Clinical Insights on Autonomous AI Implementation: A Radiologist's Perspective

Connie Lehman MD PhD FACR FSBI
Chief of Breast Imaging MGH
Professor of Radiology Harvard Medical School
Every Year worldwide:
- Out of > 2 billion adult women who are of screening age:
  - > 2 million diagnosed with and
  - > 600,000 die from breast cancer

When detected early through screening mammography, breast cancer is cured.
Key Ingredients

- Large volumes of specific exam type
- Structured reports
  - Binary labelling of exams as “normal” or “not normal”
- Auditing
  - Performance metrics at individual and group level
  - Abnormal exams per 100 screened (AIR) and cancers detected per 1000 screened (CDR)
- Follow up for outcomes
  - Binary labelling of outcome as positive or negative for cancer 365 days from index exam
Autonomous AI Image Interpretation

• What is it?
• Radiologists’ and patients’ perspectives
  – Benefits
  – Risks
• Clinical scenarios
  – Appropriate use cases
• Transition from computer assisted to autonomous reading
  – Clinical implementation
Computer Assistance vs Autonomous Interpretation

Traditional CAD

AI Triage

Fully Autonomous Interpretation of Subset of Exams

Fully Autonomous Interpretation of All Exams for Specific Case Use
Autonomous AI Image Interpretation

• What is it?

• Radiologists’ and patients’ perspectives
  – Benefits
  – Risks

• Clinical scenarios
  – Appropriate use cases

• Transition from computer assisted to autonomous reading
  – Clinical implementation
HEALTH

Most of the World Doesn't Have Access to X-Rays
One hospital in Boston has 126 radiologists. Liberia has two.

JASON SILVERSTEIN  SEPTEMBER 27, 2016
HUMAN VARIATION

Recall  PPV  Sensitivity  Specificity

Source: Buist, Workshop Summary, 2015
DL Model to Enhance Screening Test performance: Image Interpretation (any reason to suspect cancer?)

*Wide variation* in human accuracy in mammographic interpretation

- 40% of U.S. certified breast imaging radiologists perform outside recommended ranges for acceptable false positive rates
- Wide range for acceptable cancer detection rates

Lehman et al *Radiology* April 2017
Breast Cancer Screening Using Tomosynthesis in Combination With Digital Mammography

Sarah M. Friedewald, MD; Elizabeth A. Rafferty, MD; Stephen L. Rose, MD; Melissa A. Durand, MD; Donna M. Plecha, MD; Julianne S. Greenberg, MD; Mary K. Hayes, MD; Debra S. Copit, MD; Kara L. Carlson, MD; Thomas M. Cink, MD; Lora D. Barke, DO; Linda N. Greer, MD; Dave P. Miller, MS; Emily F. Conant, MD


Recall rates
- Digital mammography alone
- Digital mammography plus tomosynthesis

- Pooled performance
- Model estimate

No. of Cancers Detected per 1000 Screens

Patients Recalled After Screening, %
Comparing Diagnostic Performance of Digital Breast Tomosynthesis and Full-Field Digital Mammography

Laila R. Cochon, Catherine S. Giess, Ramin Khorasani, MD

Table 2. Individual radiologists' recall rate by period*

<table>
<thead>
<tr>
<th>Attending</th>
<th>FFDM-Only Recall Rate, n (%)</th>
<th>Hybrid Recall Rate, n (%)</th>
<th>DBT-Only Recall Rate, n (%)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>428 of 4,363 (9.8)</td>
<td>740 of 7,265 (10.2)</td>
<td>83 of 726 (11.4)</td>
<td>.46</td>
</tr>
<tr>
<td>B</td>
<td>129 of 1,882 (6.8)</td>
<td>405 of 5,280 (7.7)</td>
<td>253 of 3,468 (7.3)</td>
<td>.53</td>
</tr>
<tr>
<td>C</td>
<td>109 of 1,132 (9.6)</td>
<td>229 of 1,917 (11.9)</td>
<td>38 of 270 (14.0)</td>
<td>.09</td>
</tr>
<tr>
<td>D</td>
<td>205 of 2,123 (9.7)</td>
<td>628 of 5,512 (11.4)</td>
<td>461 of 4,383 (10.5)</td>
<td>.11</td>
</tr>
<tr>
<td>E</td>
<td>486 of 6,740 (7.0)</td>
<td>642 of 8,564 (7.5)</td>
<td>415 of 4,931 (8.4)</td>
<td>.06</td>
</tr>
<tr>
<td>F</td>
<td>323 of 4,384 (7.4)</td>
<td>550 of 6,210 (8.9)</td>
<td>741 of 7,795 (9.5)</td>
<td>&lt;.001†</td>
</tr>
<tr>
<td>G</td>
<td>479 of 4,765 (10.0)</td>
<td>1,252 of 10,032 (12.4)</td>
<td>700 of 6,301 (11.1)</td>
<td>&lt;.001†</td>
</tr>
<tr>
<td>H</td>
<td>41 of 535 (7.7)</td>
<td>218 of 2,517 (8.6)</td>
<td>31 of 380 (8.2)</td>
<td>.77</td>
</tr>
<tr>
<td>I</td>
<td>107 of 1,310 (8.2)</td>
<td>277 of 2,350 (11.8)</td>
<td>530 of 4,021 (13.2)</td>
<td>&lt;.001†</td>
</tr>
<tr>
<td>J</td>
<td>225 of 2,794 (8.1)</td>
<td>622 of 7,398 (8.4)</td>
<td>323 of 3,780 (8.5)</td>
<td>.79</td>
</tr>
<tr>
<td>K</td>
<td>162 of 1,983 (8.2)</td>
<td>59 of 656 (8.5)</td>
<td>501 of 4,527 (11.1)</td>
<td>.003†</td>
</tr>
<tr>
<td>L</td>
<td>143 of 1,875 (7.6)</td>
<td>505 of 5,514 (9.2)</td>
<td>283 of 3,451 (8.2)</td>
<td>.11</td>
</tr>
<tr>
<td>M</td>
<td>133 of 1,171 (11.3)</td>
<td>419 of 3,555 (11.8)</td>
<td>357 of 2,987 (11.9)</td>
<td>.89</td>
</tr>
<tr>
<td>Pooled</td>
<td>867 of 6,223 (13.9)</td>
<td>1,279 of 9,020 (14.1)</td>
<td>2,020 of 14,127 (14.2)</td>
<td>.83</td>
</tr>
</tbody>
</table>

*Recall rates of individual radiologists not present during one or more periods were pooled.
†Statistically significant.
By Mei-Sing Ong and Kenneth D. Mandl

National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year

ABSTRACT Populationwide mammography screening has been associated with a substantial rise in false-positive mammography findings and breast cancer overdiagnosis. However, there is a lack of current data on the associated costs in the United States. We present costs due to false-positive mammograms and breast cancer overdiagnoses among women ages 40–59, based on expenditure data from a major US health care insurance plan for 702,154 women in the years 2011–13. The average expenditures for each false-positive mammogram, invasive breast cancer, and ductal carcinoma in situ in the twelve months following diagnosis were $852, $51,837 and $12,369, respectively. This translates to a national cost of $4 billion each year. The costs associated with false-positive mammograms and breast cancer overdiagnoses appear to be much higher.
Risks

• Performance of models in virtual setting does not translate to “real world” clinical practice
  – Lessons learned from traditional CAD
Influence of Computer-Aided Detection on Performance of Screening Mammography

Study Limitations

- Data from early years of CAD integration (1998-2002)
- Didn’t control for learning curve (weeks to a year to learn to use CAD)
- Outdated “obsolete” technology (film screen CAD)
- Low numbers (25k CAD exams)
Challenges addressed by BCSC:

No improvement of digital mammography performance with CAD

Odds ratio for CAD vs. No CAD adjusted for site, age, race, time since prior mammogram and calendar year of exam using mixed effects model with random effect for exam reader and varying with CAD use found no significant difference in sensitivity, specificity or recall rate.
Risks

• “Cancers are missed by AI model that would have been diagnosed by a human reader”
  – What is our comparison given background of wide variation in human diagnostic performance?
Risks

• Excellent AI tools developed in a specific population for a specific use are used inappropriately in “wrong use” cases
  – Who will oversee/monitor “off label” use?
  – Benefit of required audits---opportunity to leverage this asset
  – Consider higher bar for performance reporting if AI used autonomously
Risks associated with fear....
“I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon. You’re already over the edge of the cliff, but you haven’t yet looked down. There’s no ground underneath. It’s just completely obvious that in five years deep learning is going to do better than radiologists. It might be ten years.”

— Geoffrey Hinton, *A.I. Versus M.D.*
Hinton now qualifies the provocation. “The role of radiologists will evolve from doing perceptual things that could probably be done by a highly trained pigeon to doing far more cognitive things,” he told me.

“Early and accurate diagnosis is not a trivial problem. We can do better. Why not let machines help us?”
Autonomous AI Image Interpretation

- Radiologists’ and patients’ perspectives
  - Benefits
  - Risks
- Clinical scenarios
  - Appropriate use cases
- Change management from computer assisted to autonomous reading
  - Clinical implementation
Mammography as a Screening Examination in Breast Cancer¹

JOHN N. WOLFE, M.D.²


² Supported by grants from the Michigan Cancer Foundation and Woman’s Hospital Research Fund.

Associate Radiologist, Woman’s Hospital, Detroit, Mich.

The tedious task of examining about 250 women to detect one cancer seems relatively unrewarding unless it is realized that the cancer found is most likely to be in a curable stage. If left until it is clinically evident, the likelihood of salvage diminishes rapidly.
Interpreting/Triaging Mammograms

1. Routine Screening
   1000 Patients/200,000 images

2. Call back for Additional Imaging
   100 Patients

3. Biopsy
   20 Patients

4. Diagnosis
   5 Patients dxed
   < 1 Patient missed cancer
Mammography Interpretation and AI

- Breast Density?
- Normal or Not?

ACR BI-RADS® ATLAS
Breast Imaging Reporting and Data System

2013

[Images of mammograms showing different breast densities]
DL Model to Enhance Screening Test performance:

Image Interpretation (breast density)

Early Matters

Breast density is one of the strongest predictors of the failure of mammography screening to detect cancer.

Brian L. Sprague, PhD; Emily F. Conant, MD; Tracy Onega, PhD; Michael P. Garcia, MS; Elisabeth F. Beaber, PhD; Sally D. Herschorn, MD; Constance D. Lehman, MD, PhD; Anna N.A. Tosteson, ScD; Ronilda Lacson, MD, PhD; Mitchell D. Schnall, MD, PhD; Despina Kontos, PhD; Jennifer S. Haas, MD, MSc; Donald L. Weaver, MD; William E. Barlow, PhD; on behalf of the PROSPR Consortium *
97% agreement with expert readers in test setting
94% agreement with human readers in routine clinical practice
Radiologist vs Model
Radiologist vs Model

- Fatty
- Scattered
- Heterogeneous
- Dense
DL Model to Enhance Screening Test performance:

Image Interpretation (any reason to suspect cancer?)

**Training Set:**
- Patients: 56,835
- Exams: 212,020
- No Exclusions

**Testing Set:**
- Patients: 7,234
- Exams: 26,736
- No Exclusions

**Model AUC:** 0.85

**Community Certified MDs:** 0.87

**MGH Expert MDs:** 0.92

**Radiologist False Positive Assessments by Risk Percentile**
- FP triaged below threshold
- FP triaged above threshold
Machine-driven triage is pathway to autonomous AI interpretation.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>% Exams Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGH Radiologist</td>
<td>90.6% (86.7, 94.8)</td>
<td>93.0% (92.7, 93.3)</td>
<td>100% (100, 100)</td>
</tr>
<tr>
<td>MGH Radiologist + AI Triage</td>
<td>90.1% (86.1, 94.5)</td>
<td>93.7% (93.0, 94.4)</td>
<td>80.7% (80.0, 81.5)</td>
</tr>
<tr>
<td>BCSC Community Rad</td>
<td>86.9% (86.3, 87.6)</td>
<td>88.9% (88.8, 88.9)</td>
<td>100% (100, 100)</td>
</tr>
</tbody>
</table>

A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. 
Yala A¹, Schuster T¹, Miles R¹, Barzilay R¹, Lehman C¹.


Knowledge of effective strategies for clinical implementation essential

- Breast density DL platform in place now at MGH and implemented in routine clinical care
  - 50,000 screening mammograms/year performed/processed
- 1 (triage), 2 and 5 year risk assessment DL model platform in place at MGH and under evaluation for performance
- Rigorous peer reviewed original scientific publications
Future

– Experts in the imaging sciences are best prepared to lead the paradigm shift from human reading to AI autonomous reading

– Quality and safety measures developed and implemented for humans reading images can translate to autonomous settings
  • MQSA audit: recall, cancer detection rates, sensitivity and specificity
    – Opportunity to exceed current strategies
Imaging sciences will join the autonomous AI revolution and can benefit from lessons learned in domains of automotive, military and banking enterprises.
Multidisciplinary approach needed to leverage strengths from domains of healthcare, computer science, regulatory bodies
Thank you