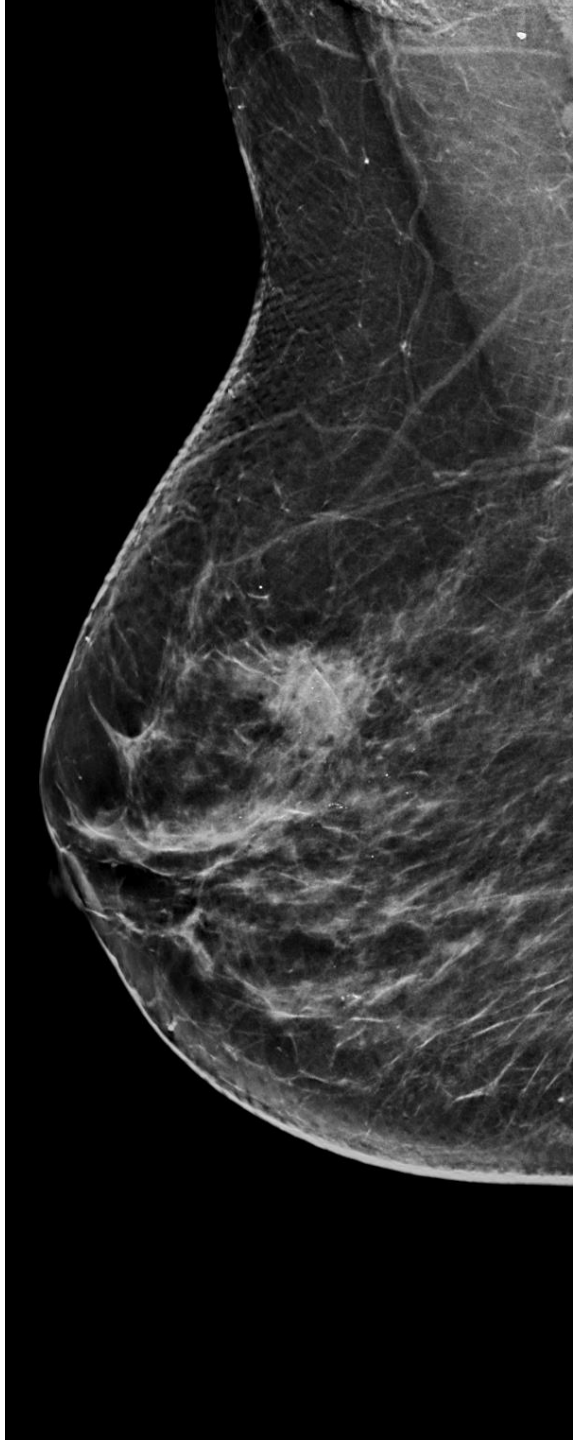
Implementing ML in Lung CT Screening: Challenges + Hurdles

October 2020

A. Gregory Sorensen MD
DeepHealth, Inc.

Disclosures: Listed on LinkedIn

DeepHealth
RadNet AI Solutions



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)

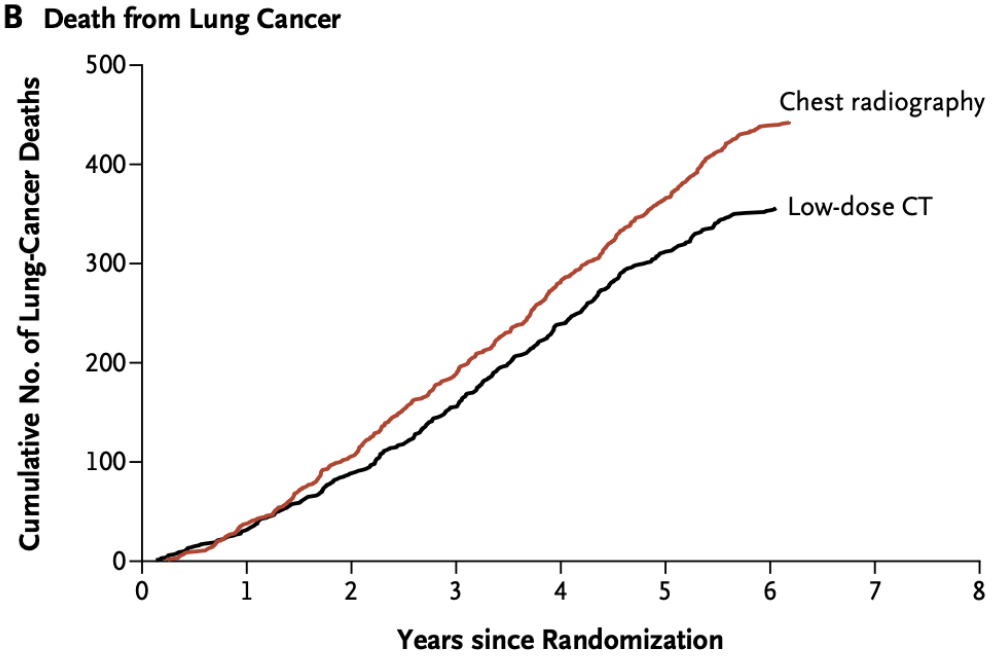
The good news: Screening works

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 4, 2011 VOL. 365 NO. 5

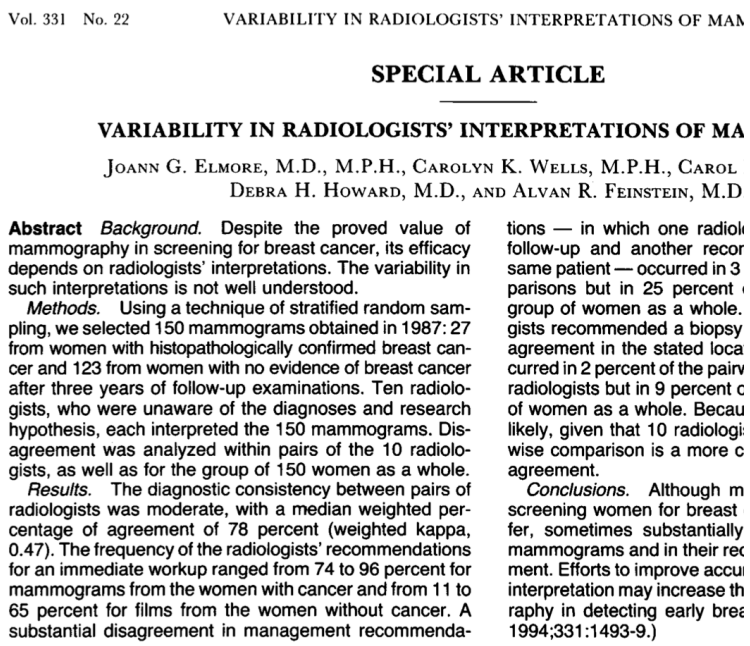
Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening

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



Vol. 331 No. 22 VARIABILITY IN RADIOLOGISTS' INTERPRETATIONS OF MAM

SPECIAL ARTICLE

VARIABILITY IN RADIOLOGISTS' INTERPRETATIONS OF MA

JOANN G. ELMORE, M.D., M.P.H., CAROLYN K. WELLS, M.P.H., CAROL
DEBRA H. HOWARD, M.D., AND ALVAN R. FEINSTEIN, M.D

Abstract *Background.* Despite the proved value of mammography in screening for breast cancer, its efficacy depends on radiologists' interpretations. The variability in such interpretations is not well understood.

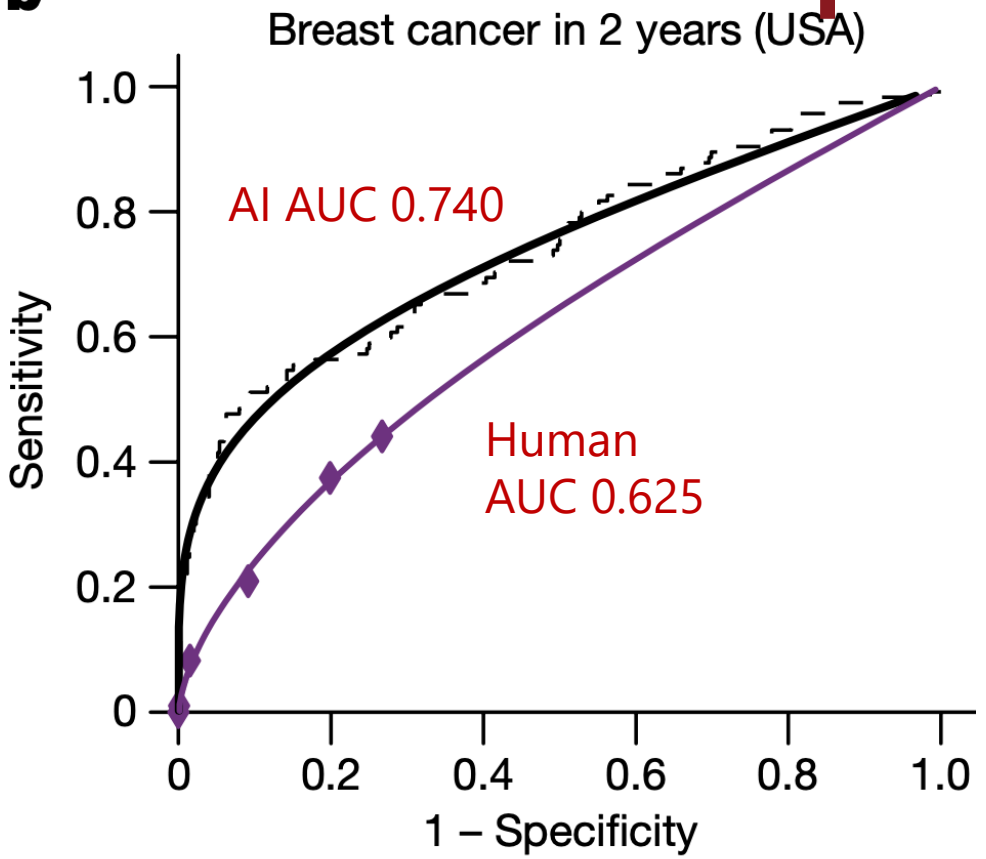
Methods. Using a technique of stratified random sampling, we selected 150 mammograms obtained in 1987: 27 from women with histopathologically confirmed breast cancer and 123 from women with no evidence of breast cancer after three years of follow-up examinations. Ten radiologists, who were unaware of the diagnoses and research hypothesis, each interpreted the 150 mammograms. Disagreement was analyzed within pairs of the 10 radiologists, as well as for the group of 150 women as a whole.

Results. The diagnostic consistency between pairs of radiologists was moderate, with a median weighted percentage of agreement of 78 percent (weighted kappa, 0.47). The frequency of the radiologists' recommendations for an immediate workup ranged from 74 to 96 percent for mammograms from the women with cancer and from 11 to 65 percent for films from the women without cancer. A substantial disagreement in management recommenda-

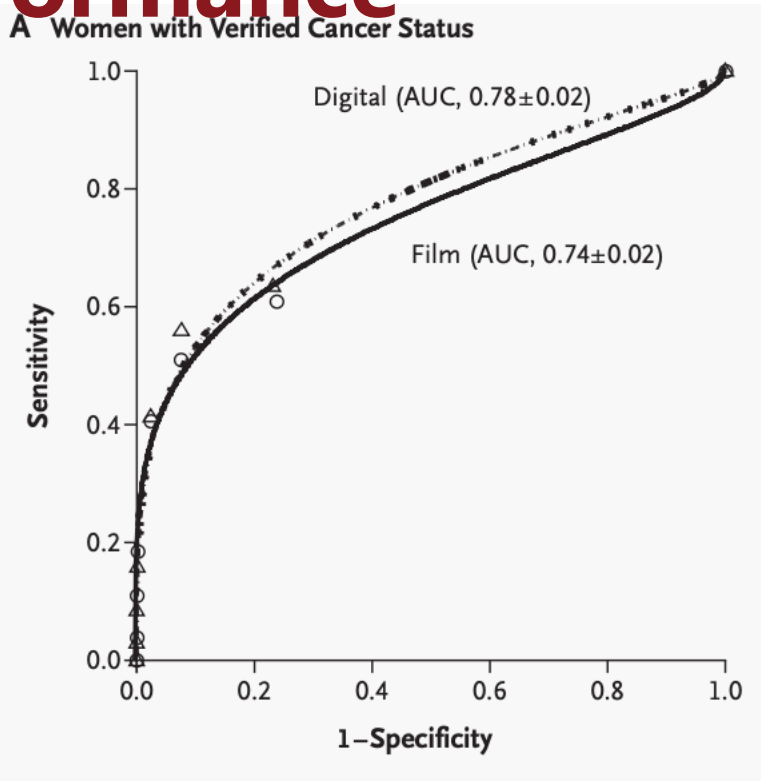
tions — in which one radiol
follow-up and another recor
same patient — occurred in 3
parisons but in 25 percent
group of women as a whole.
gists recommended a biopsy
agreement in the stated loca
curred in 2 percent of the pairv
radiologists but in 9 percent c
of women as a whole. Becau
likely, given that 10 radiologi
wise comparison is a more c
agreement.

Conclusions. Although m
screening women for breast
fer, sometimes substantially
mammograms and in their rec
ment. Efforts to improve accu
interpretation may increase th
raphy in detecting early brea
1994;331:1493-9.)

NEJM 1994:
10 expert readers, kappa = 0.47



Nature 2020:
AI versus Human performance



DMIST NEJM 2005:
also shows human performance no better than 70% sens/spec

In breast cancer screening, challenges to measure and describe human performance

Table 3

Performance Measures for 1 682 504 Screening Digital Mammography Examinations from 2007 to 2013

Performance Measure	1996–2005	2004–2008	2007–2013*	NMD 2008–2012†
AIR (recall rate) (%)	10.9	10.0	11.6 (11.5, 11.6)	10.0
CDR (per 1000 examinations)	4.8	4.3	5.1 (5.0, 5.2)	3.43
Sensitivity (%)	78.7	84.9	86.9 (86.3, 87.6)	NA
Specificity (%)	89.5	90.3	88.9 (88.8, 88.9)	NA
FNR (per 1000 examinations)			0.8 (0.7, 0.8)	NA

Radiology: Volume 283: Number 1—April 2017

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

Measure	Low Performance Range	Percentage of the BCSC Radiologists in Low Performance Range
Sensitivity	<75	18.0%
Specificity	<88 or >95	47.7%
Recall rate	<5 or >12	49.1%
PPV_1	<3 or >8	38.4%
PPV_2	<20 or >40	34.0%
Cancer detection rate	<2.5/1,000	28.4%

NOTE: BCSC = Breast Cancer Surveillance Consortium; PPV = positive predictive value.

- 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+)

In breast cancer screening, challenges to measure and describe human performance

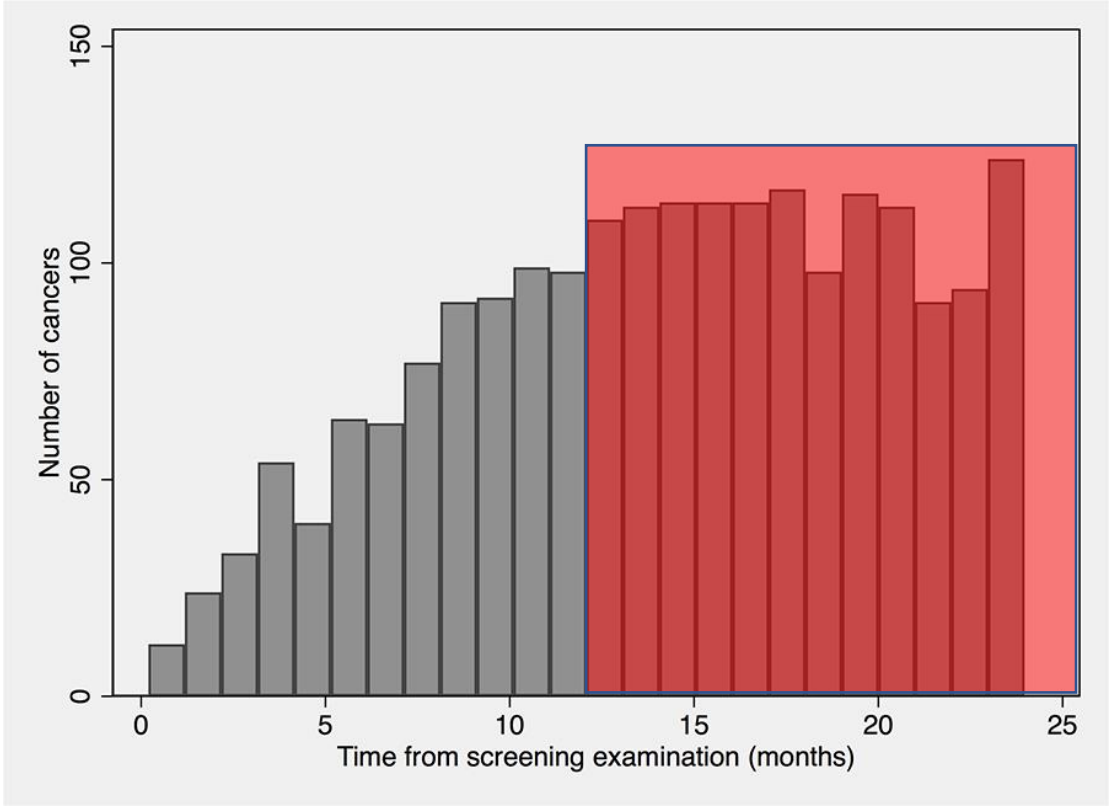
Radiology

ORIGINAL RESEARCH · BREAST IMAGING

Range of Radiologist Performance in a Population-based Screening Cohort of 1 Million Digital Mammography Examinations

Parameter	All Interpreting Radiologists (n = 110)
No. of examinations	1 186 045
Sensitivity (%)	73
Specificity (%)	96
AIR	39
CDR	3.0
FNR (%)	29
Accuracy (%)	96
Positive predictive value (%)	8

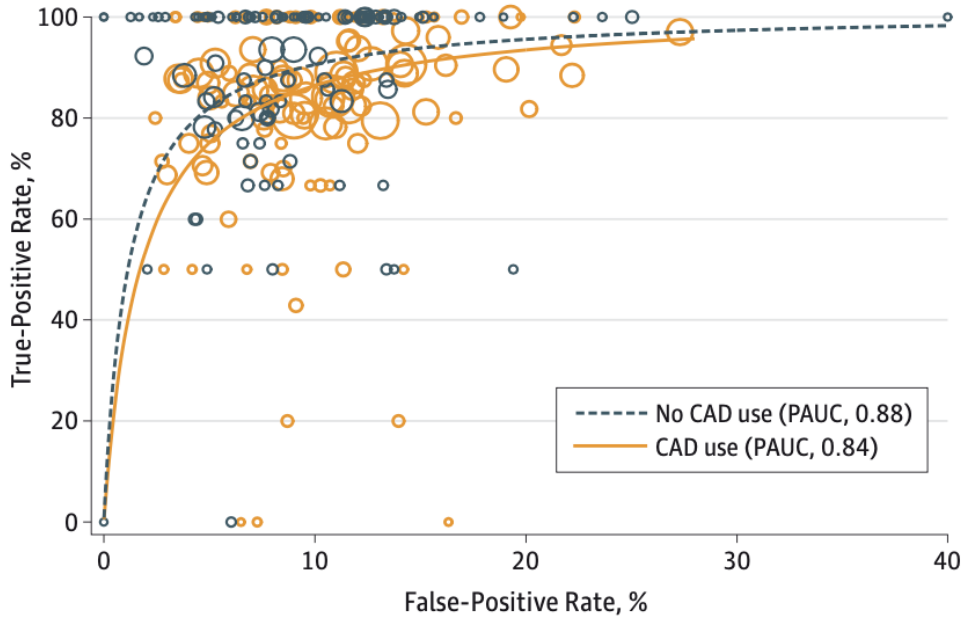
Radiology 2020; 297:33–39



- Do we radiologists not want to report our performance?

In breast cancer screening, laboratory performance does not always equal "real world"

Figure 2. Receiver Operating Characteristic Curves for Digital Screening Mammography With and Without the Use of CAD, Estimated From 135 Radiologists Who Interpreted at Least 1 Examination Associated With Cancer



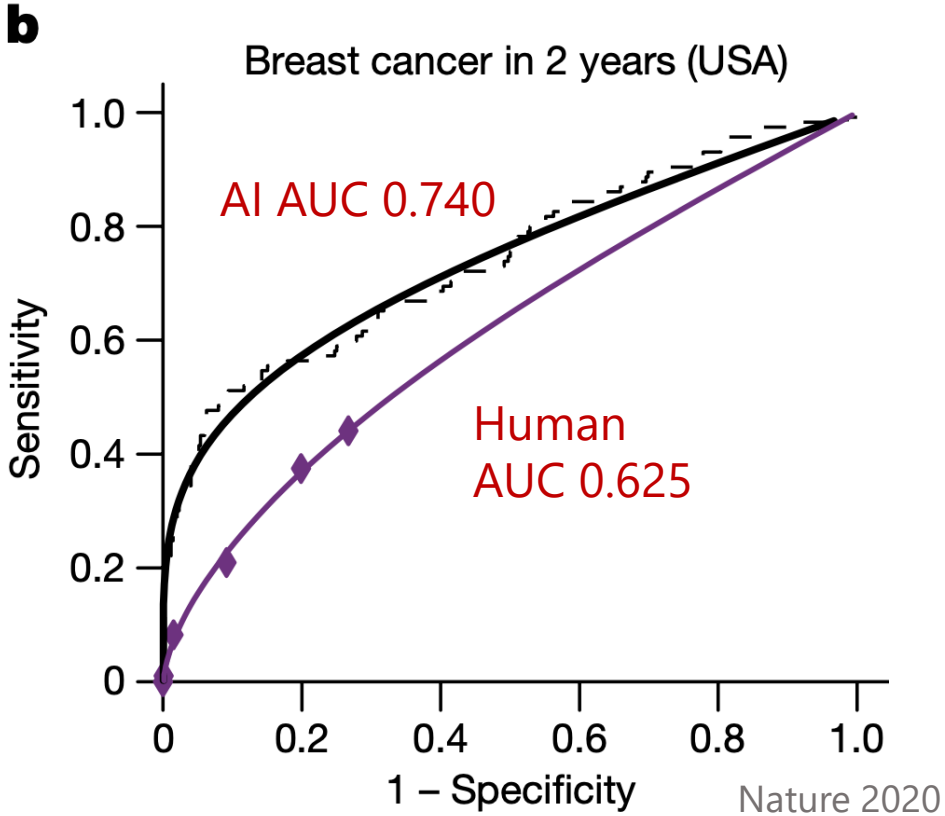
Each circle represents the true-positive or false-positive rate for a single radiologist, for examinations interpreted with (orange) or without (blue) computer-aided detection (CAD). Circle size is proportional to the number of mammograms associated with cancer interpreted by that radiologist with or without CAD. PAUC indicates partial area under the curve.

JAMA Intern Med. 2015;175(11):1828-1837.

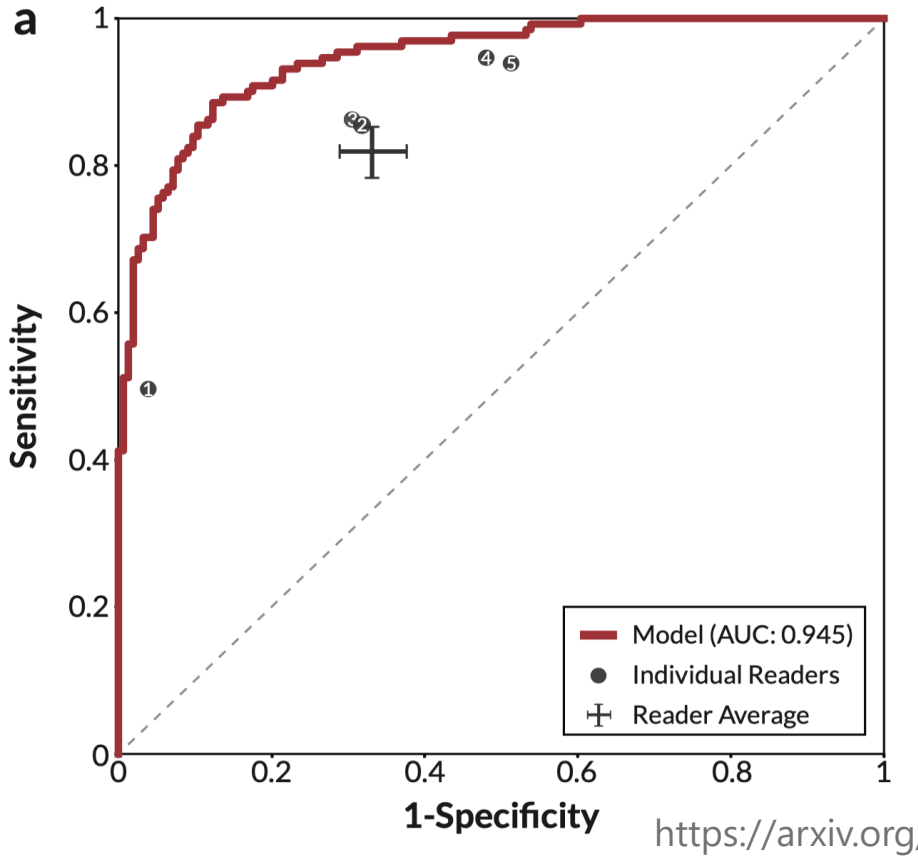
- While CAD had shown benefit in FDA studies, in practice this benefit was not evident
- Many radiologists individually had already had this perception
- Reimbursement, perceived legal risks still drive usage even today

- What went wrong?

Human interpreters are not just “good readers” or “bad readers” and AI should probably reflect that…



Nature 2020



https://arxiv.org/abs/1912.11027

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

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