Implementing ML in Lung CT Screening: Challenges + Hurdles

October 2020

A. Gregory Sorensen MD
DeepHealth, Inc.

Disclosures: Listed on LinkedIn
Q: What is the most powerful force in American medical practice?

A: Newton’s First Law  =  Inertia

Why?

• Revenue involved?
• Physicians intuition (wrongly or rightly)
Some leading reasons for inertia
(or mistrust, or concern...)
Based on what we’ve seen with breast cancer screening

- Baseline performance ("Why do I need medicine if I’m not sick?")
- Improvement not present in actual practice (aka “the real world”)
- Variability (Not as simple as “good” or “bad” reader)
The good news: Screening works

Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening

The National Lung Screening Trial Research Team*

- What can we say about *human interpretation* process?
In breast cancer screening, a certain reluctance to measure and describe human performance

**NEJM 1994:**
10 expert readers, kappa = 0.47

**Nature 2020:**
AI versus Human performance

**Breast cancer in 2 years (USA)**

- **AI AUC 0.740**
- **Human AUC 0.625**

**DMIST NEJM 2005:**
also shows human performance no better than 70% sens/spec
In breast cancer screening, challenges to measure and describe human performance

While screening works overall, many cancers are missed; gatekeeper bias

Exactly how many is disputed: DMIST found 7.8 cancers per 1000 women followed for 15 months after screening (5.9 in 365 days or less)

Breast cancer incidence (from SEER) is ~4.5 per 1000 women per year (age 50+)

**Table 3**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIR (recall rate) (%)</td>
<td>10.9</td>
<td>10.0</td>
<td>11.6 (11.5, 11.6)</td>
<td>10.0</td>
</tr>
<tr>
<td>CDR (per 1000 examinations)</td>
<td>4.8</td>
<td>4.3</td>
<td>5.1 (5.0, 5.2)</td>
<td>3.43</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>78.7</td>
<td>84.9</td>
<td>86.9 (86.3, 87.6)</td>
<td>NA</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>89.5</td>
<td>90.3</td>
<td>88.9 (88.8, 88.9)</td>
<td>NA</td>
</tr>
<tr>
<td>FNR (per 1000 examinations)</td>
<td>0.8 (0.7, 0.8)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

**TABLE 3** Final Cut-Points for Screening Mammography Using the Angoff Method

<table>
<thead>
<tr>
<th>Measure</th>
<th>Low Performance Range</th>
<th>Percentage of the BCSC Radiologists in Low Performance Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>&lt;75</td>
<td>18.0%</td>
</tr>
<tr>
<td>Specificity</td>
<td>&lt;88 or &gt;95</td>
<td></td>
</tr>
<tr>
<td>Recall rate</td>
<td>&lt;5 or &gt;12</td>
<td>47.7%</td>
</tr>
<tr>
<td>PPV&lt;sub&gt;1&lt;/sub&gt;</td>
<td>&lt;3 or &gt;8</td>
<td>38.4%</td>
</tr>
<tr>
<td>PPV&lt;sub&gt;2&lt;/sub&gt;</td>
<td>&lt;20 or &gt;40</td>
<td>34.0%</td>
</tr>
<tr>
<td>Cancer detection rate</td>
<td>&lt;2.5/1,000</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

NOTE: BCSC = Breast Cancer Surveillance Consortium; PPV = positive predictive value.
In breast cancer screening, challenges to measure and describe human performance

- Do we radiologists not want to report our performance?
In breast cancer screening, laboratory performance does not always equal “real world”

- While CAD had shown benefit in FDA studies, in practice this benefit was not evident
- Many radiologists individually had already had this perception
- Reimbursement, perceived legal risks still drive usage even today
- What went wrong?
Human interpreters are not just “good readers” or “bad readers” and AI should probably reflect that...

- Human readers do not easily shift where they operate on the ROC curve

AI AUC 0.740
Human AUC 0.625

Nature 2020
https://arxiv.org/abs/1912.11027
Recommended priorities to consider

• Measure performance pre-intervention ("Pre-AI"). Take an honest look, be it under QA protection (or whatever)
  • This is harder than it sounds; note that MQSA requires performance reporting already for screening mammography, and yet...

• Measure performance along the way (whatever “real world” you live in...)

• Consider designing tailored interventions to boost reader acceptance
Conclusion

• Quality and cost considerations strongly drive adoption of AI... *in theory*
• Barriers to adoption are real but can be overcome

Priorities
• Measure performance pre-intervention (“Pre-AI”)  
• Measure performance along the way (whatever “real world” you live in...)
• Tailored interventions