transpara®

By ScreenPoint Medical

Life after academia: How a physicist can contribute to breast cancer detection

Ruud Peeters

SCREENPOINT Medical

Contents

Background

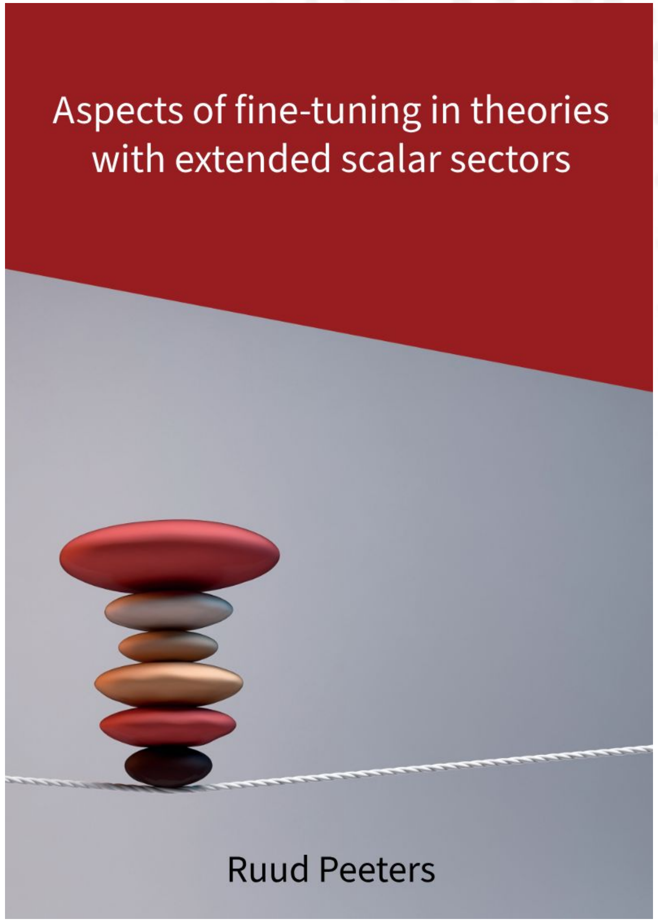
2011 - 2016: Studied physics in Nijmegen

2016 - 2021: PhD in Groningen:

Part of Nikhef program Higgs as a Probe and Portal

2021 - now: Screenpoint Medical in Nijmegen

- AI Research Scientist
- AI Software Engineer



transpara®

By ScreenPoint Medical

Breast cancer and screening

SCREENPOINT
Medical

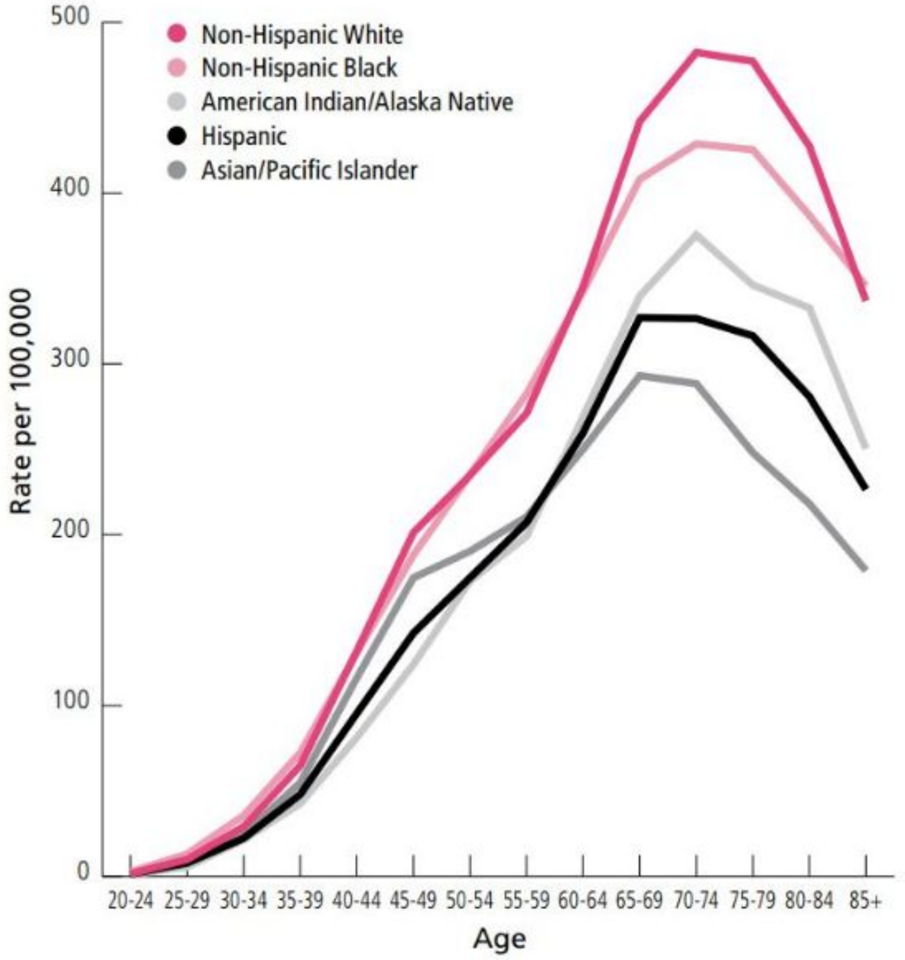
Breast cancer

Most common cancer for women

1 in 8 women will develop breast cancer at some point in their lives

Early detection is critical:

90% average 5Y survival rate for non-metastatic invasive breast cancer



Note: Rates are per 100,000 and age adjusted to the 2000 US standard population.

Source: NAACCR, 2019. Data for American Indians/Alaska Natives are based on Purchased/Referred Care Delivery Area (PRCDA) counties.

Breast cancer screening

Regular examination of healthy population

Screening is common in many countries nowadays

- 75M examinations worldwide per year
- Protocols vary by region
    - Age group (~50-75 years)
    - Interval (1-3 years)
    - Reading (Europe 2 radiologists, US 1 radiologist)

Breast cancer screening

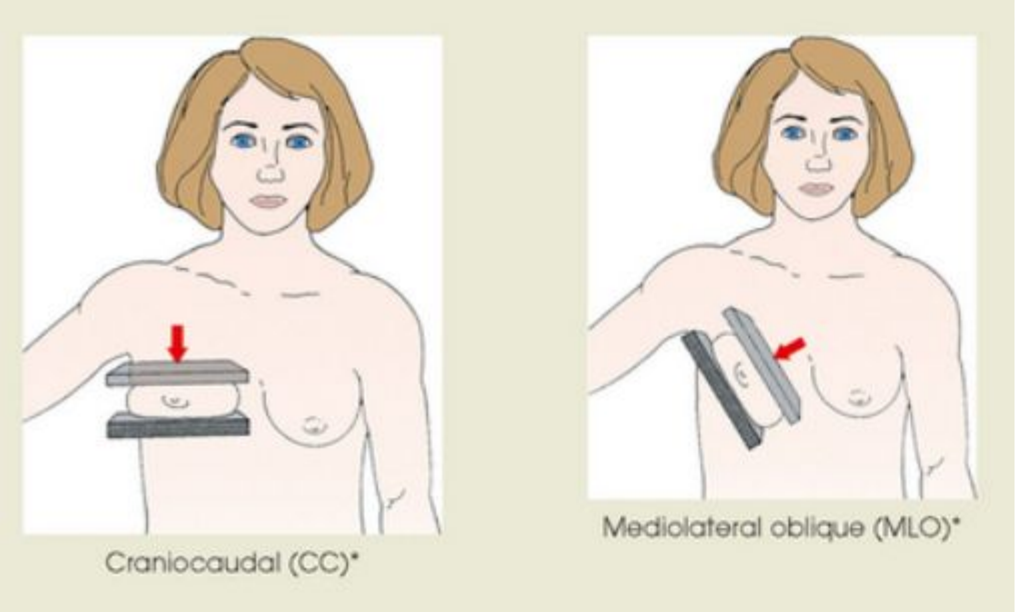
Suspicious finding: woman is recalled for further examination

- In the Netherlands: recall rate 2.35%
- Cancer incidence in screening populations about 0.5%

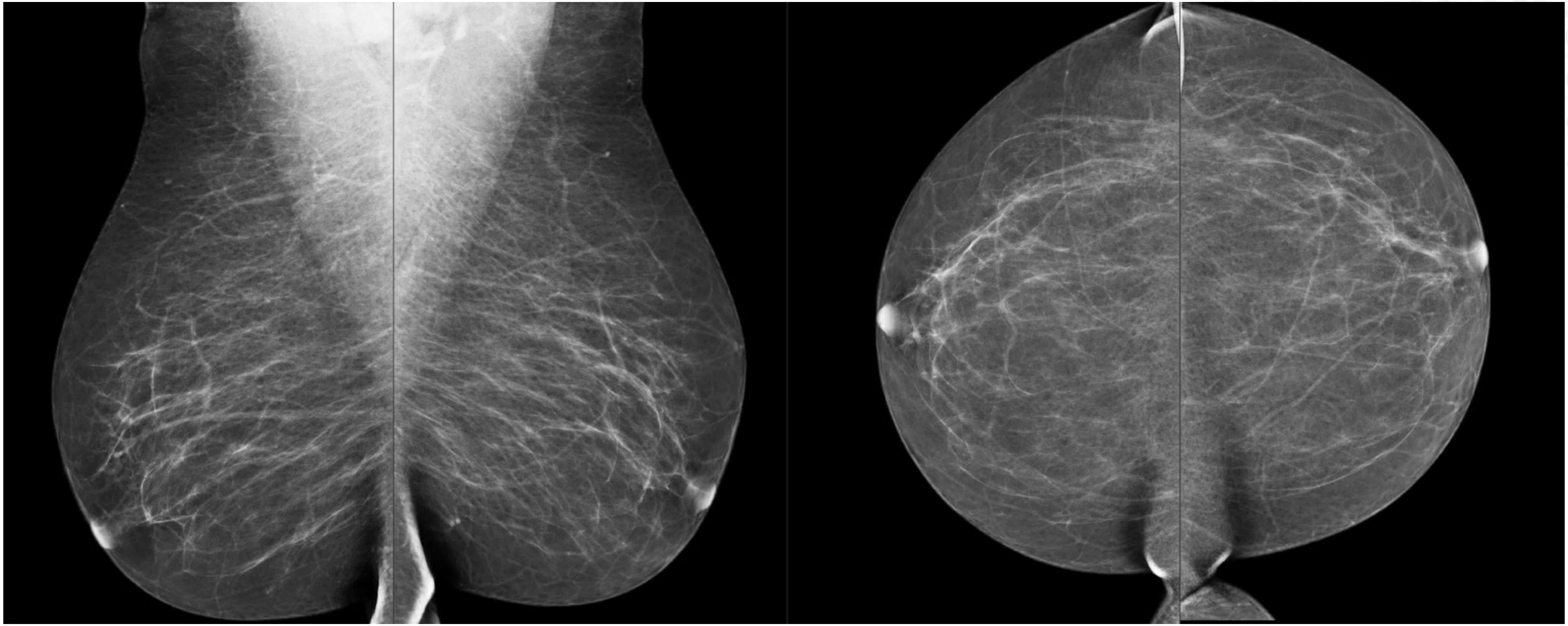
Screening in practice

Mammography: X-ray images of breast tissue

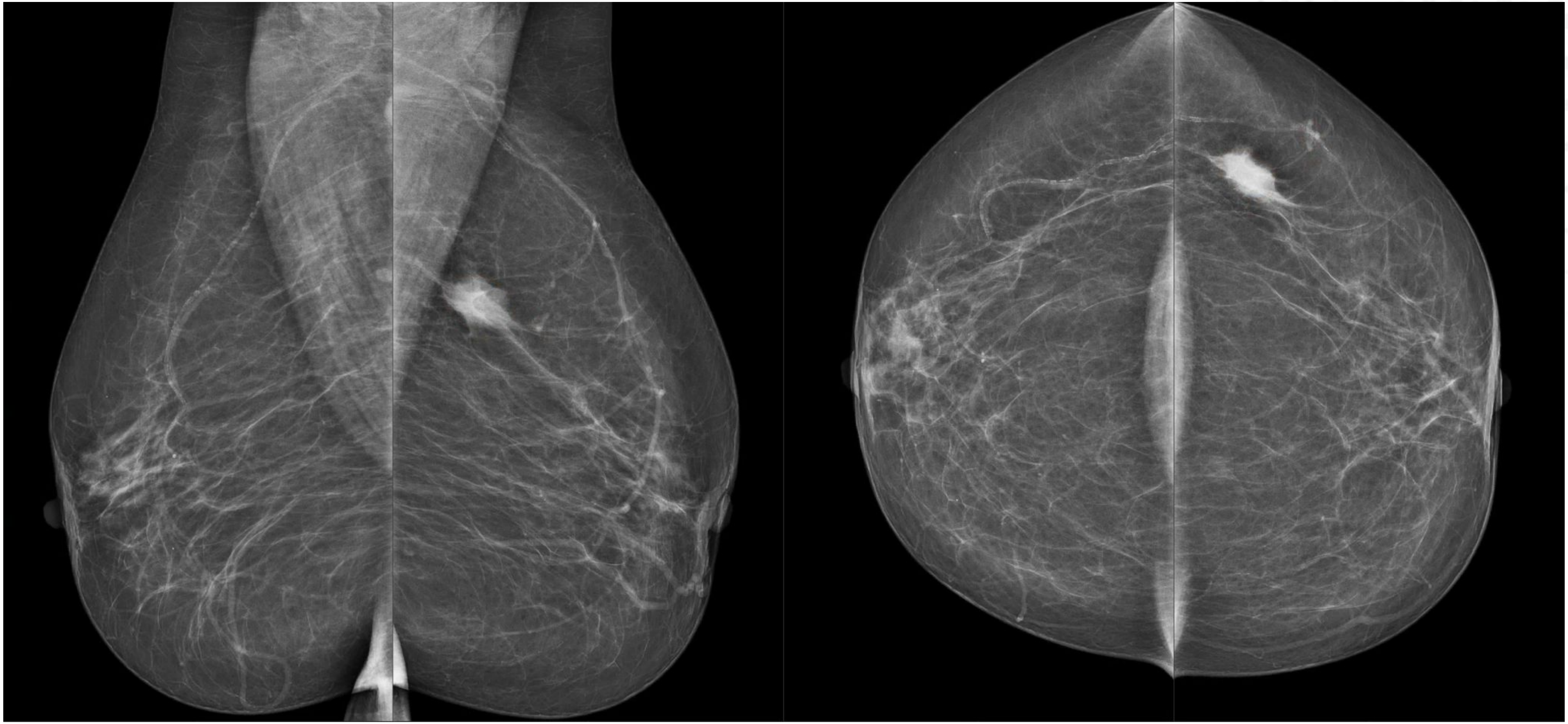
Two projections of each breast: CC/MLO



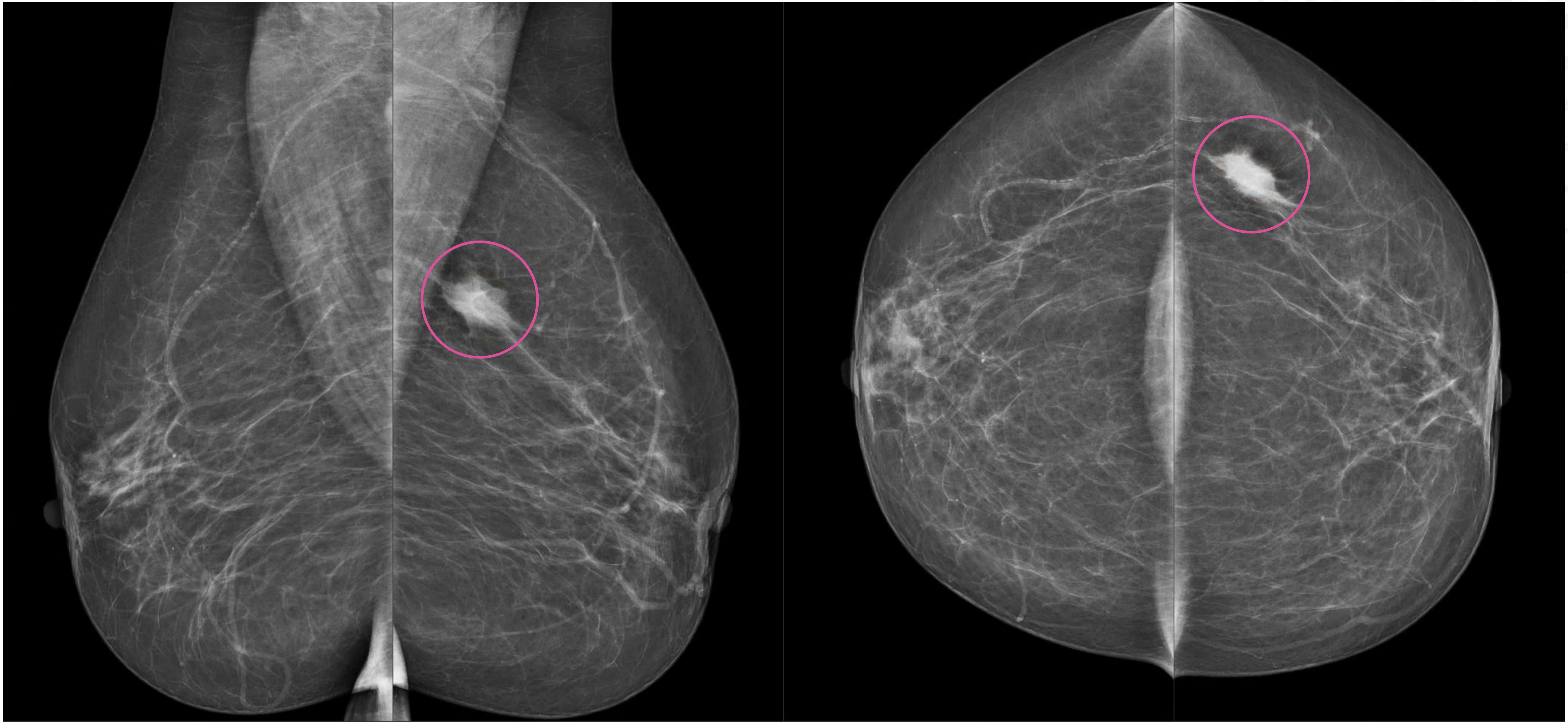
Right MLO Left MLO Right CC Left CC

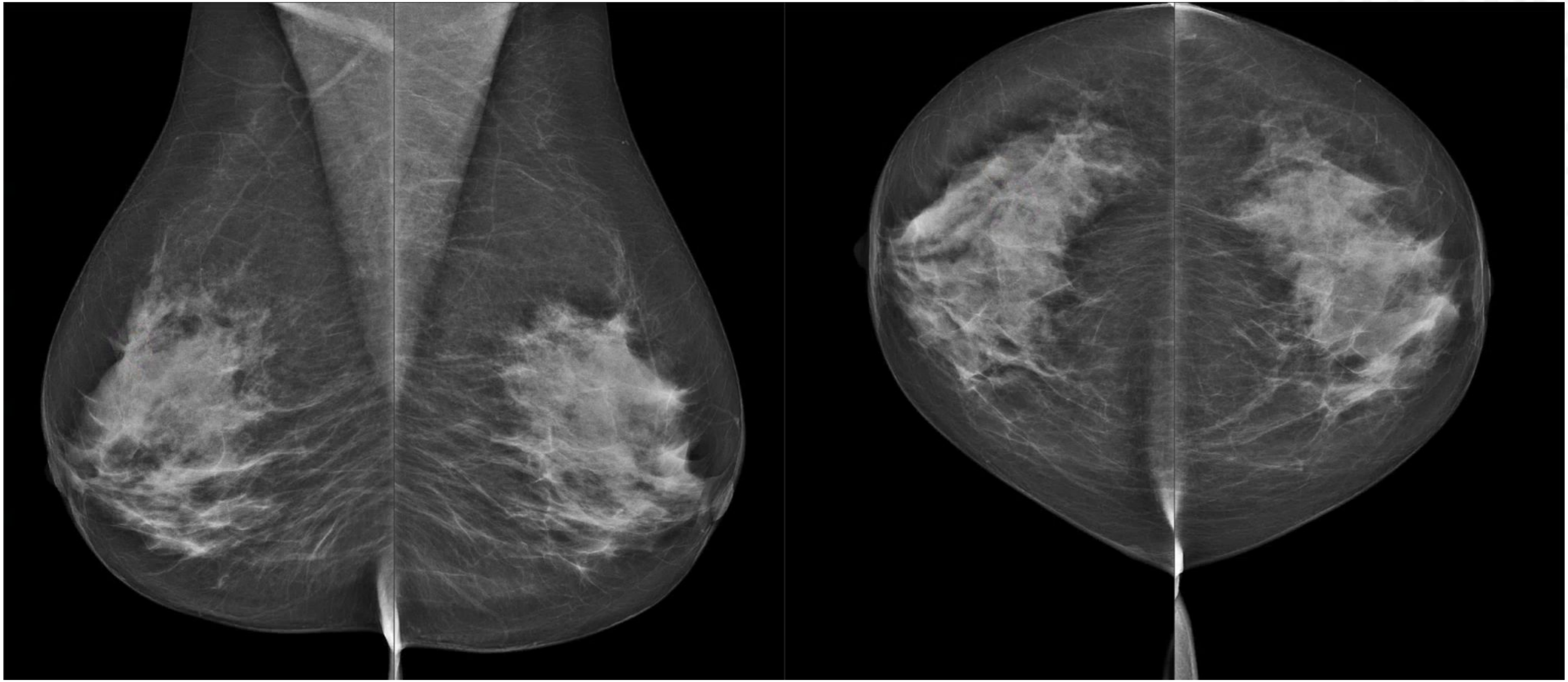


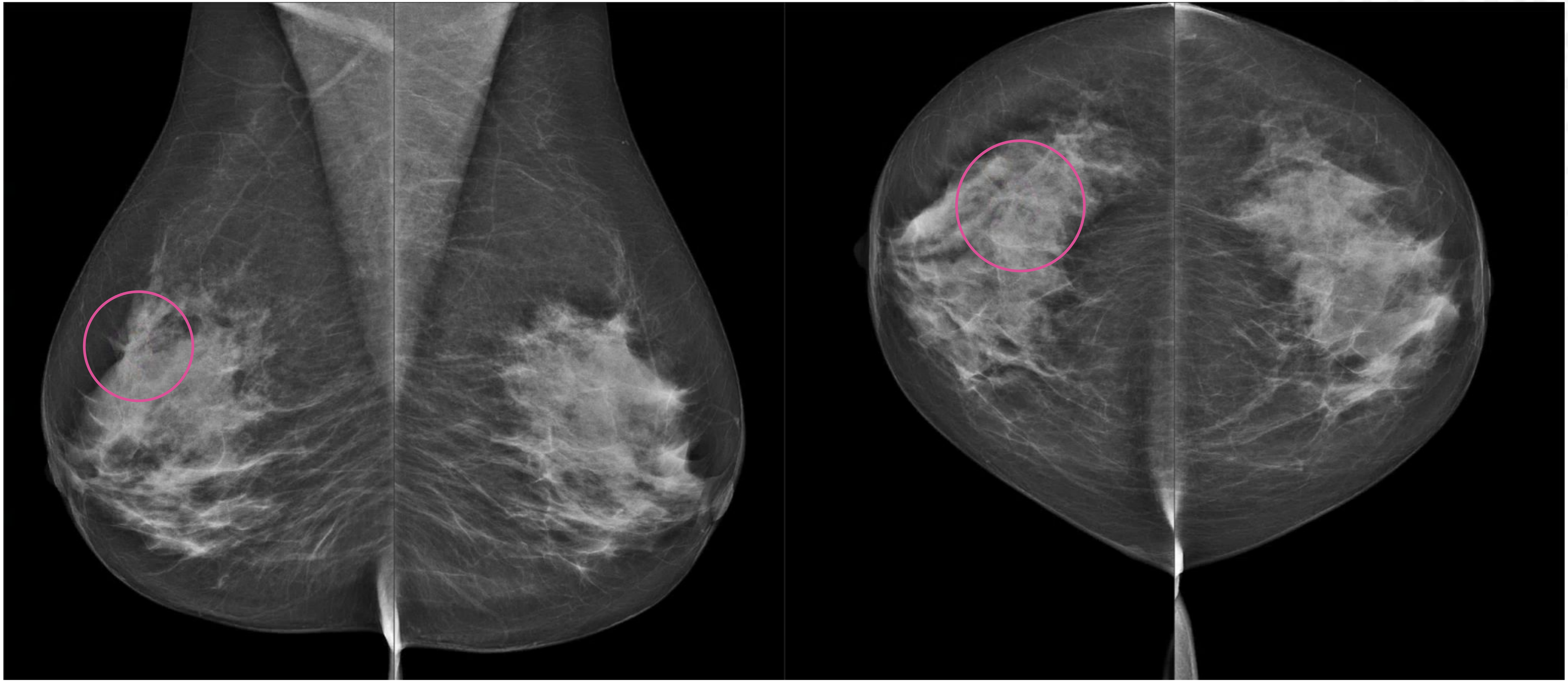
Right MLO Left MLO Right CC Left CC



Right MLO Left MLO Right CC Left CC







Challenges of screening

Difficult work: 5/1000 exams have a cancer

United States: liability (recall rate ~10-15%)

Shortage of radiologists

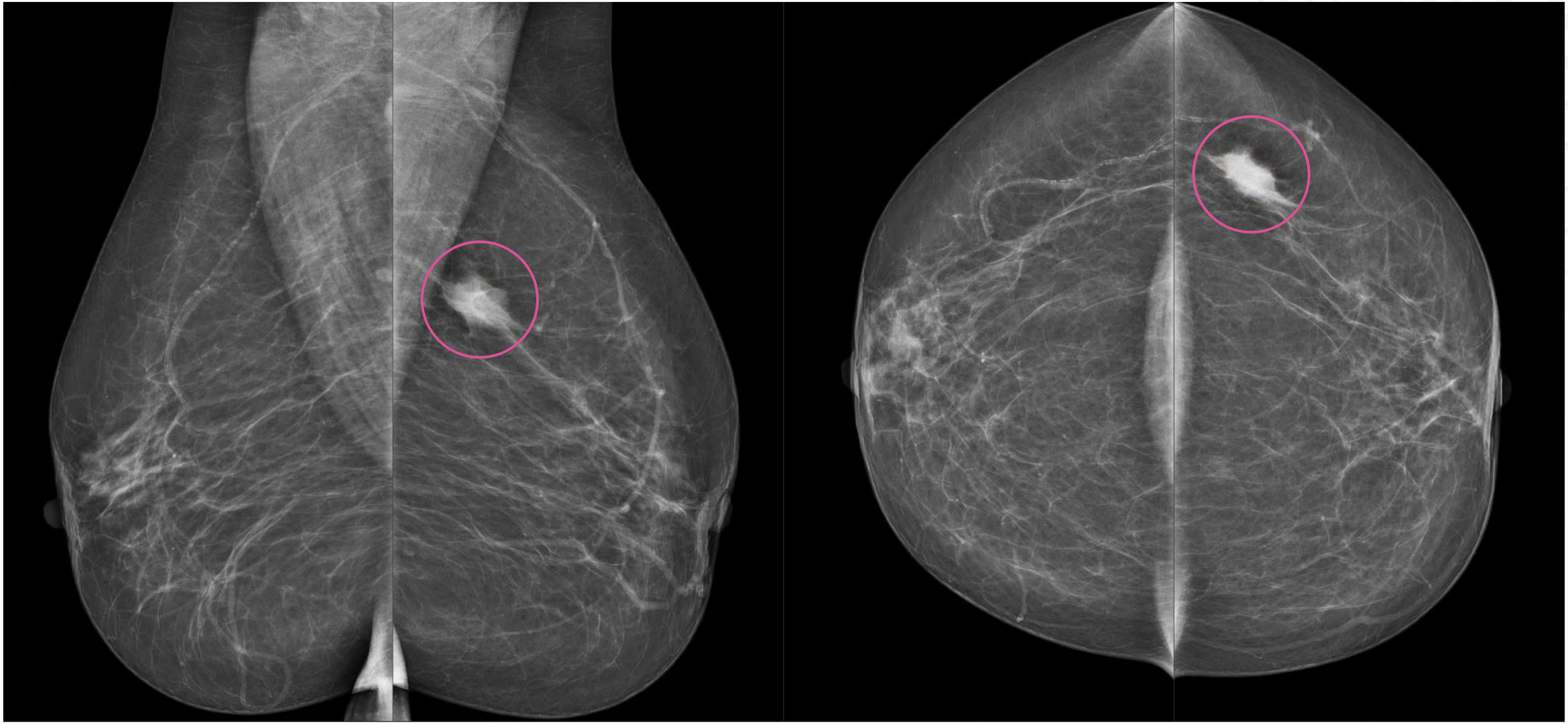
transpara®

By ScreenPoint Medical

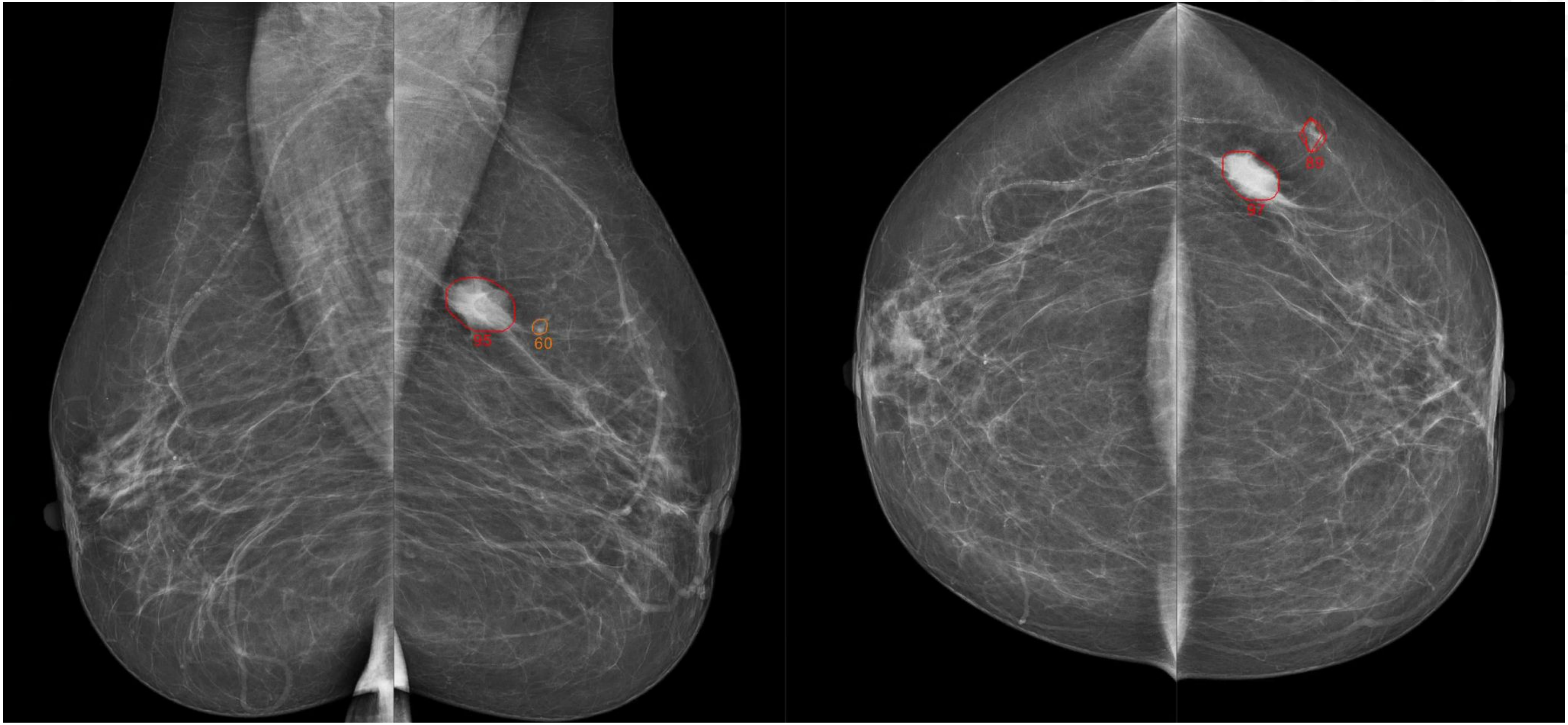
AI in breast cancer detection

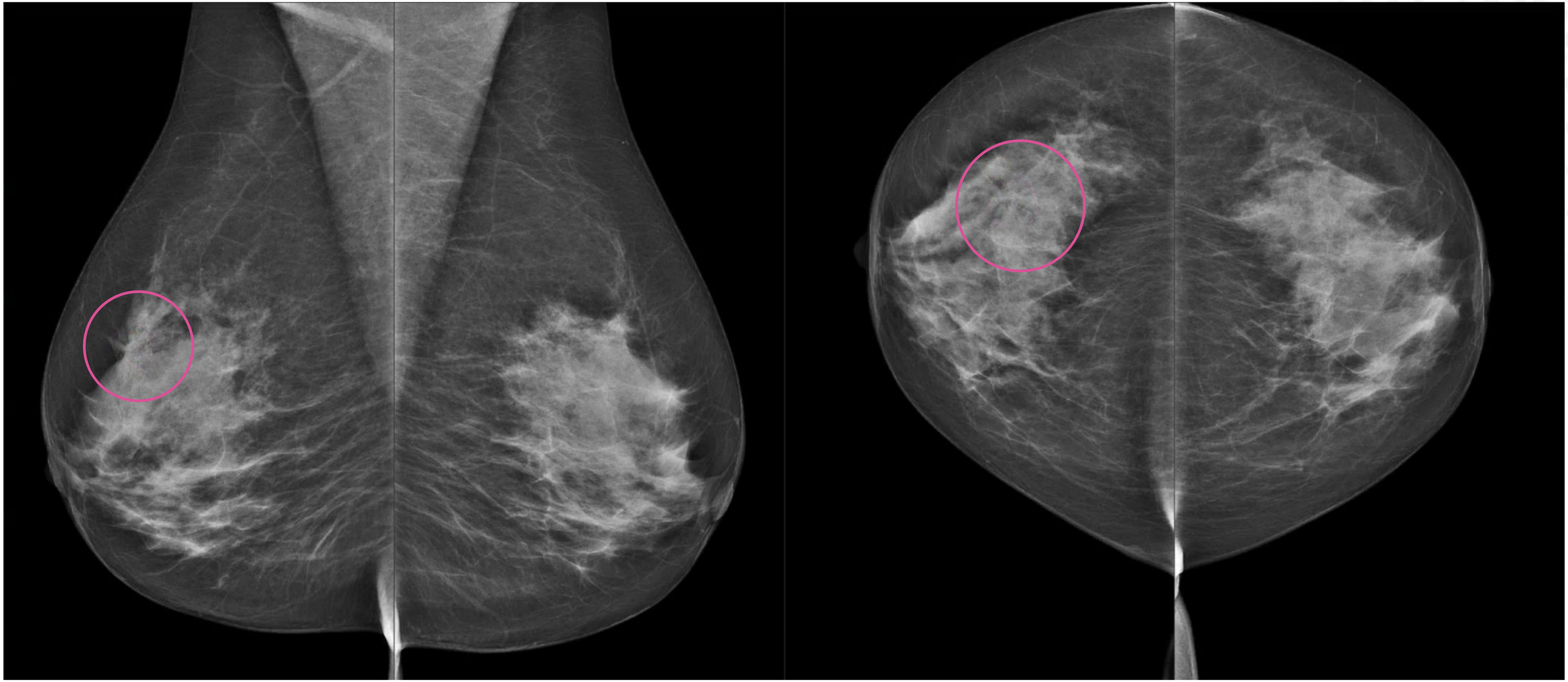
SCREENPOINT
Medical

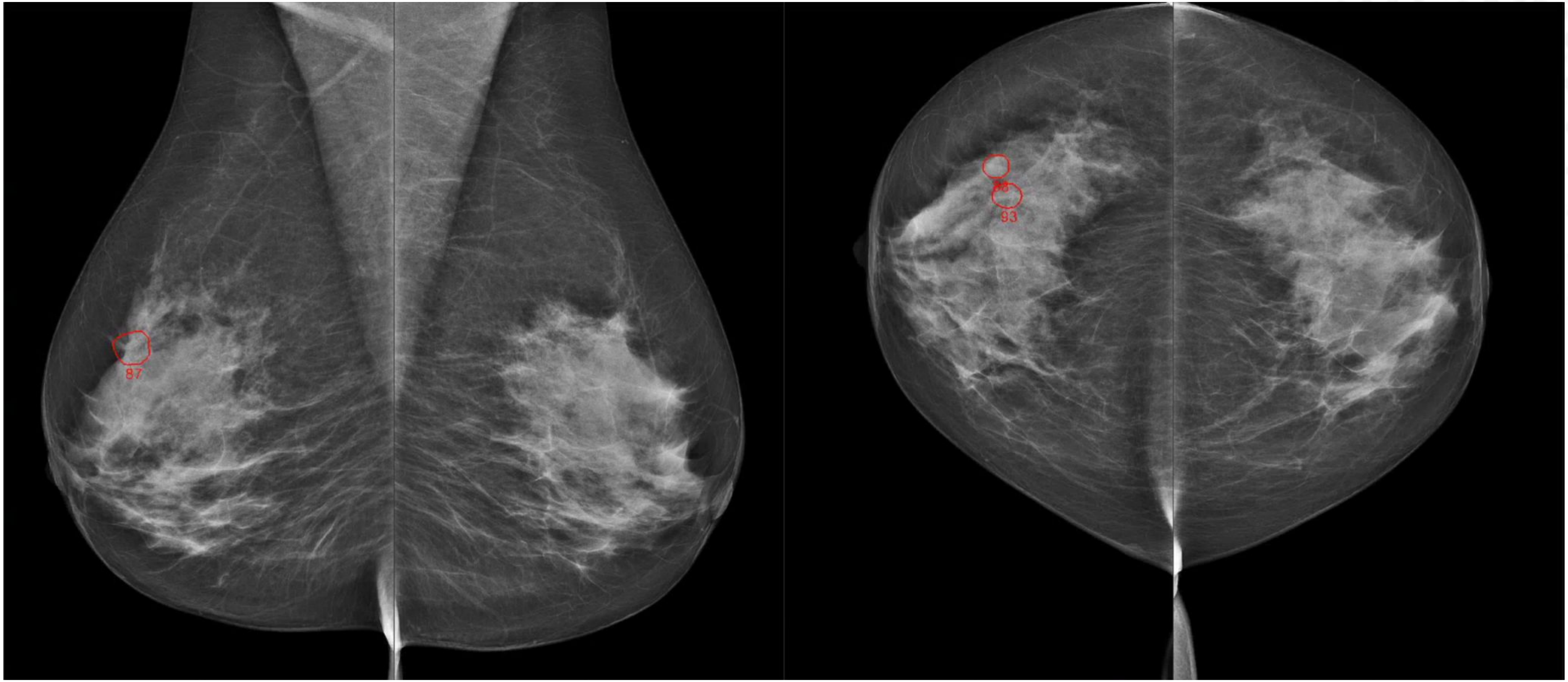
Right MLO Left MLO Right CC Left CC

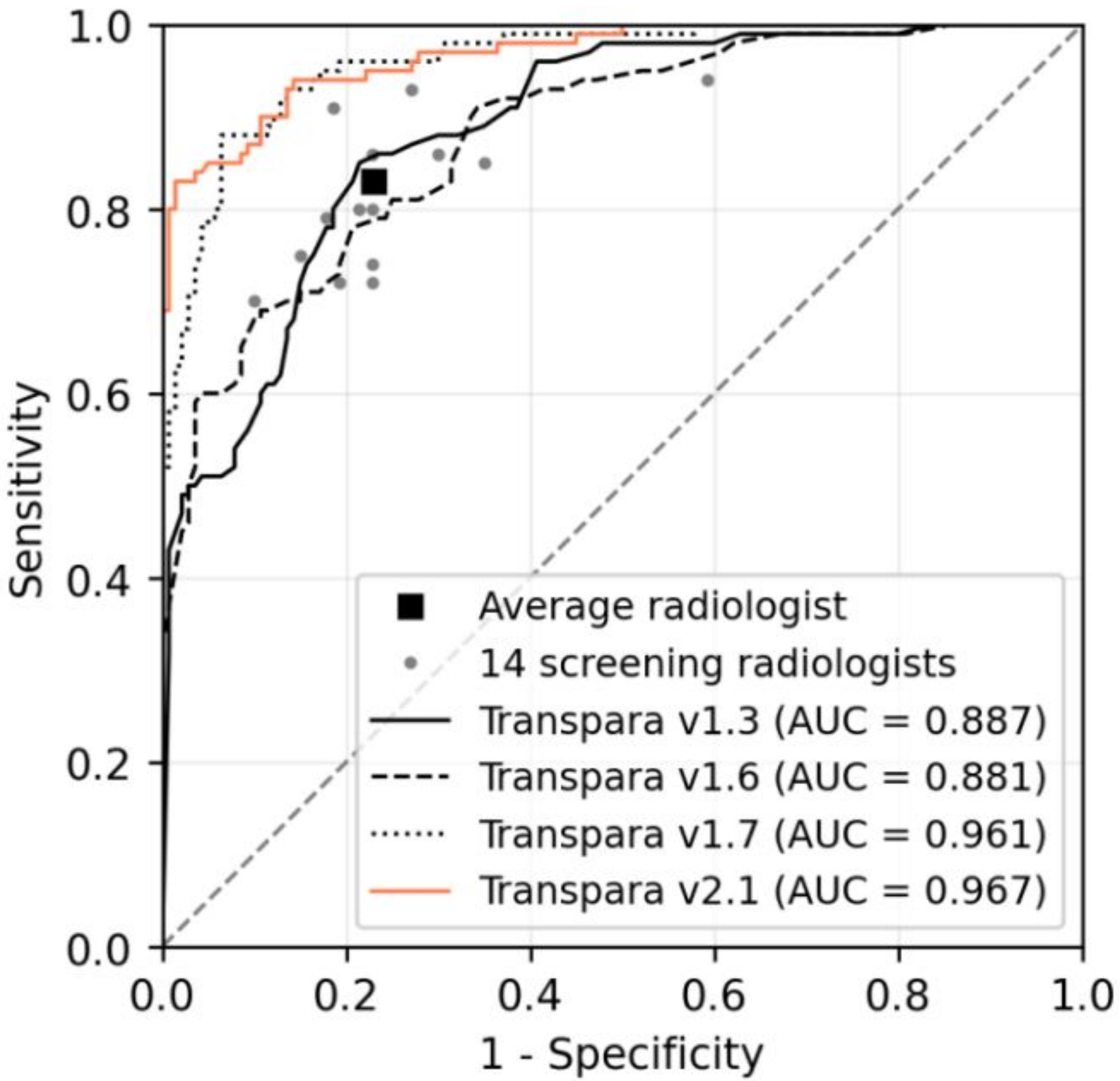


Right MLO Left MLO Right CC Left CC






87
93


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Average radiologist
14 screening radiologists
Transpara v1.3 (AUC = 0.887)
Transpara v1.6 (AUC = 0.881)
Transpara v1.7 (AUC = 0.961)
Transpara v2.1 (AUC = 0.967)

First randomized control trial

MASAI trial in Sweden

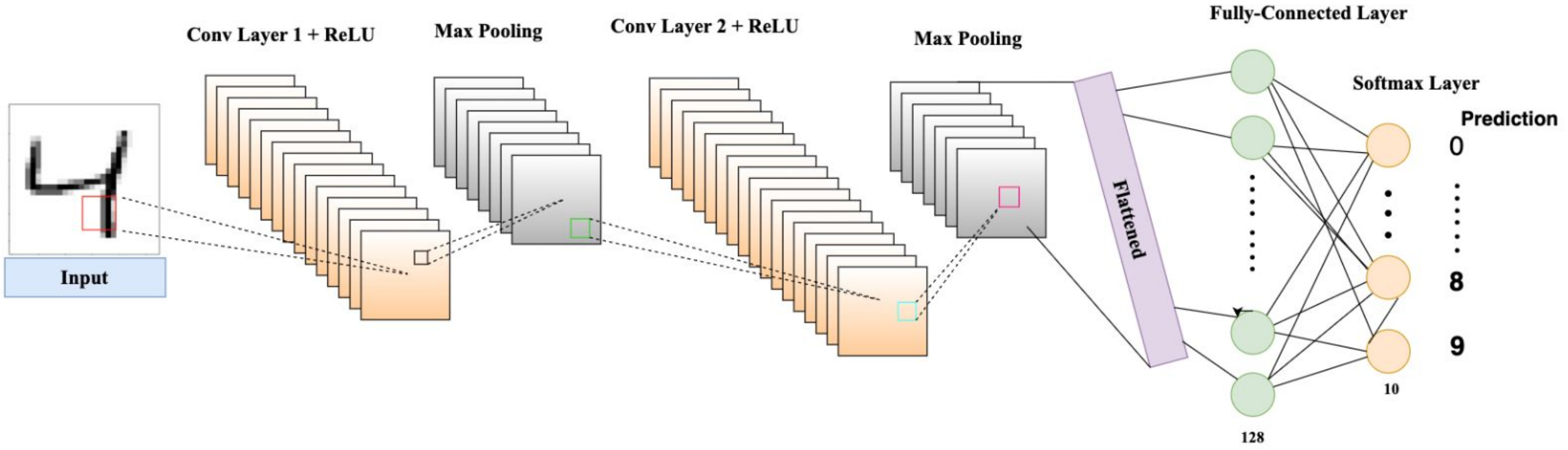
Population randomly divided in two: regular screening vs screening with AI

	Regular screening	Screening with AI
Cancer detection	5.1/1000	6.4/1000
Recall	2.0%	2.2%

Workload reduction: 44%

AI algorithms for breast cancer detection

Convolutional neural network



transpara®

By ScreenPoint Medical

Company details

SCREENPOINT Medical

Company history

Founded in 2014

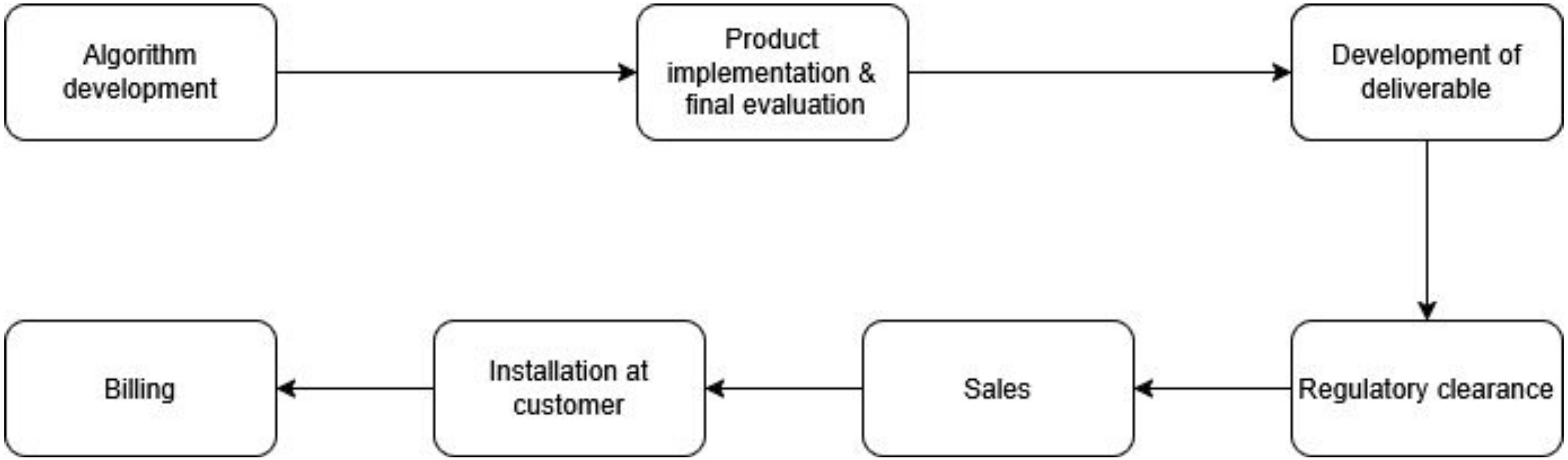
Spin-off from Radboud University Medical Center

55 employees in Nijmegen

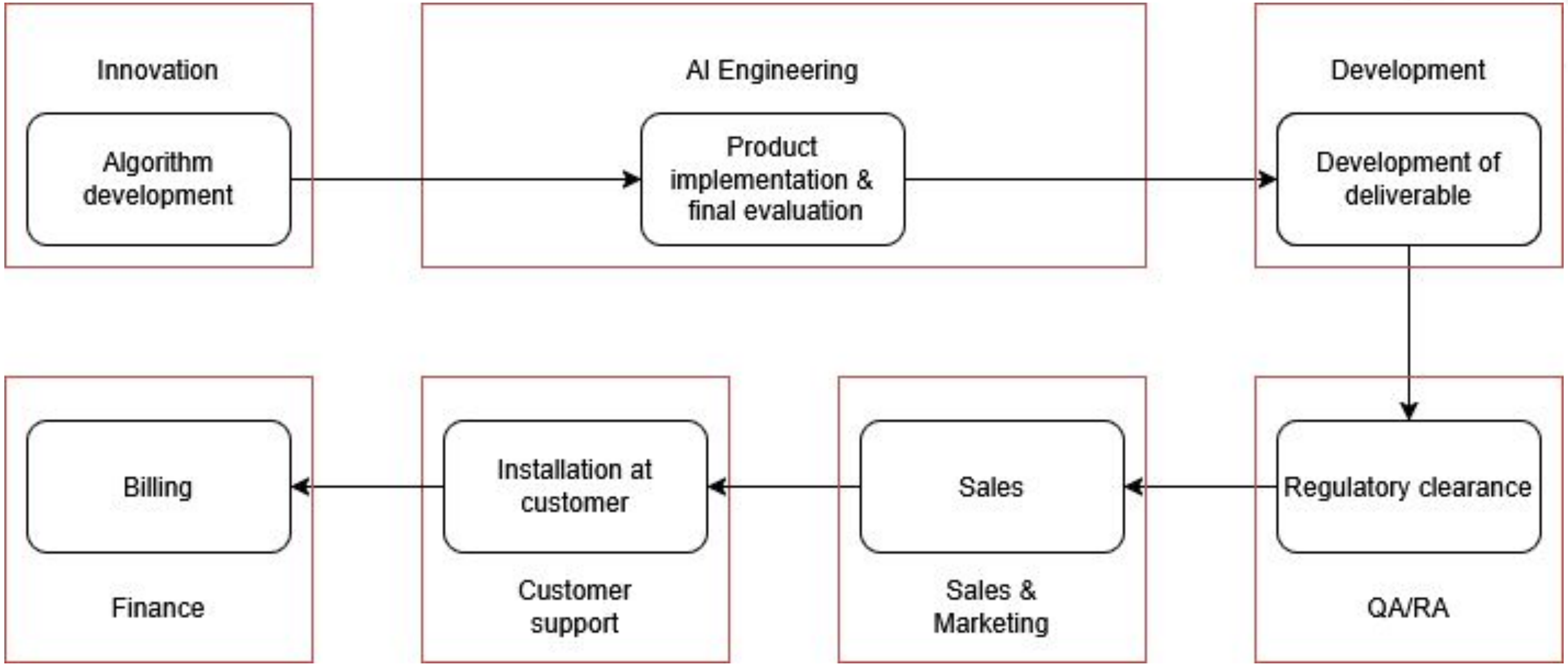
13 US employees

3 interns

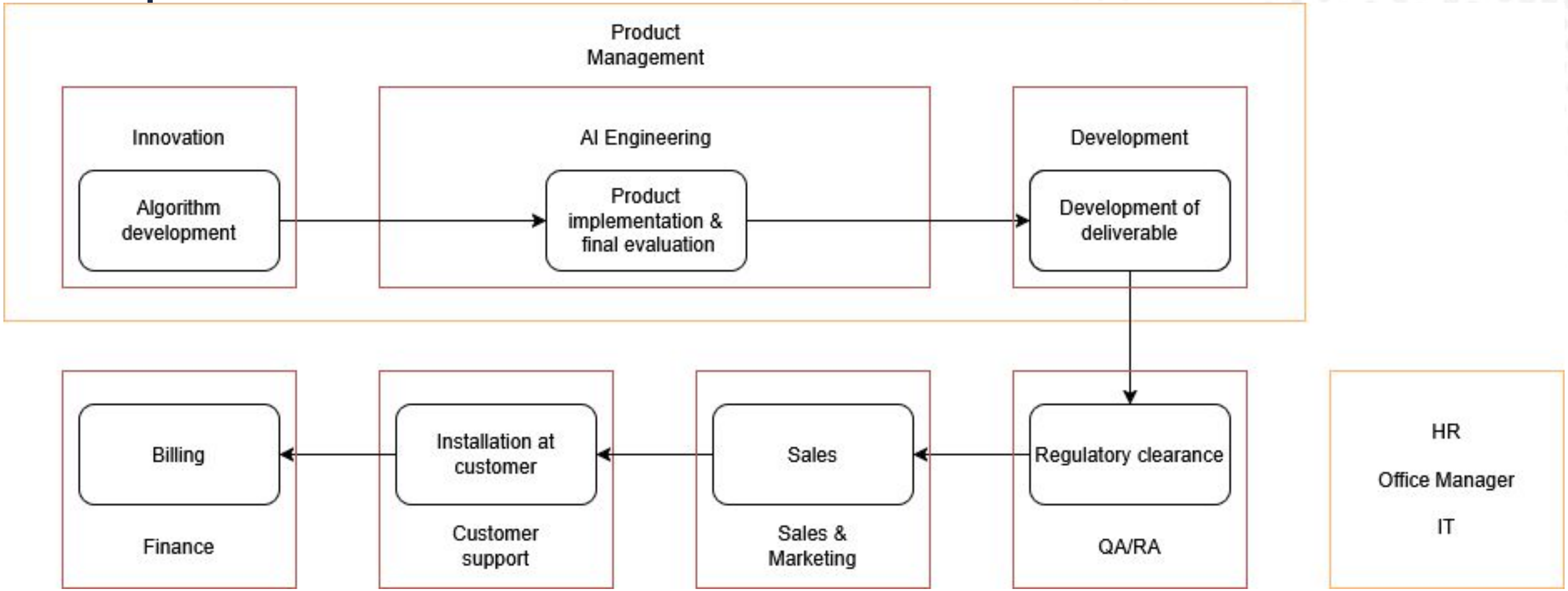
Company structure



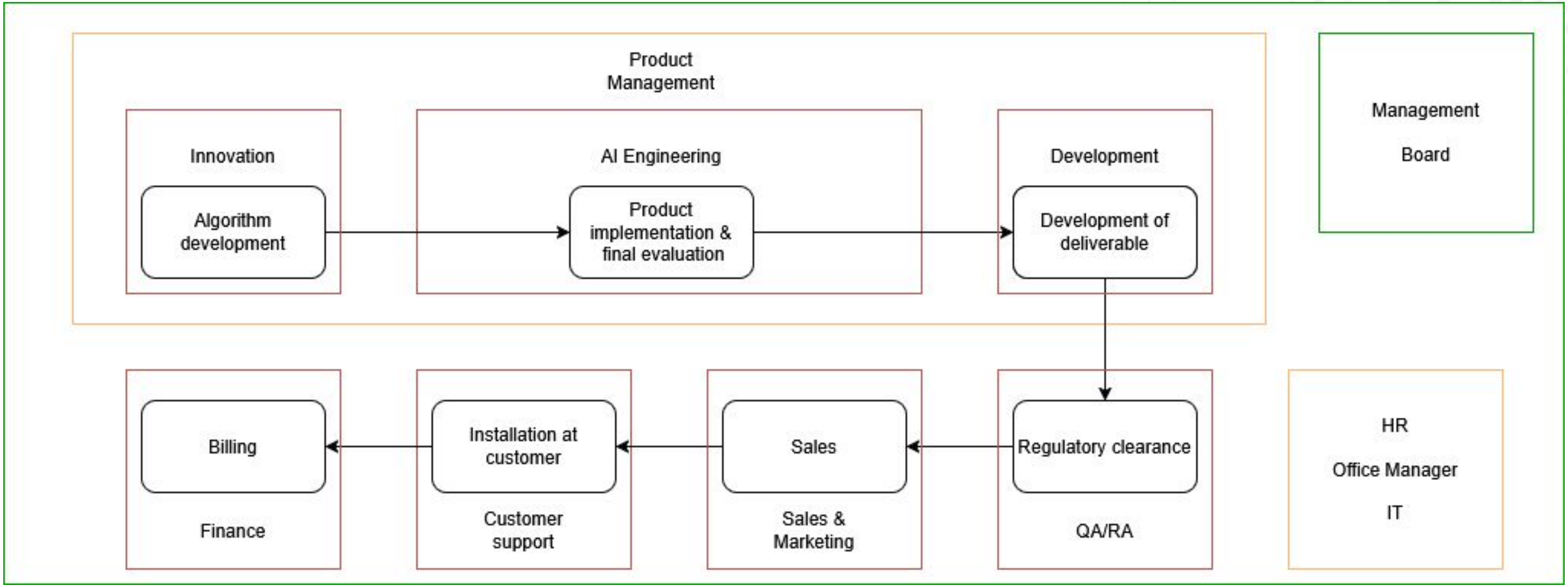
Company structure

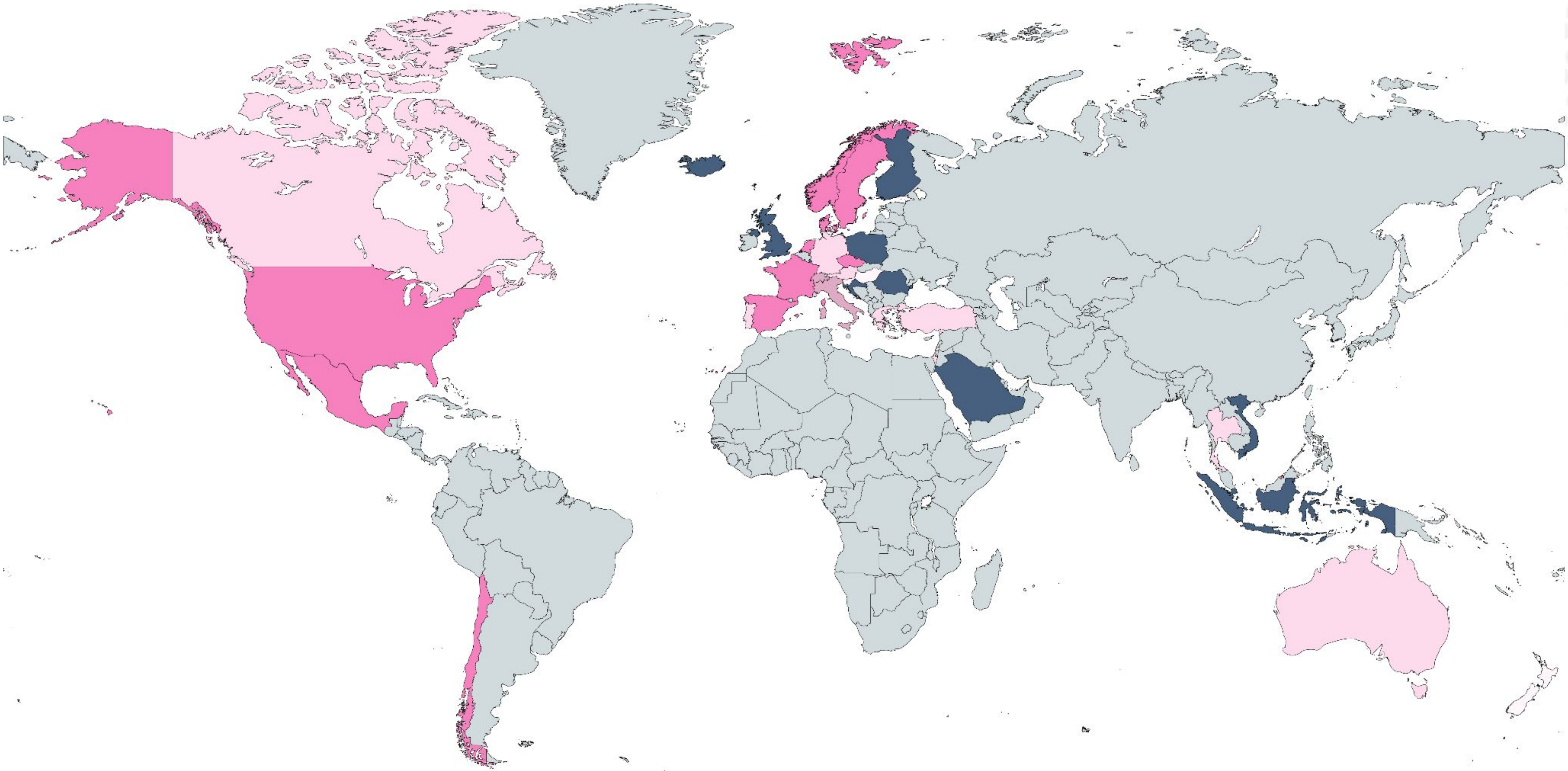


Company structure



Company structure





My team

AI Engineering: team lead + 7 engineers

Responsibilities:

- Product implementation of new features
- Evaluation of the product
- Data
- Calibration
- …

3 physicists (2 HEP)

transpara®

By ScreenPoint Medical

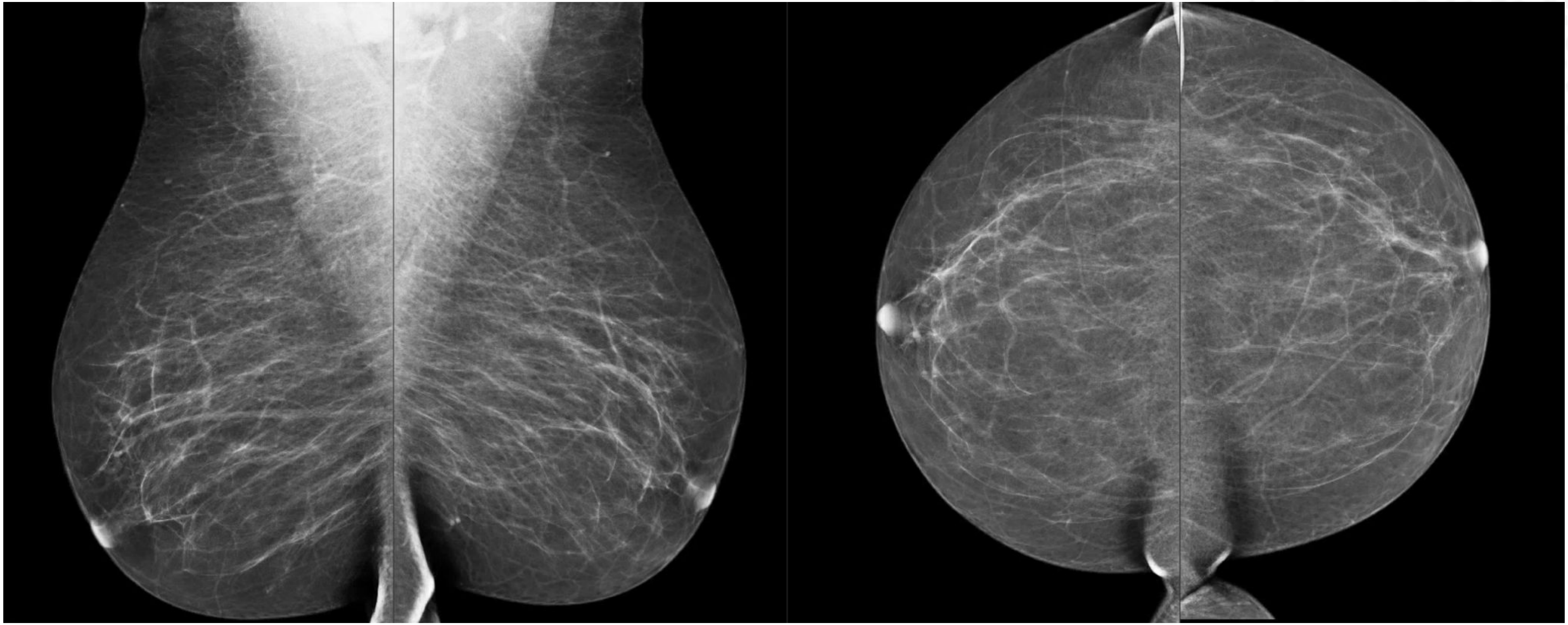
My daily work

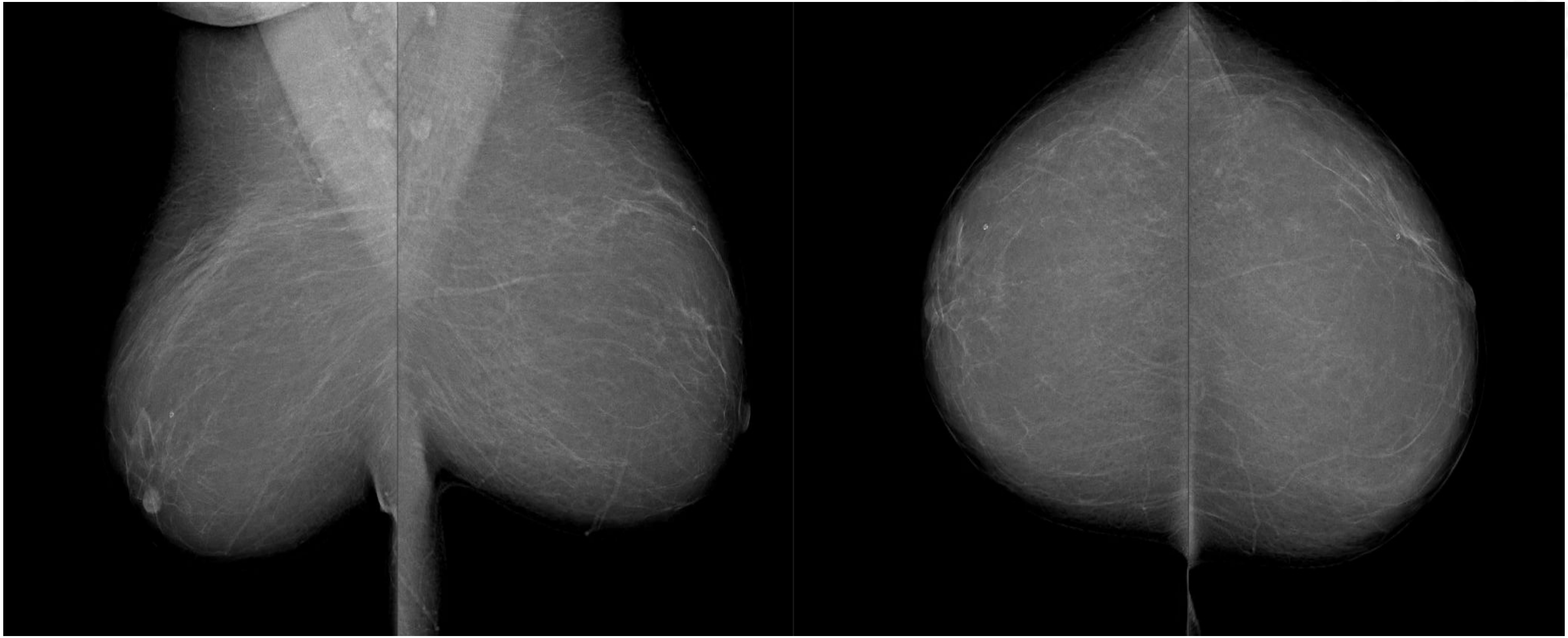
SCREENPOINT Medical

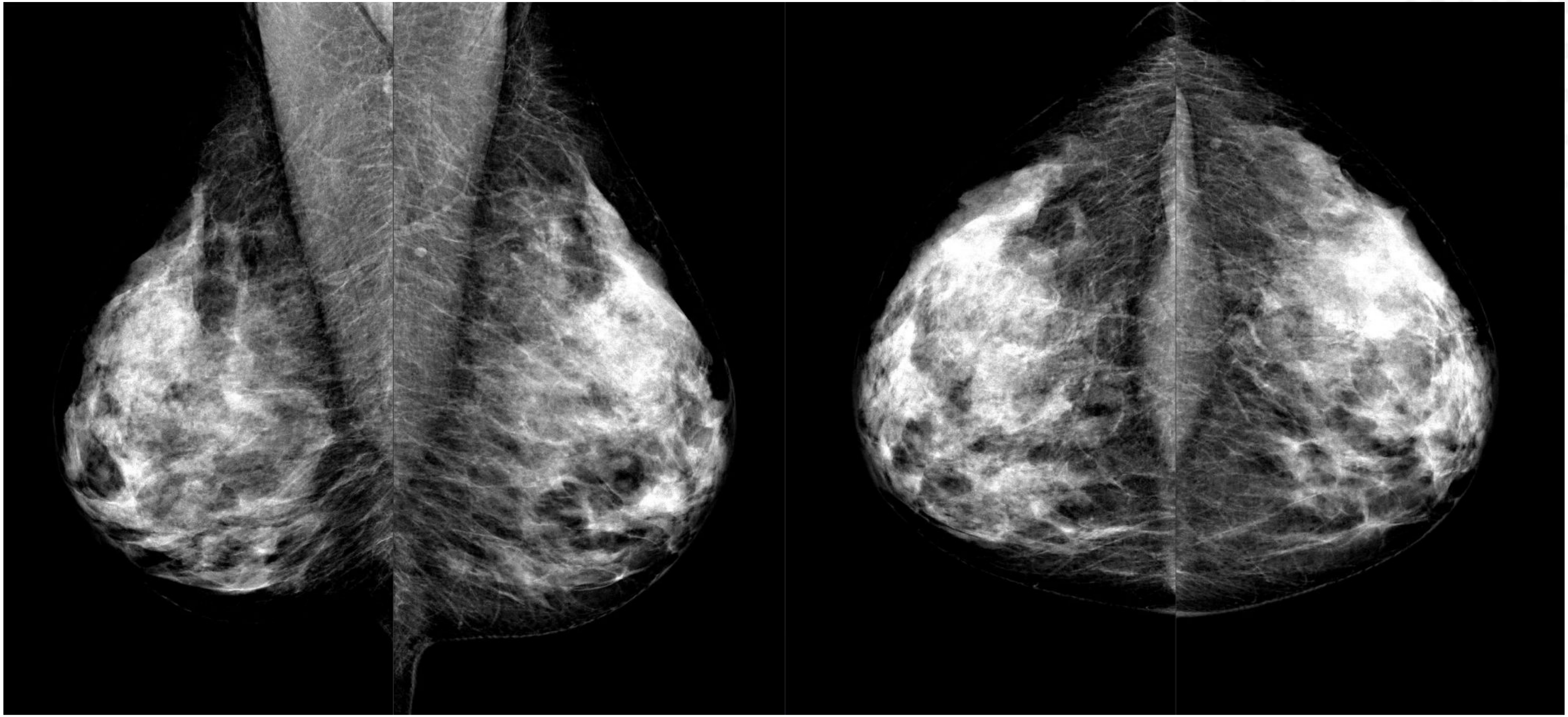
Calibration

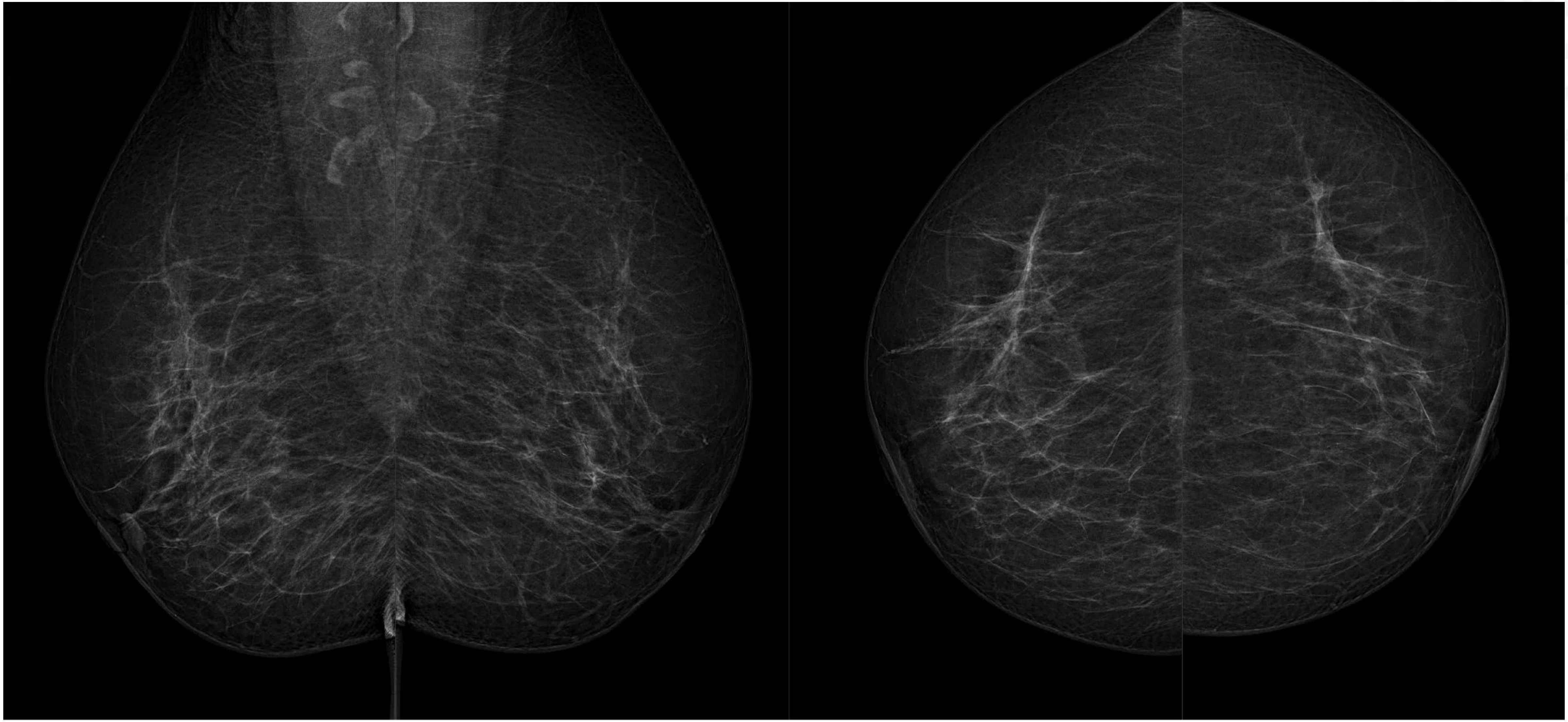
Large variety in data in the field:

- Population differences (age/ethnicity/…)
- Screening vs diagnostics
- Image types









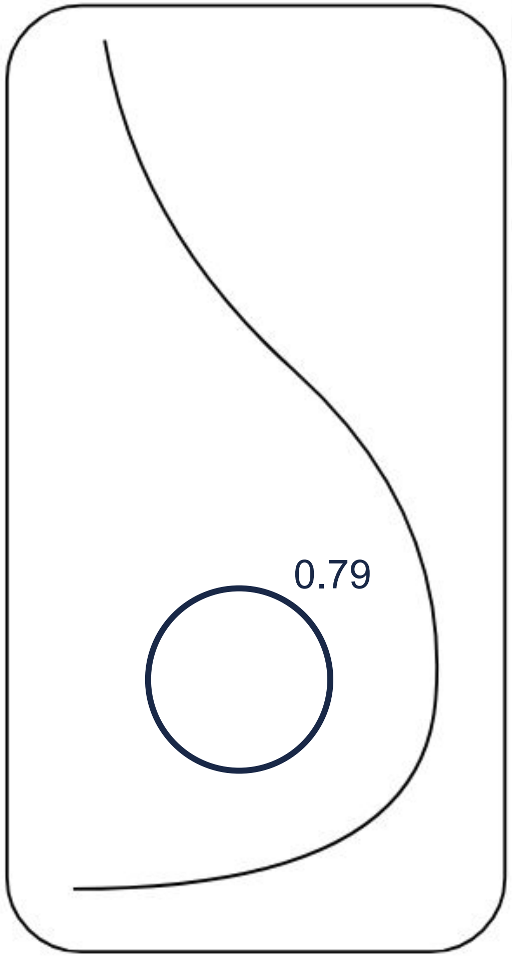
Calibration

Our product needs to work reliably on all these images

Robustness:
- Ability of algorithms to handle different image types

Calibration:
- Convert algorithm output to interpretable number
- Desired output distribution: 10% of exams in each 10-point bin
- 2 steps

Level of suspiciousness (fp-level)



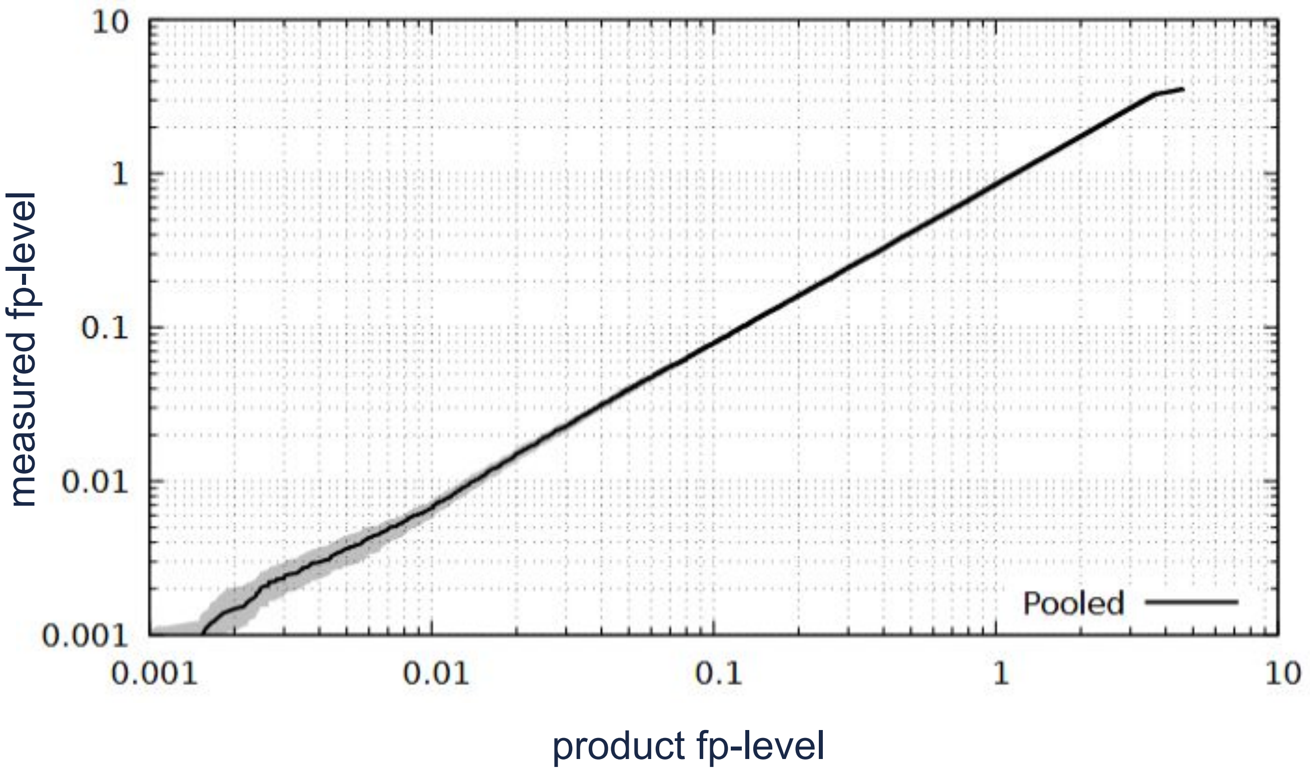
Calibration

Calibration is done on fixed dataset

On a set of normal exams:

Map algorithm output to measured fp-level

Tested on independent data



Calibration

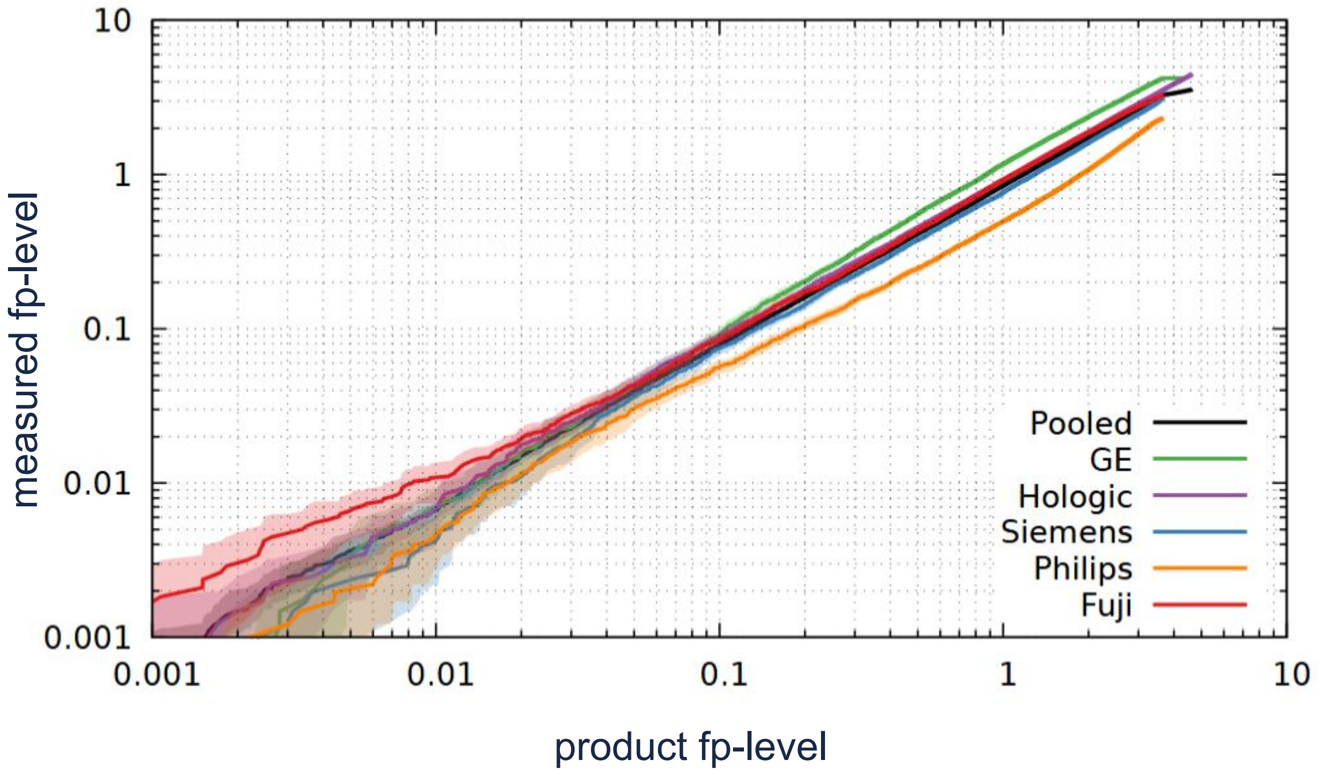
Calibrate on fixed dataset

On a set of normal exams:
    Map algorithm output to
    measured fp-level

Test on independent data

Customer doesn't care about
  Pooled!

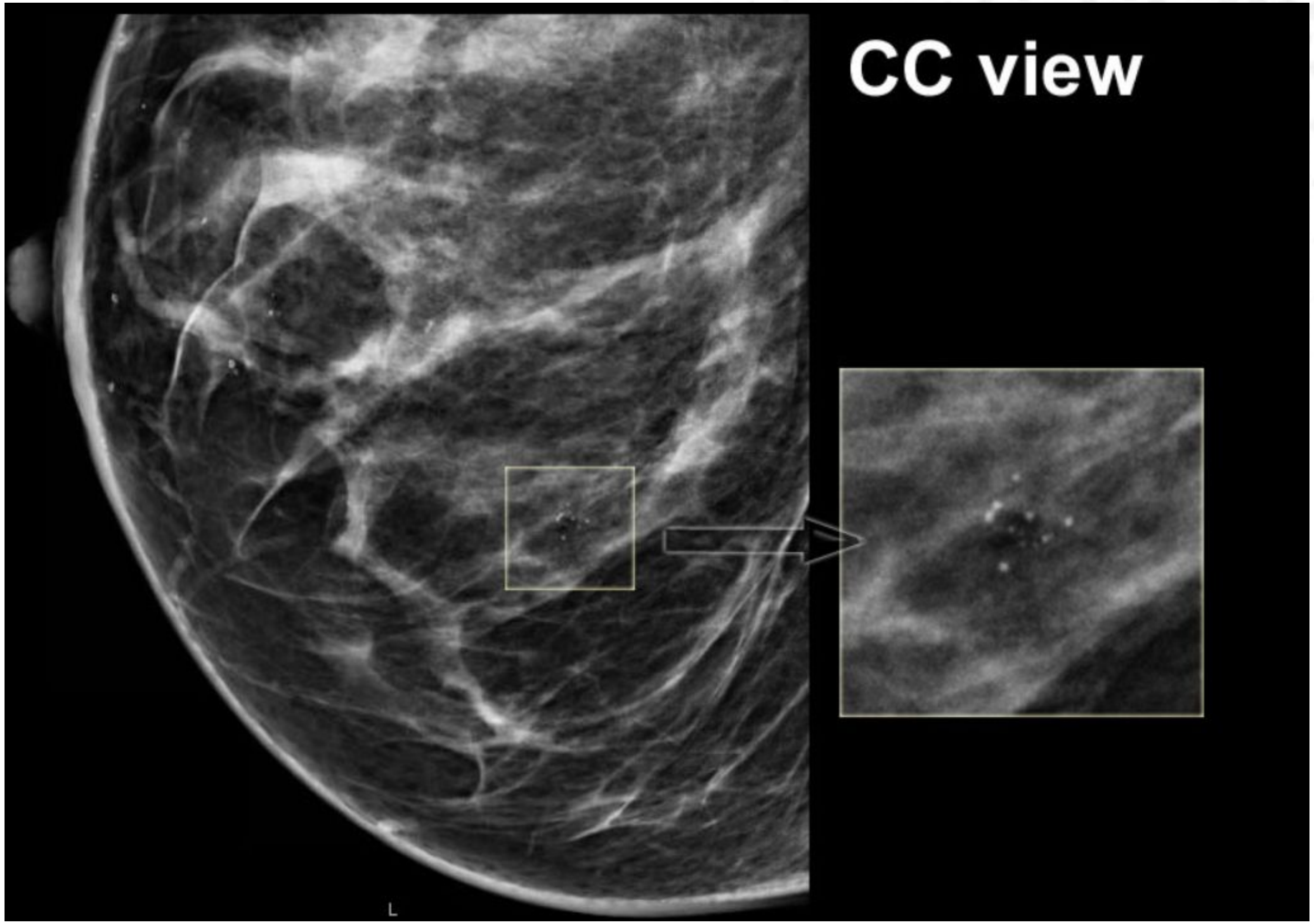
Manufacturer-specific calibration



Calibration

But that's not the full story…

Micro calcifications

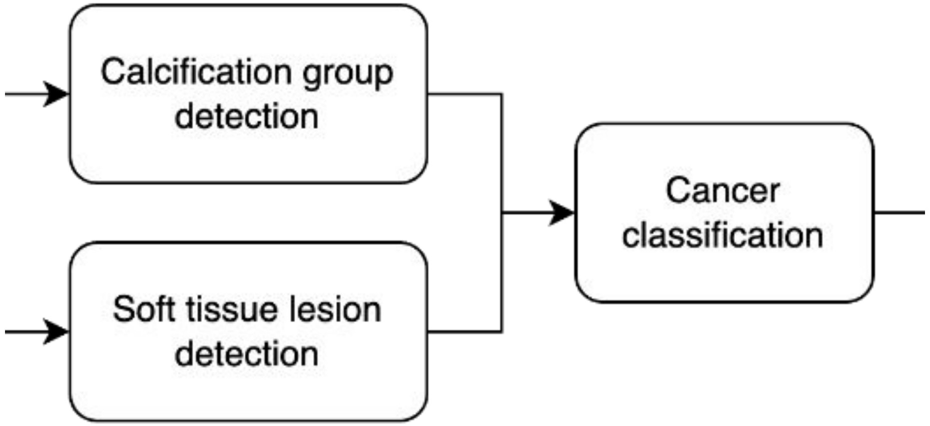


Calibration

Separate algorithms for calcifications and soft tissue lesions

Cannot be combined directly

=> More calibration steps necessary



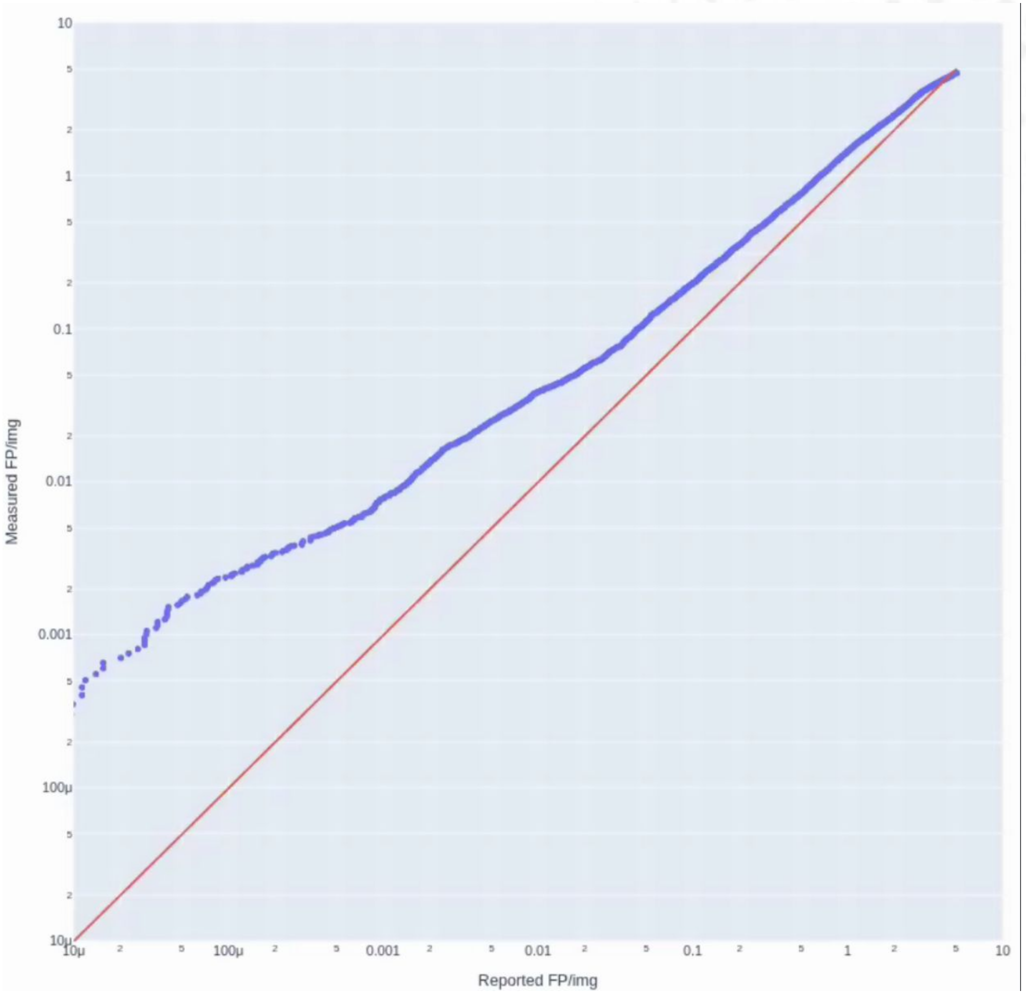
Calibration

Customer complaint:

- New machine
- Population?
- Screening?
- Number of images in exam?
- …

Options to address it?

- Custom calibration
- Retrain algorithms
- Retract compatibility



Calibration

Conclusion:

- Complicated pipeline
  - Need to understand every step
- Many considerations
- Why you need physicists!

Life in a company

Very different from academia:

- Way of working: agile (ish)
- Short term gratification
- Business
- Colleagues with different backgrounds
- Everything is a collaboration

Life in a company

Very different from academia:

- Not developing new technologies.
  Instead: bring new technology to market

- Very little physics in my daily life

- Code quality / documentation

- Amount of meetings

Life in a company

But also quite similar

- International environment
- Smart people
- Conferences/workshops(/trade shows)
- Complexity of work (though on different timescale)
- Understand all the details of a problem
- Certain amount of freedom

Contact details

Email: ruud.peeters@screenpointmed.com

LinkedIn: https://www.linkedin.com/in/ruud-peeters/

Feel free to reach out!

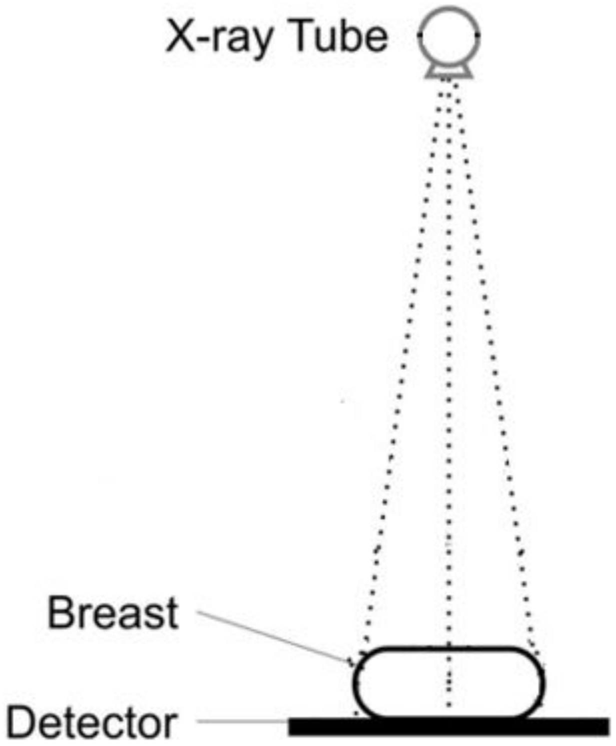
Calibration

For 2D images, life is relatively easy

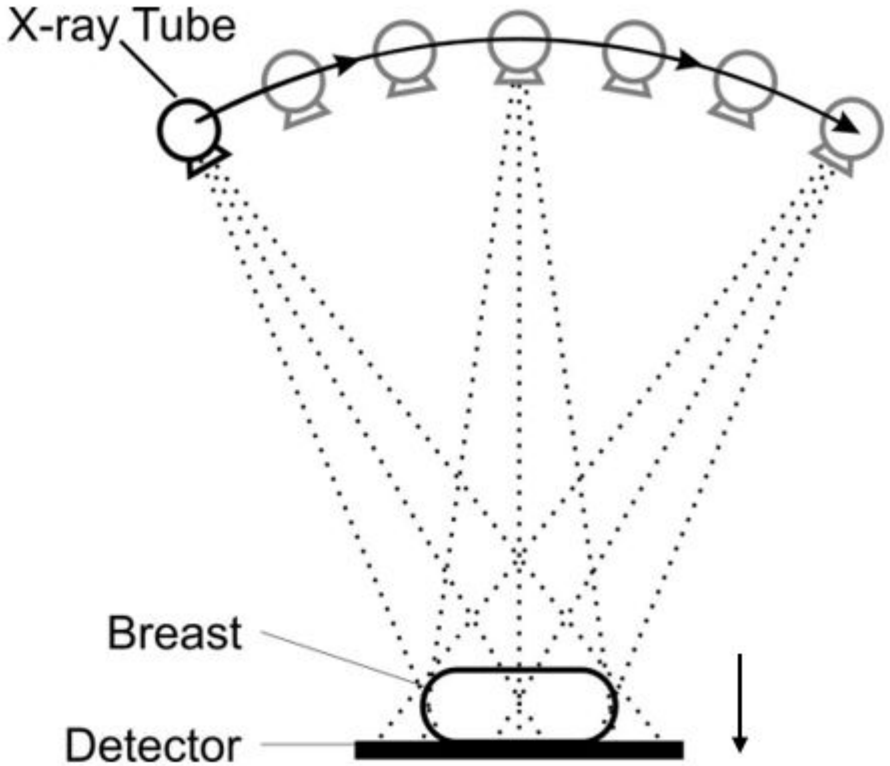
- Lot of training data: algorithms “learn” to handle different inputs

But that’s not the full story…

Calibration

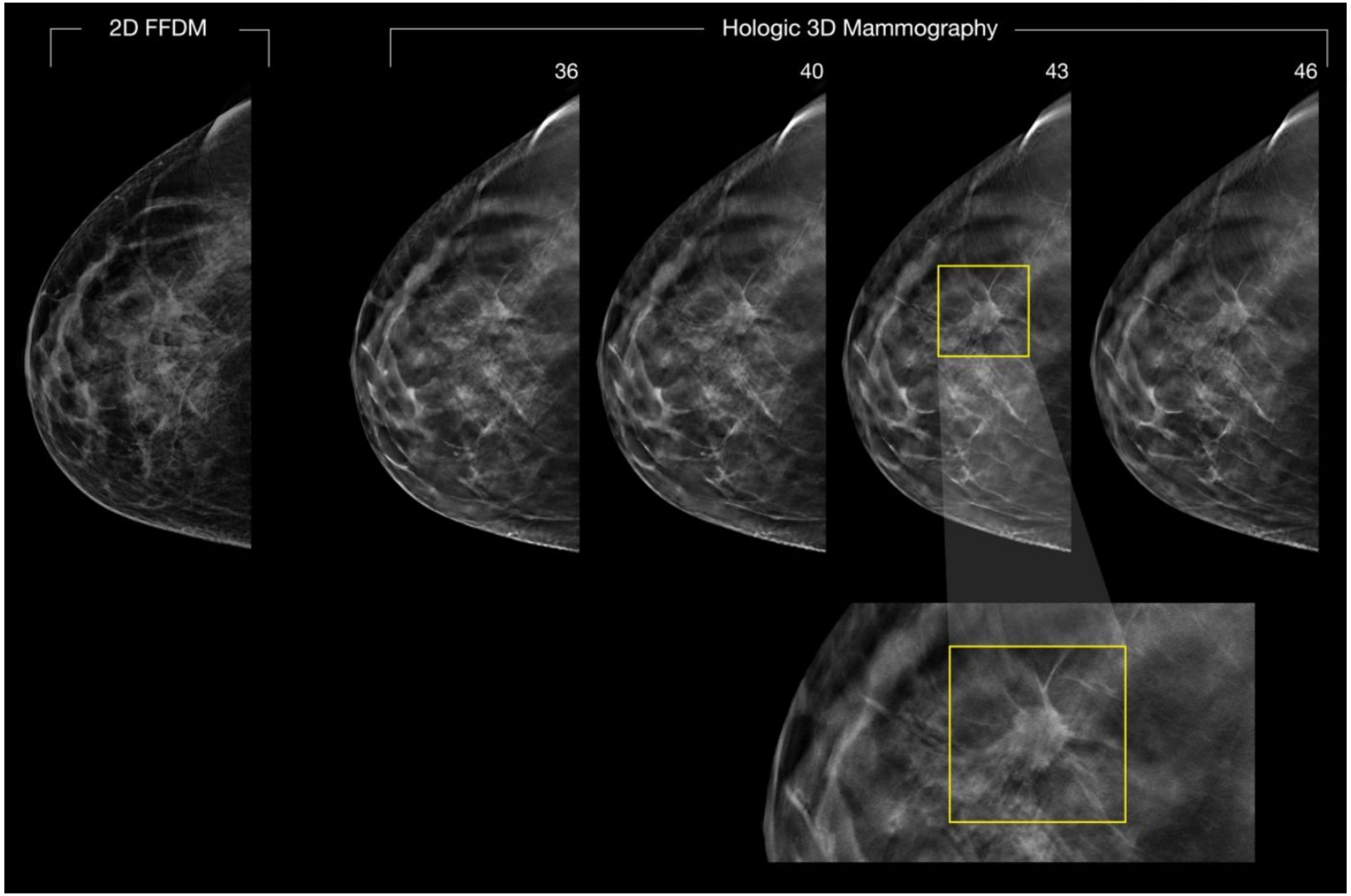


2D Planar digital mammography

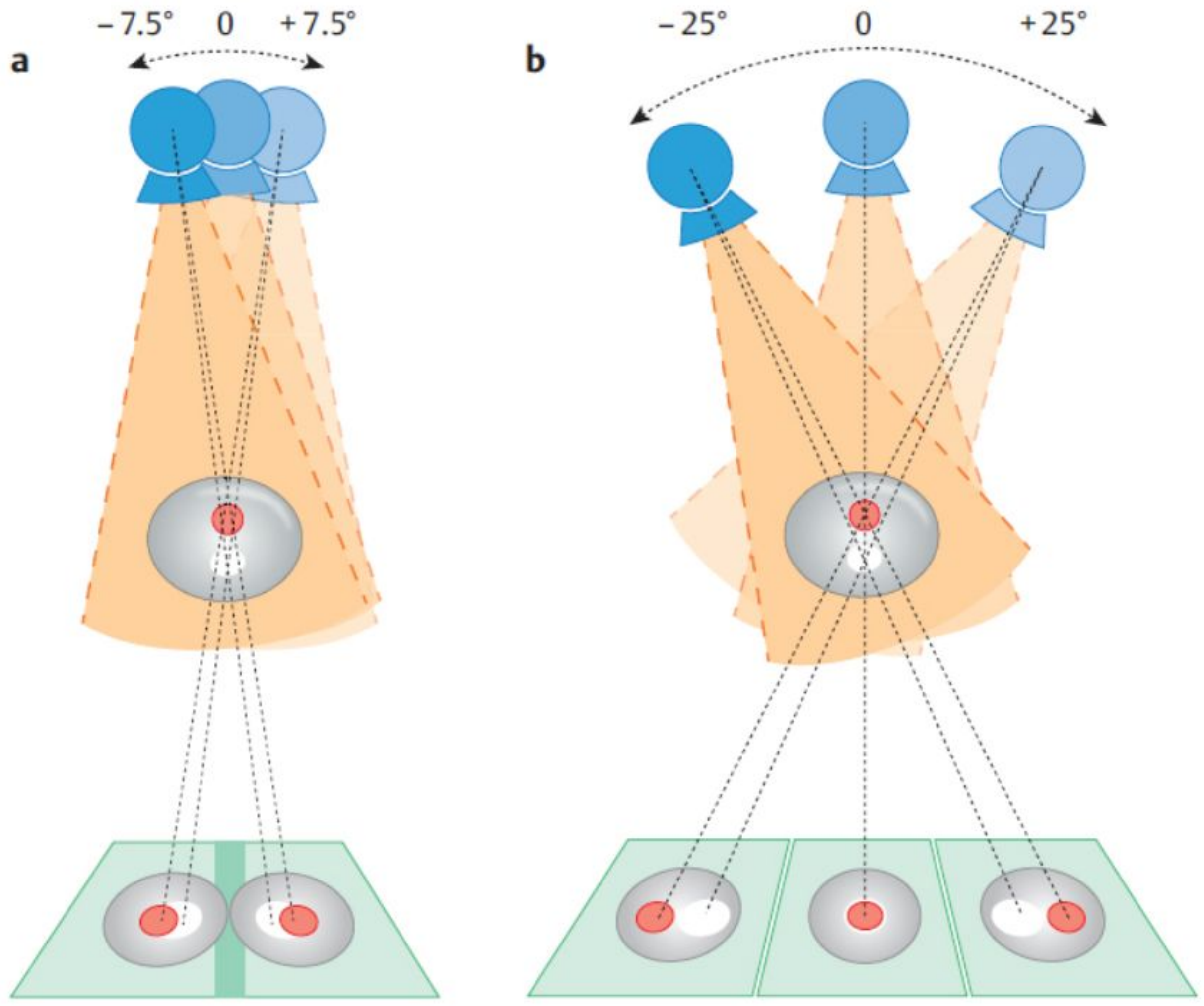


Digital Breast Tomosynthesis

Calibration



Calibration



Calibration

For 3D images:

Separate calibration for narrow-angle and wide-angle images

Separate calibration for specific manufacturers

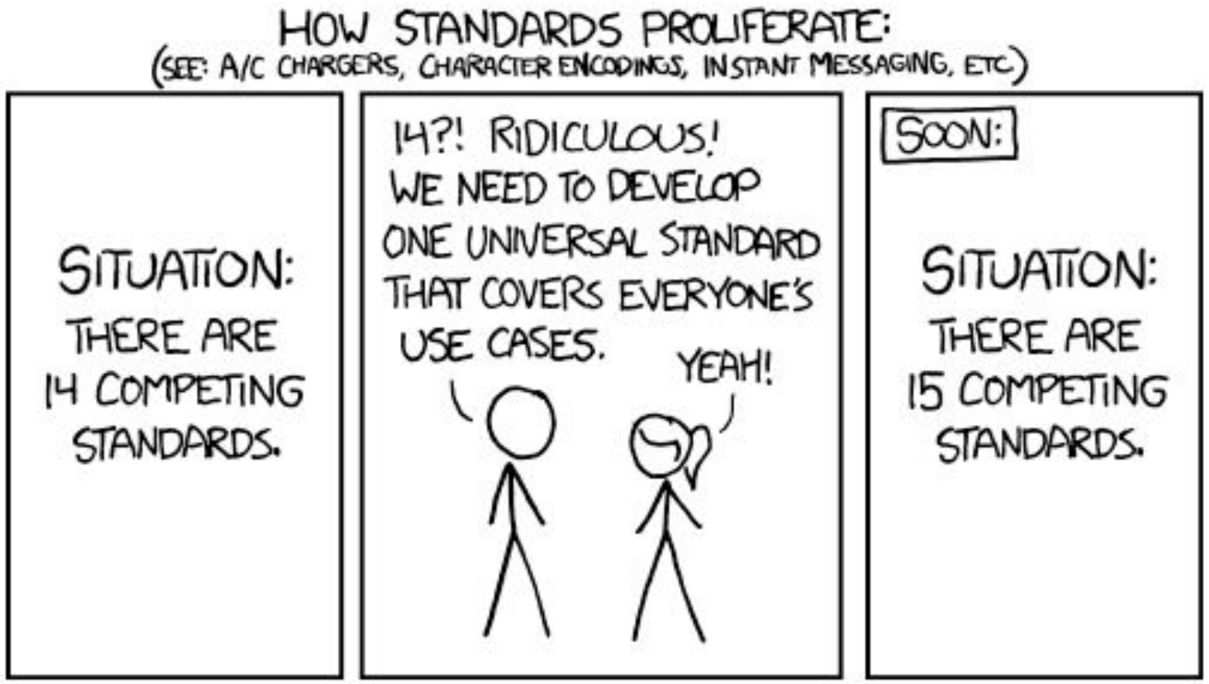
Data

>600k 2D studies

>100k DBT studies

92 TByte

DICOM file format



Less nice parts

Processes / regulations

Time to market

Changing priorities

Context switching

Lessons learned

Talk to people

If you're eager to learn and interested you can learn anything

It helps when a company knows how valuable physicists are

My journey

Was looking for clearer contribution to society

Interested in machine learning, though no hands-on experience

Kaggle

Found company via LinkedIn

False positives

"A false-positive mark is a mark made by the CAD system that does not correspond to the location of a lesion"

Only look at **normal exams**

- So every finding is a false positive

False positives

Algorithm output: [0,1]

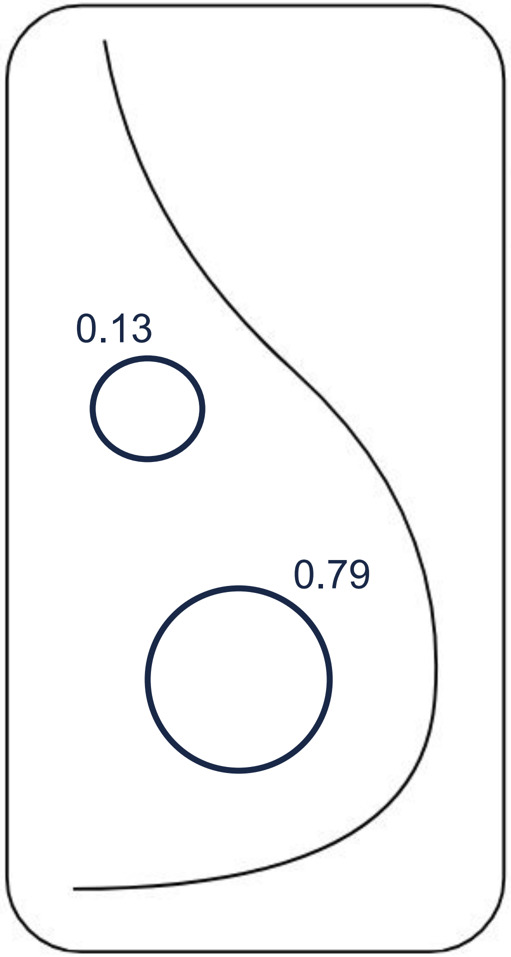
What does a score mean?

- By itself: nothing

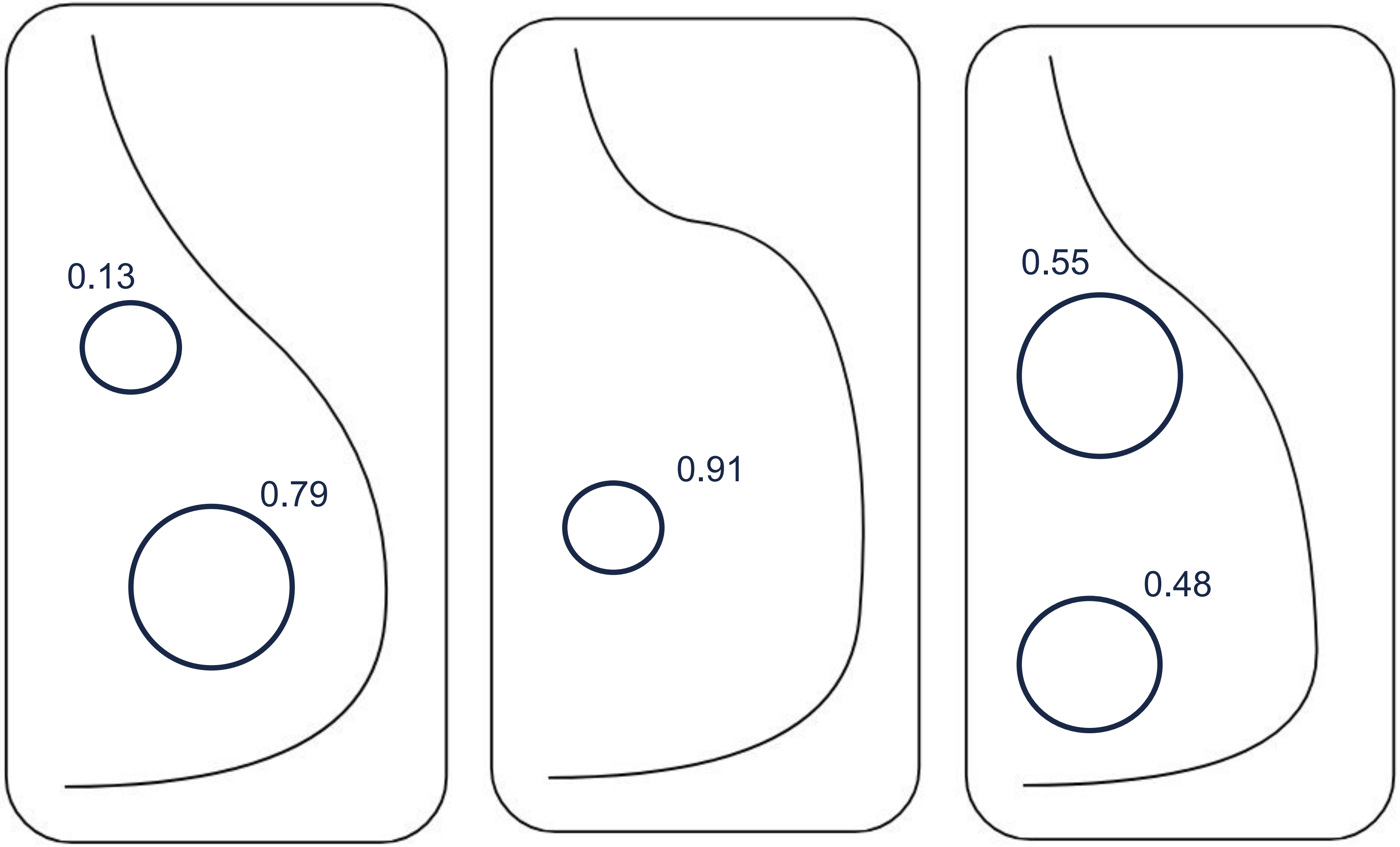
Need a way to 'interpret' what score means

- Calibration

Toy example



From network output to FP-level



Calibration dataset: 3 images

Threshold	# FPs	fp/image
0.0	5	1.667
0.13	5	1.667
0.45	4	1.333
0.55	3	1
0.79	2	0.667
0.91	1	0.333
1.0	0	0

From network output to FP-level

Created a lookup-table (LUT)

For new findings:
- Convert network output into expected number of false positives per image with at least that score

fp-level: expected number of false positives per image

Measure for how suspicious a finding is

Threshold	# FPs	fp/image
0.0	5	1.667
0.13	5	1.667
0.45	4	1.333
0.55	3	1
0.79	2	0.667
0.91	1	0.333
1.0	0	0

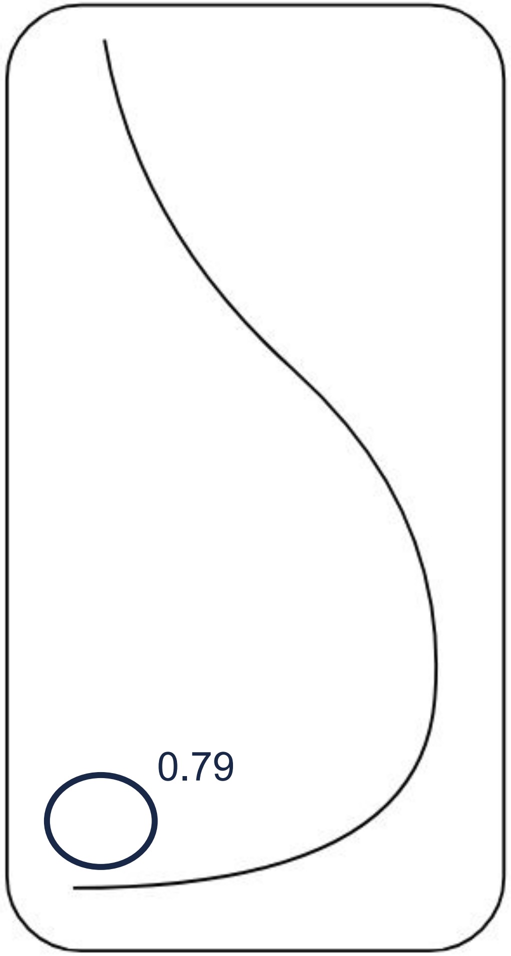
From network output to FP-level

score: 0.79
fp-level: 0.667

Expect to see 0.667 findings per image with score = 0.79

Interpretable!

Very coarse:
  Use more data



Score	fp-level
0.0	1.667
0.13	1.667
0.48	1.333
0.55	1
0.79	0.667
0.91	0.333
1.0	0

From network output to FP-level

In practice:

Calibrate on 1000 exams per vendor

Pooled of ~5000 normal exams

Score	fp-level
0.0000025	0.6420745
…	…
0.0013939	0.1458729
0.0013967	0.1457999
0.0013981	0.1457268
0.0013990	0.1456538
0.0013993	0.1455807
0.0014044	0.1455077
0.0014062	0.1454346
0.0014122	0.1453616
0.0014263	0.1452885
0.0014271	0.1452155
…	…
0.9976540	0.0000730

Calibration curves

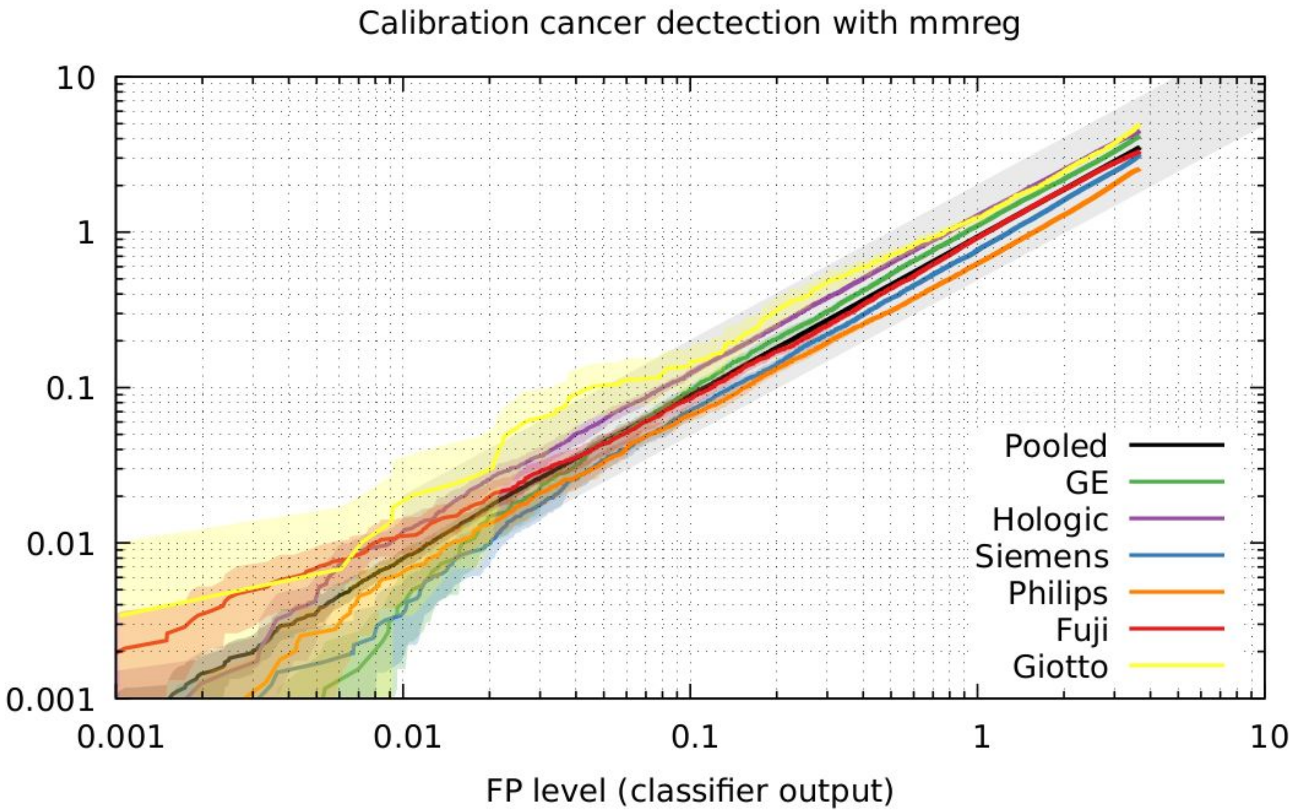
Calibration set:
-> Perfect by definition

Test set:
-> Some degree of deviation is acceptable

-> If a flavour is off, we can apply a flavour specific calibration

Careful, datasets are sensible to population changes (population, denser or bigger breast)

Measured (real)



Expected by look up table (LUT)