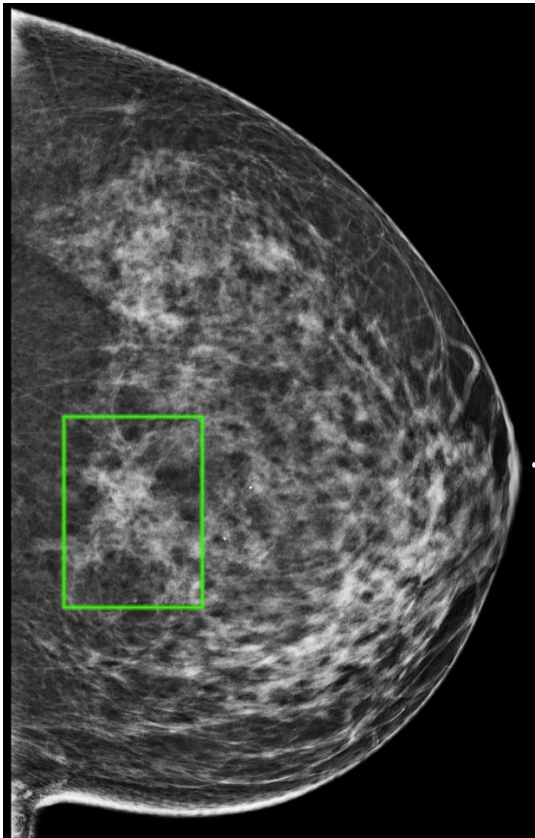
## Discussion

The radiologists were given a target region of interest for their assessment. This was not a test of the radiologist's ability to detect a lesion; it was only a test of their assessment of the seriousness of a pre-identified area. The incidence rate of BI-RADS 2 is higher than in the public because all studies had some lesion pre-identified for the readers.

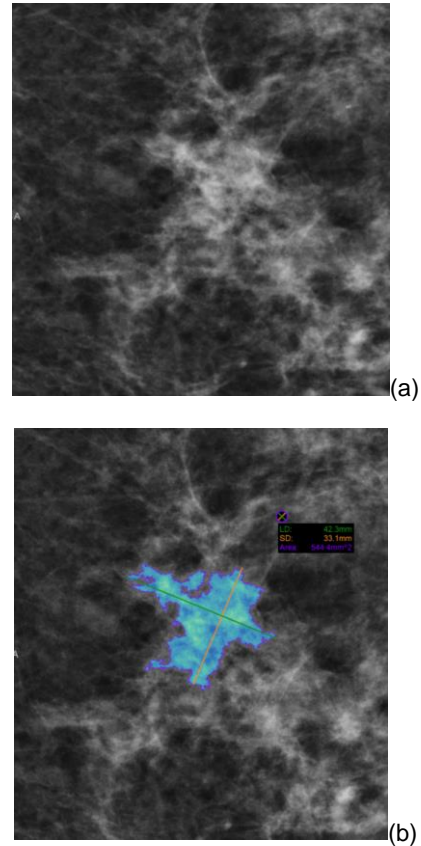
The dominant change to lower utilization occurred from BI-RADS 0 (Requiring Follow-Up) to BI-RADS 2 (Benign Finding). The dominant change to higher utilization was in the reverse direction. The difference between the sums of those two change types, divided by the 102 case-reader pairs, accounts for a 12% favorable change in utilization rate.

Patients 2 and 5 are notable for the number of downgrades from BI-RADS 0 to 2. See Figure 4 for an illustrative comparison of patient 5. Under normal viewing conditions the lesion would be magnified for evaluation.

ROI Identified for the study participants



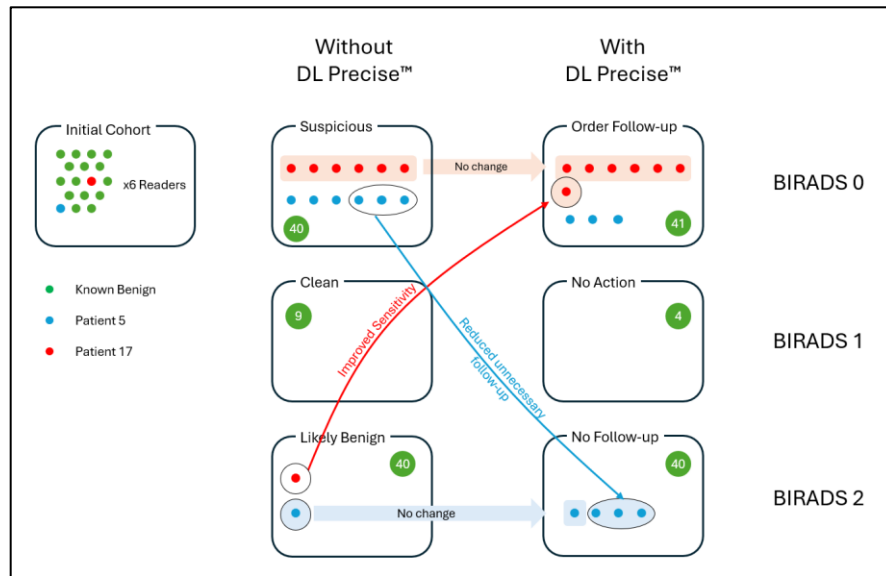
Without (a) and with (b) DL Precise



**Figure 4** - The identified region of interest for patient 5 and the lesion highlighted by DL Precise™

- Patient 15 is notable for two reasons. First, it is the only one of the images with a known positive biopsy. Furthermore, it is notable because Reader 1 (See Table 2) scored it initially as BI-RADS = 2 and reclassified it as BI-RADS = 0, effectively swapping a false negative for a true positive, and increasing her overall detection rate. If by coincidence, Patient 15's mammogram was read by Reader 1, she would have gone another year without cancer detection. Note that Reader 1 is also the radiologist with the historically lowest call-back rate.
- Reader 6 is also notable because she reduced the number of call backs by three, from 9-of-17 to 6-of-17. She is also the reader with the highest historical call-back rate.

The impact of DL Precise™ on these two patients is illustrated in Figure 5.



**Figure 5** - Changes in patient care path associated with the intervention of DL Precise™.

It is notable that even in the presence of identified ROIs, 9 reads were assessed as BI-RADS Category 1 in the “without DL Precise” phase of the test. Essentially there is a “nothing to see here” decision even when an area of concern is implied by the ROI. Five of these 9 reads were reclassified to 0 or 2 following the introduction of DL Precise™. This suggests increased sensitivity, even in a group that was tacitly primed to find something.

## Conclusion

With only 6 readers and only 17 cases, this simulated use testing was not powered for sweeping generalizations. However, taken as a low-power study plus two notable case reports, the data points to an important impact of DL Precise™.

First, **reimbursement and capitation motivate management to press for their radiologists to perform statistically within the ACR guidelines** (See Figure 1). When radiology groups consistently perform outside of the “Acceptable Range”, CMS can discount their reimbursement until they return to compliance. This is translated into management mandates to fix the departmental performance.

Generally, this feels like to Radiologists as pressure to manage a number rather than pressure to make the right clinical call. **Without tools to better differentiate borderline cases, radiologists will either cave to the pressure to order fewer exams or stand out as the non-compliant problem in the department.** Broadly speaking radiologists are caught between two risks: being sued for under-calling suspicious lesions and being scolded for over-ordering follow-up. DL Precise™ may be a tool for improving the statistical department performance while giving the radiologists information that helps them to be comfortable with the management nudge to make those calls.

Furthermore, in at least one case, **the intervention of DL Precise™ was correlated with a new cancer detection.**

DL Precise™ differs from other software in the broad category of Quality Improvement applications. Many quality improvement applications are simply data collection and reporting tools that provide alerts when processes move outside of predetermined limits. The application of DL Precise™ is an intervention at the point-of-care in the normal breast imaging workflow that may nudge clinical decisions toward improved quality outcomes.

A properly powered, prospective study may help to solidify these initial impressions.