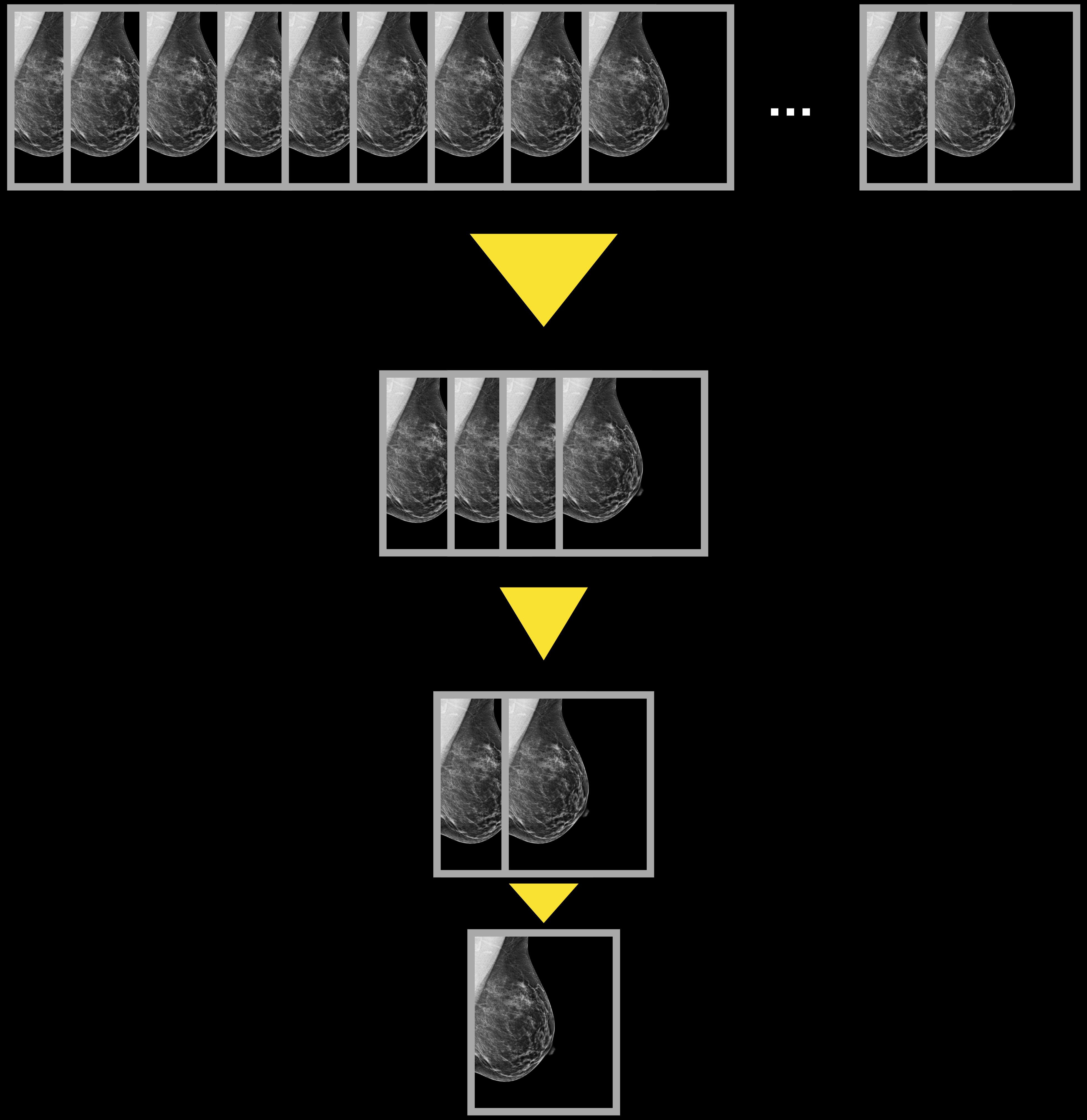
Agenda

- **Interpreting Mammograms**
    - Cancer Detection and Triage
- Assessing Breast Cancer Risk
- How to Mess up
- How to Deploy

Triaging Mammograms



1. Routine Screening

   1000 Patients

2. Called back for Additional Imaging

   100 Patients

3. Biopsy

   20 Patients

4. Diagnosis

   6 Patients

Triaging Mammograms

- **>99%** of patients are **cancer-free**
- Can we use a cancer model to automatically **triage** patients as **cancer-free**?
  - Reduce False positives, improve efficiency.
- Overall Idea:
  - Train a cancer detection model and pick a **cancer-free** threshold
    - chosen by **min probability** of a **caught-cancer** on the dev set
  - Radiologists can **skip** reading mammograms bellow threshold

Triaging Mammograms

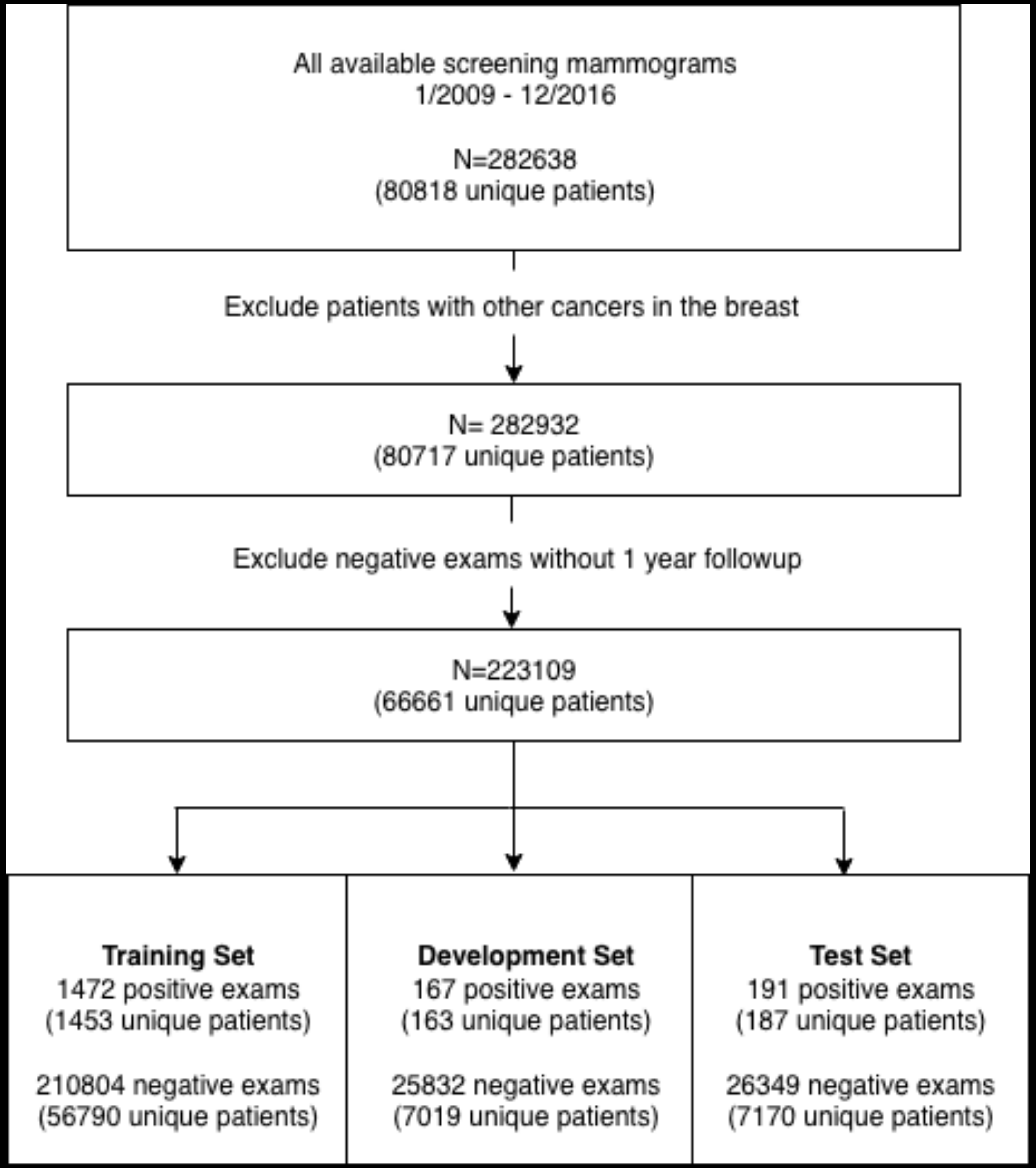
- The plan
    - **Dataset Collection**
    - Modeling
    - Analysis

Dataset Collection

- Consecutive Screening Mammograms
  - 2009-2016
- Outcomes from Radiology EHR, and Partners

5 Hospital Registry

- No exclusions based on race, implants etc.
- Split into Train/Dev/Test by Patient

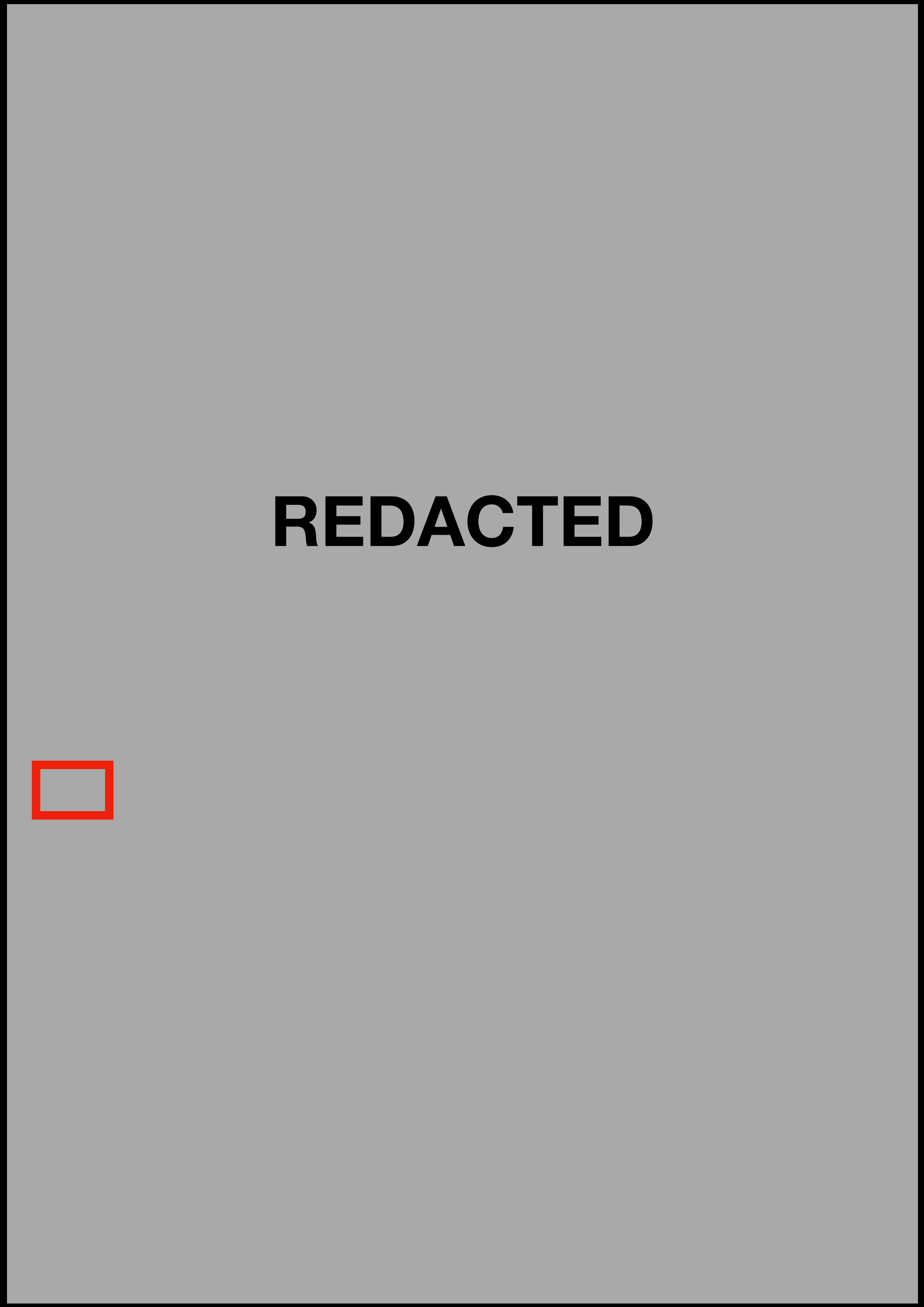


Triaging Mammograms

- The plan
  - Dataset Collection
  - **Modeling**
    - General challenges in working with Mammograms
    - Specific methods for this project
  - Analysis

Modeling: Is this just like ImageNet?

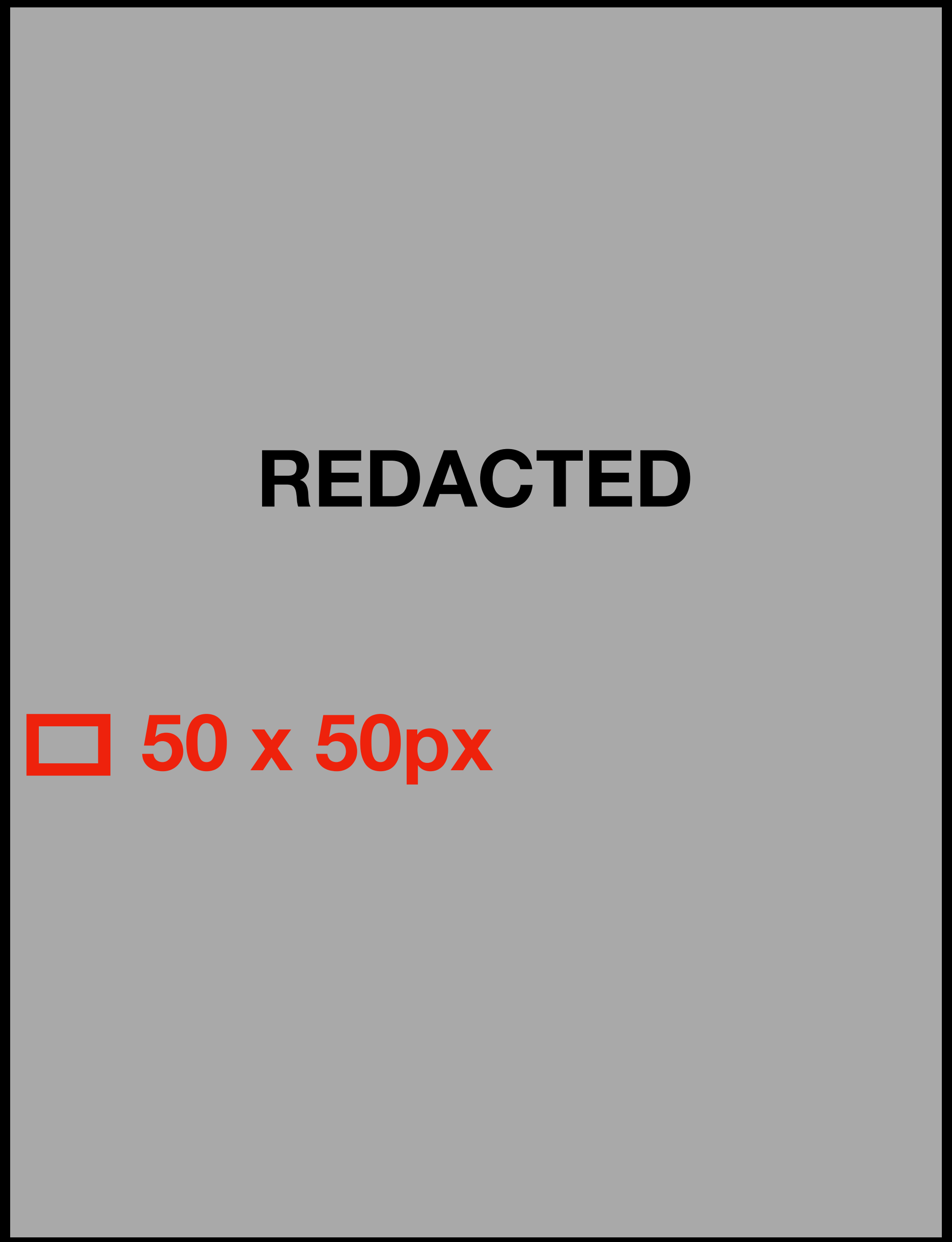
Modeling: Is this just like ImageNet?



Modeling: Is this just like ImageNet?

Many shared lessons, but important differences
in-size and nature of signal.

3200 px



2600 px

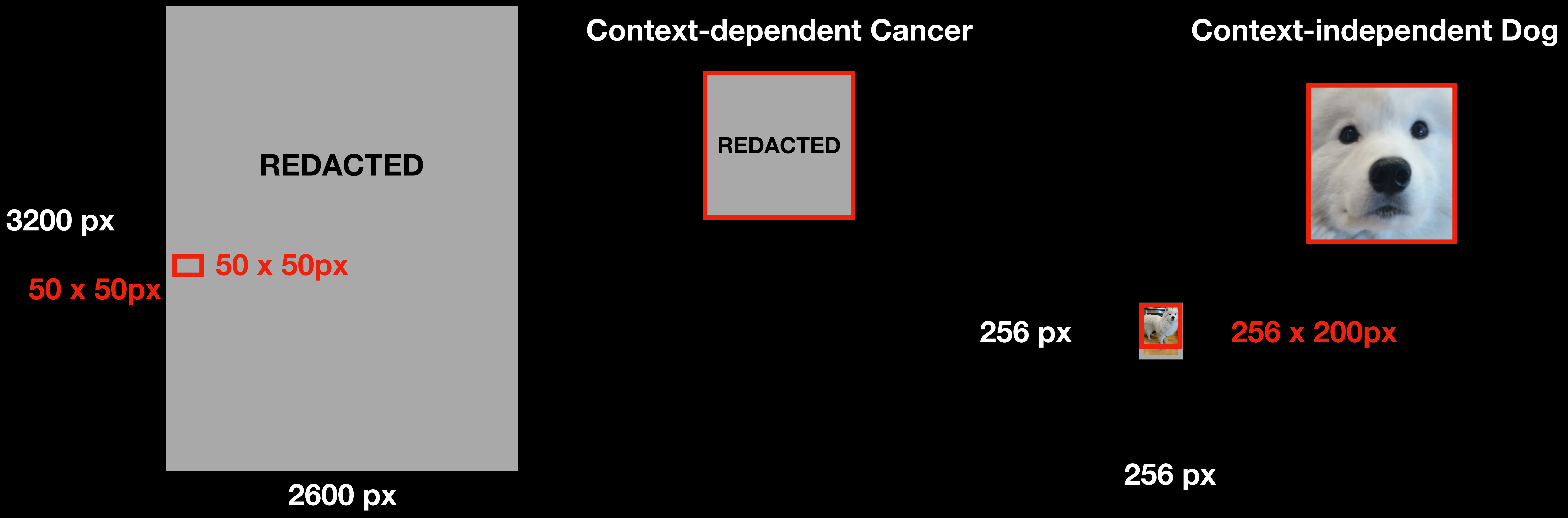
256 px

256 x 200px

256 px

Modeling: Is this just like ImageNet?

Many shared lessons, but important differences in-size and nature of signal.



Modeling: Challenges

- **Size of Object / Size of Image:**
    - Mammo: **~1%**
- **Class Balance:**
    - Mammo: **0.7%** Positive
    - **220,000** Exams, **<2,000** Cancers
- **Images per GPU:**
    - **3** Images (< 1 Mammogram)
    - **128** ImageNet Images
- Dataset Size
    - **12+** TB

The data is too small!

The data is too big!

Modeling: Key Choices

- How do we make the model actually **learn**?
  - **Initialization**
  - Optimization / Architecture Choice
- How to use the model?
  - Aggregation across images
  - Triage Threshold
  - Calibration

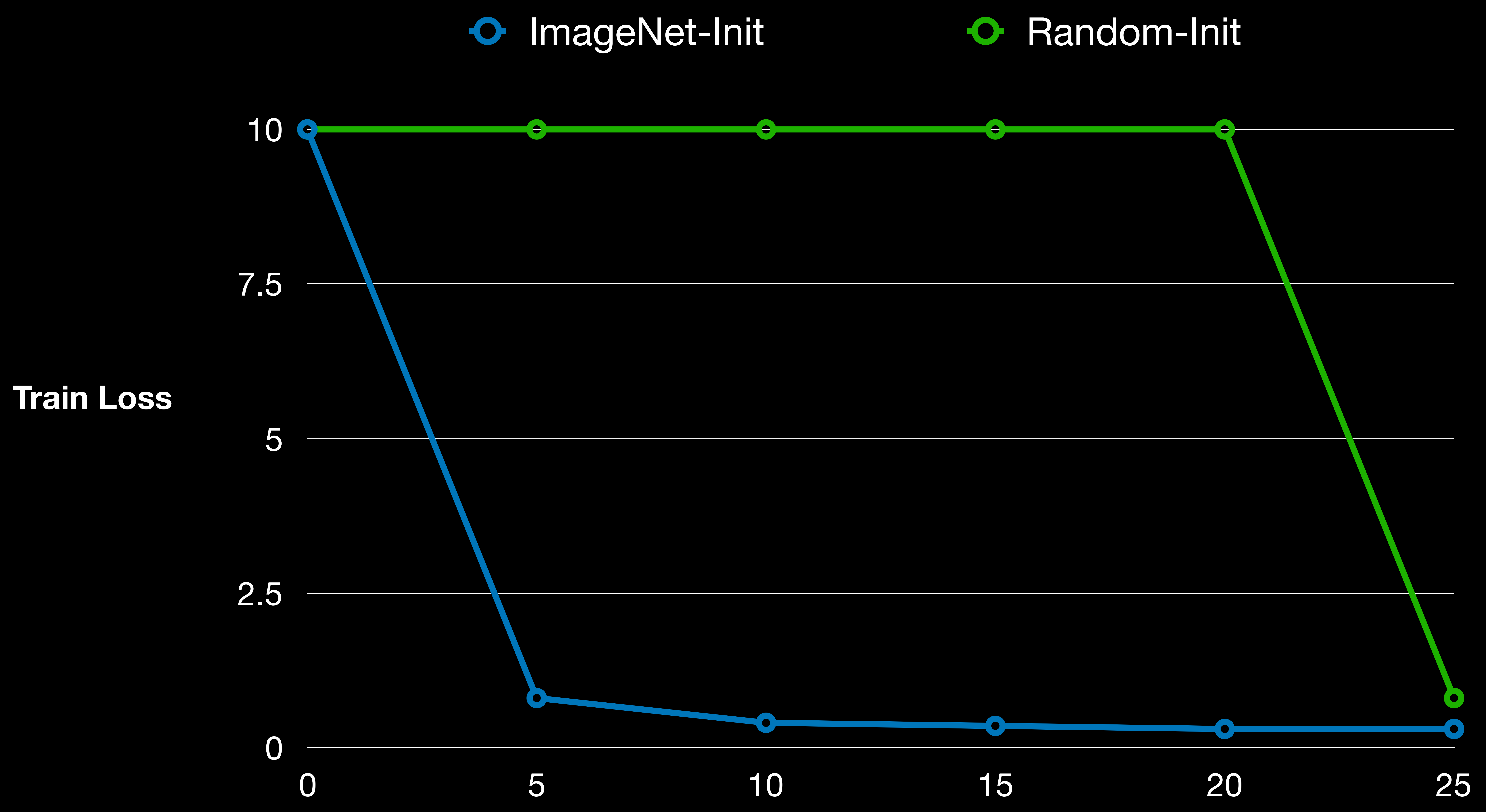
Modeling: Actual Choices

- How do we make the model learn?
  - Initialization
    - **ImageNet Init**
  - Optimization
    - Batch size: **24**
    - **2** steps on **4** GPUs for each optimizer step
    - Sample **balanced batches**
  - Architecture Choice
    - **ResNet-18**

Modeling: Key Choices

- How do we make the model actually **learn**?
  - Initialization
  - Optimization / Architecture Choice
- How to use the model?
  - Aggregation across images
  - Triage Threshold
  - Calibration

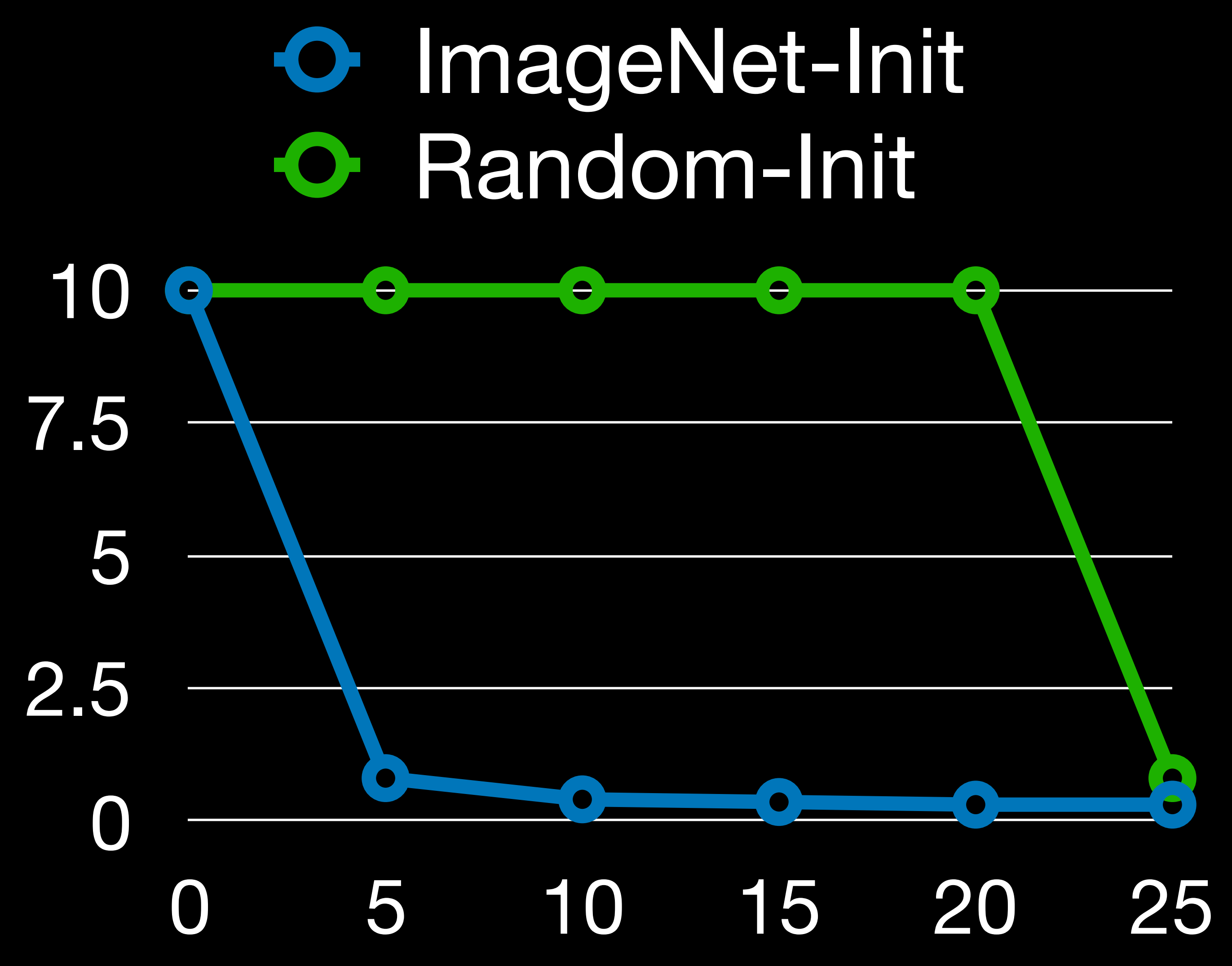
Modeling: Initialization

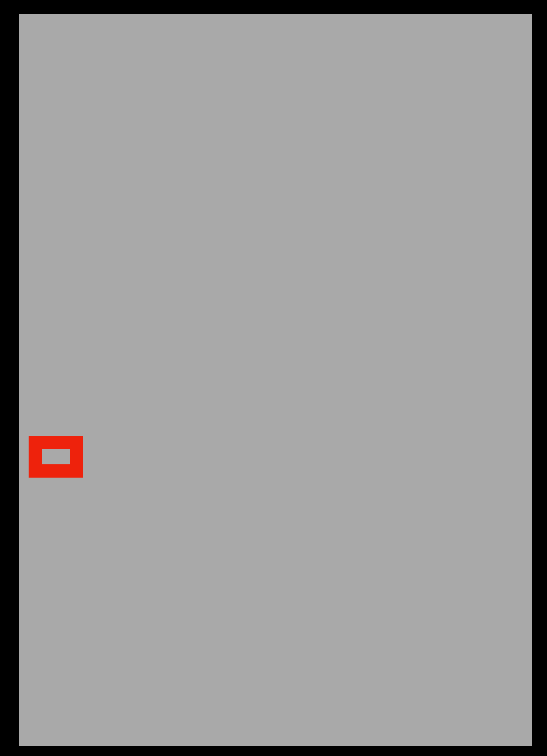


Modeling: Initialization

Empirical Observations

- ImageNet initialization learns immediately.
    - Transfer of particular filters?
        - Hard edges / shapes not shared
    - Transfer of BatchNorm Statistics
- Random initialization doesn't fit for many epochs until sudden cliff.
    - Unsteady BatchNorm statistics (3 per GPU)



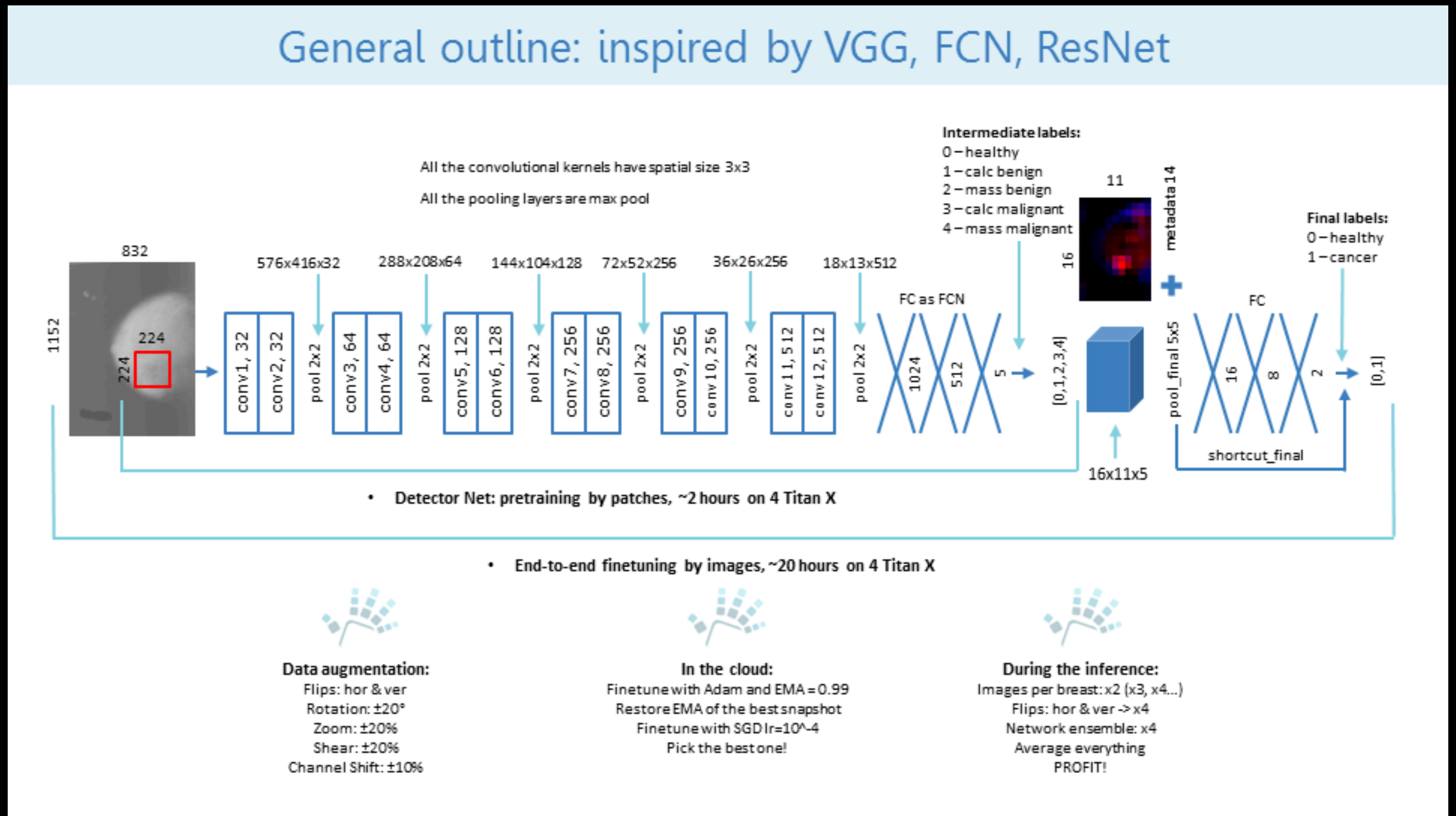


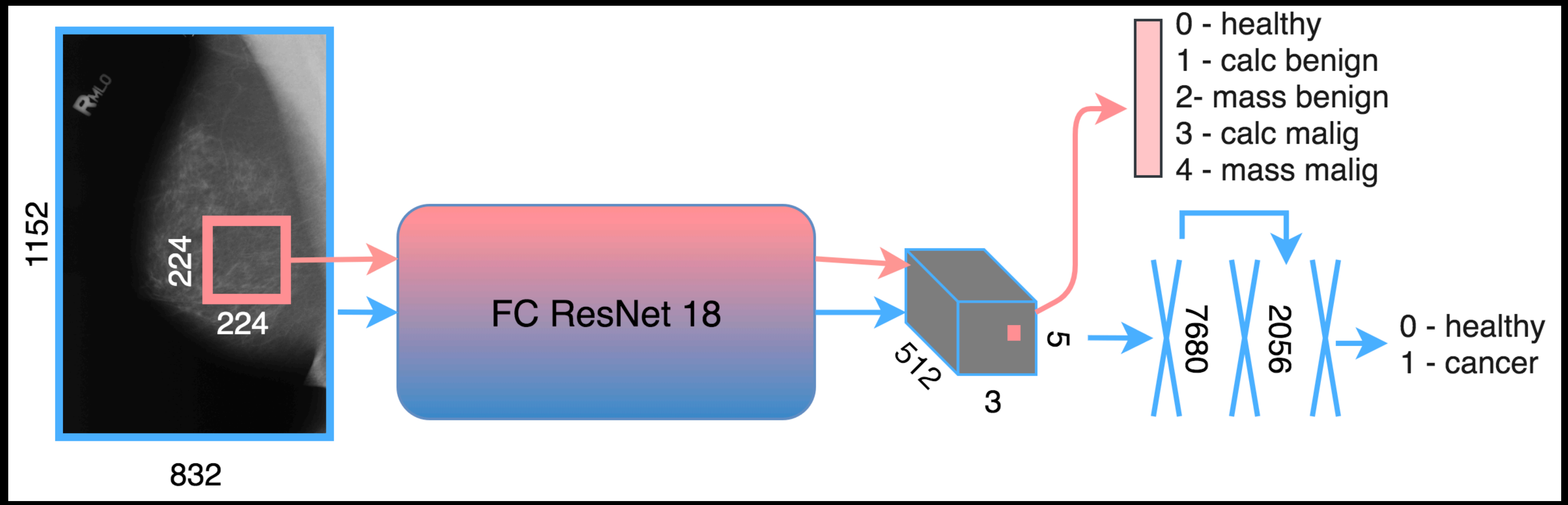
Modeling: Key Choices

- How do we make the model actually **learn**?
  - Initialization
  - **Optimization / Architecture Choice**
- How to use the model?
  - Aggregation across images
  - Triage Threshold
  - Calibration

Modeling: Common Approaches

- **Core problem:**
  - Low signal-to-noise ratio
- **Common Approach:**
  - Pre-Train at Patch level
    - High batch-size > 32
  - Fine-tune on full images
    - Low batch-size < 6





Modeling: Base Architecture

- Many valid options:
  - VGG, ResNet, Wide-ResNet, DenseNet…
- Fully convolutional variants (like ResNet) are the easiest to transfer across resolutions.
  - Use ResNet-18 as base for speed/performance trade-off.

Modeling: Building Batches

- **Build Balanced Batches:**
  - Avoid model forgetting
- Bigger batches means **less noisy stochastic gradients**

$$w := w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{n} \nabla Q_i(w)/n,$$

  - Makes 2-stage training unnecessary.
- Trade-off: the bigger the batches, the slower the training

bs	tr acc	dev acc	dev auc	test acc	test auc
PACNN					
2	73.98%	72.32%	0.80	70.61%	0.74
4	85.84%	81.19%	0.89	77.33%	0.83
10	85.25%	80.64%	0.89	77.60%	0.83
16	84.79%	79.72%	0.89	77.47%	0.84
ResNet18 on image size 832 × 1152					
2	65.09%	67.60%	0.71	68.28%	0.63
4	77.74%	74.62%	0.82	71.58%	0.75
10	85.34%	79.29%	0.87	79.16%	0.83
16	82.44%	79.53%	0.89	74.67%	0.82

Old Experiments on Film Mammography Dataset

Modeling: Key Choices

- How do we make the model actually **learn**?
  - Initialization
  - Optimization / Architecture Choice
- **How to use the model?**
  - Aggregation across images
  - Triage Threshold
  - Calibration

Modeling: Actual Choices

- How do we make the model learn?
  - Initialization
    - **ImageNet Init**
  - Optimization
    - Batch size: **24**
    - **2** steps on **4** GPUs for each optimizer step
    - Sample **balanced batches** with **data augmentation**
  - Architecture Choice
    - **ResNet-18**

Modeling: Actual Choices (Continued)

- Overall Setup:
  - Train Independently per Image
    - From each image, predict cancer in that breast
  - Get prediction for whole mammogram exam by taking max across Images
  - At each Dev Epoch, evaluate ability of model to Triage
  - **Use the model that can do Triage best on the development set.**

Not necessarily the highest AUC

Modeling: How to actually Triage?

- **Goal:**
  - Don’t miss a single cancer the radiologist would have caught.
- **Solution:**
  - Rank radiologist true positives by model-assigned probability
  - Return min probability of radiologist true positive in development set.

Modeling: How to calibrate?

- **Goal:**
    - Want model assigned probabilities to correspond to real probability of cancer.
        - Why is this a problem?
            - Model trained artificial incidence of 50% for optimization reasons.
- **Solution:**
    - Platt's Method:
        - Learn sigmoid to scale and shift probabilities to real incidence on the development set.

Triaging Mammograms

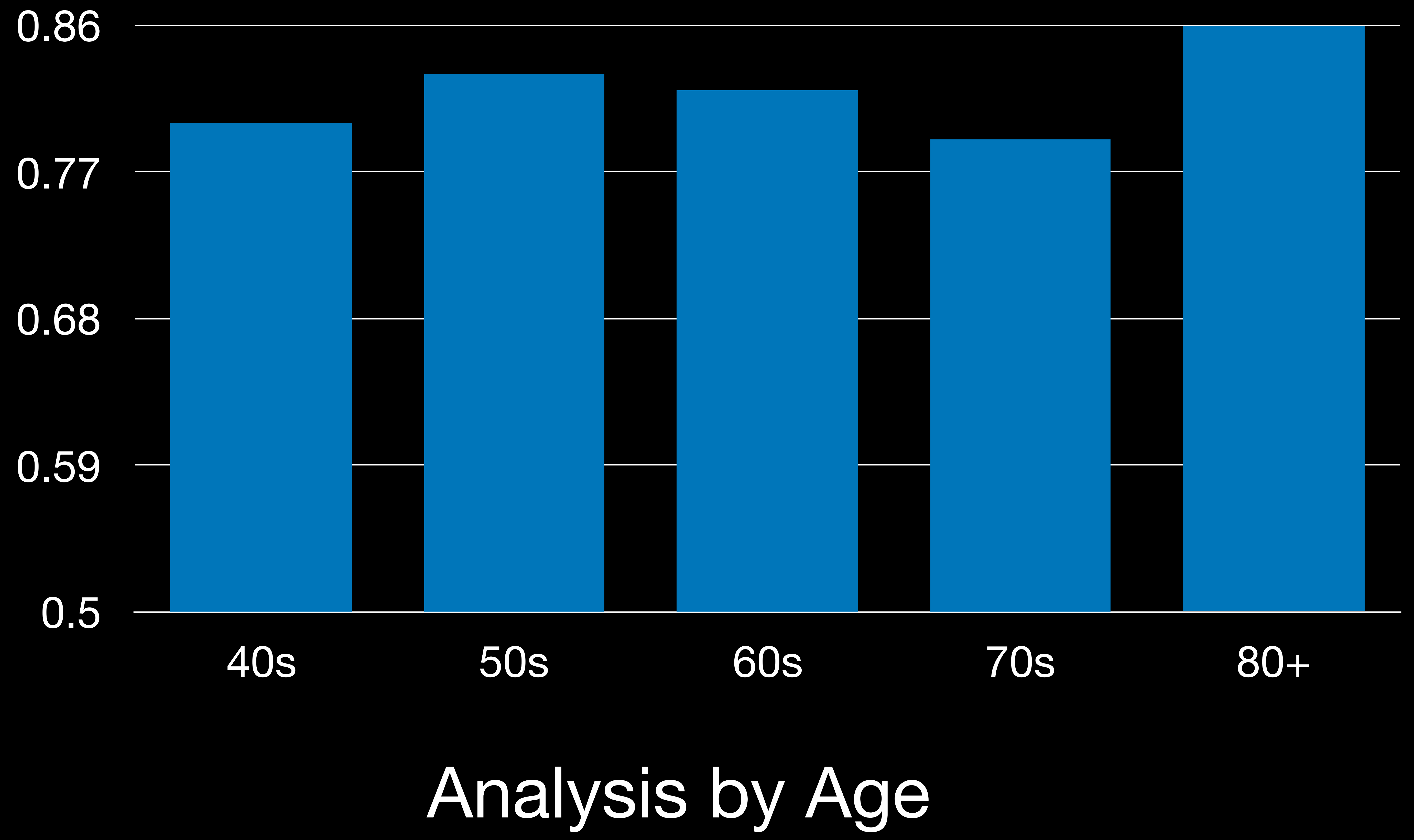
- The plan
  - Dataset Collection
  - Modeling
  - **Analysis**

Analysis: Objectives

- Is the model discriminative across all populations?
  - Subgroup Analysis by **Race**, **Age**, **Density**
- How does model relate to radiologist assessments?
- Simulate actual use of Triage on the Test Set

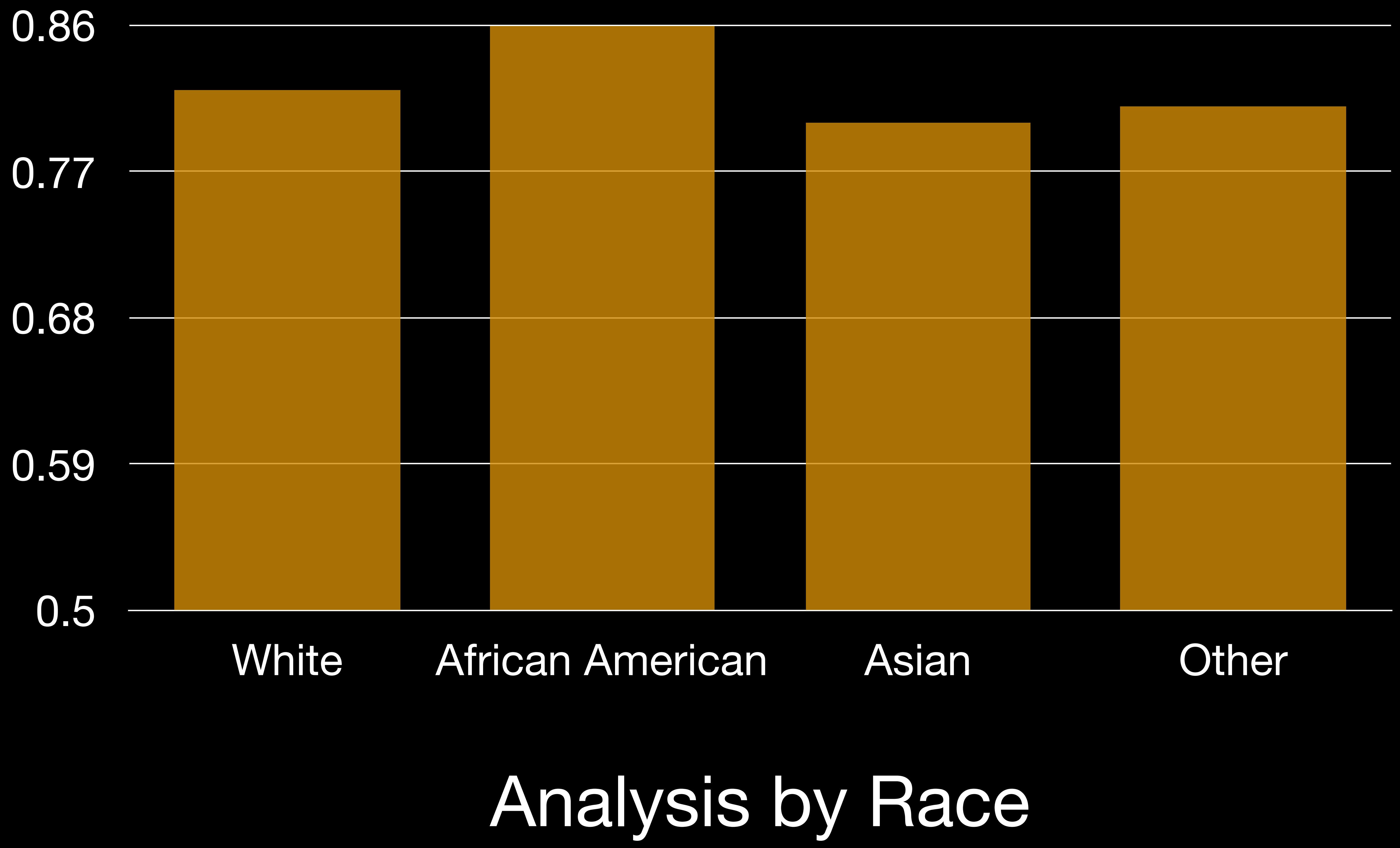
Analysis: **Model AUC**

Overall AUC: **0.82 (95%CI .80, .85)**



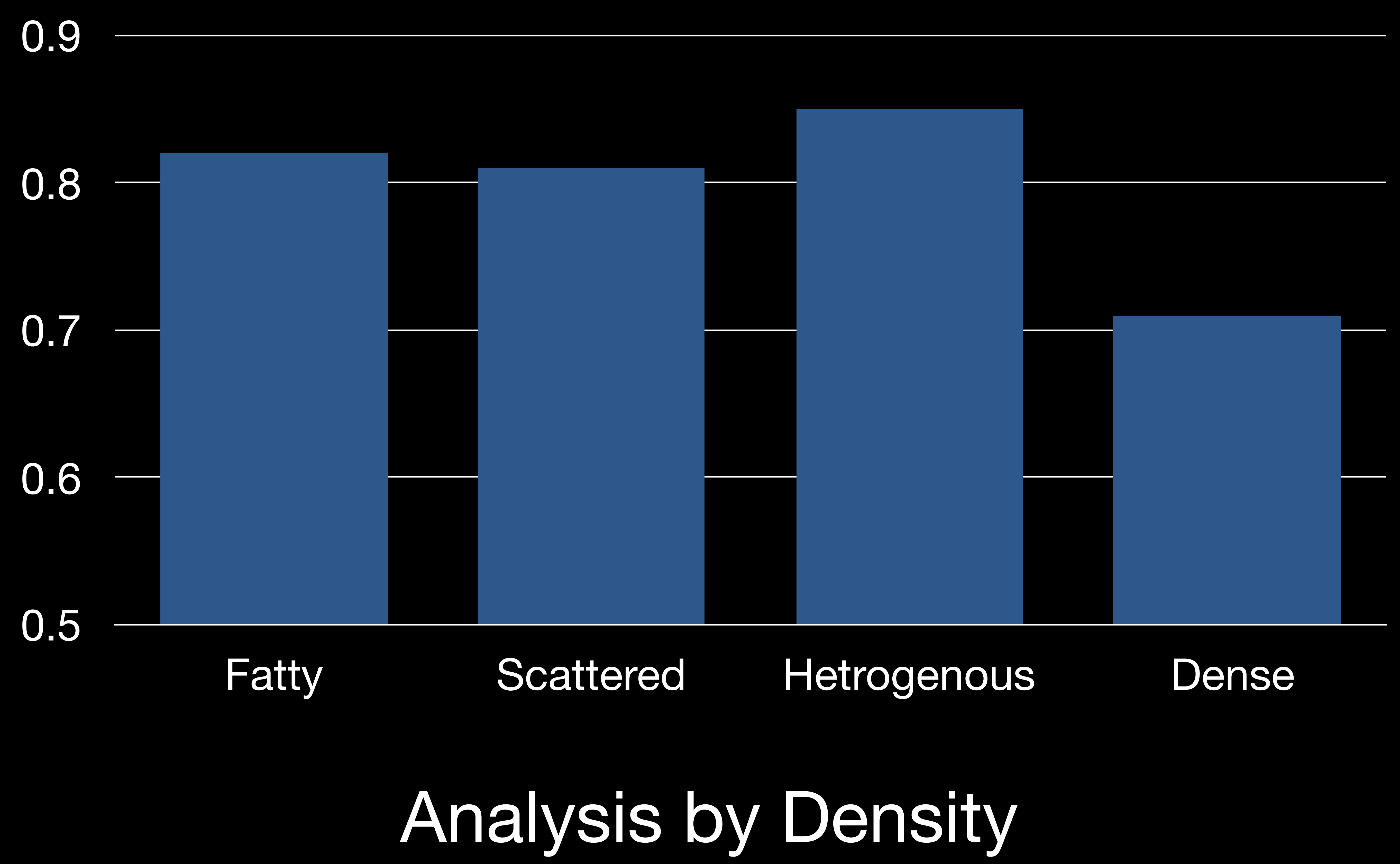
Analysis: Model AUC

Overall AUC: **0.82 (95%CI .80, .85)**

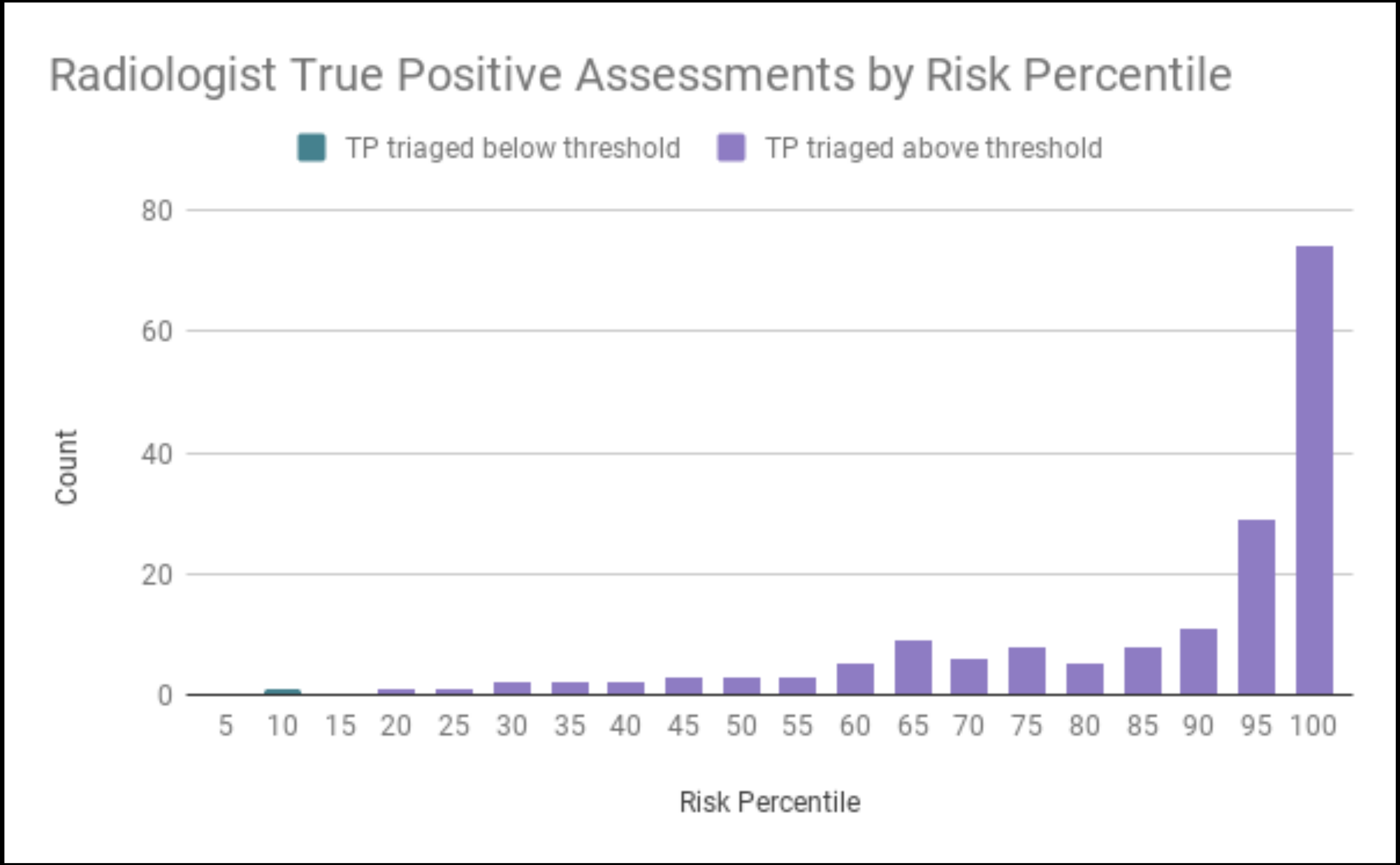


Analysis: **Model AUC**

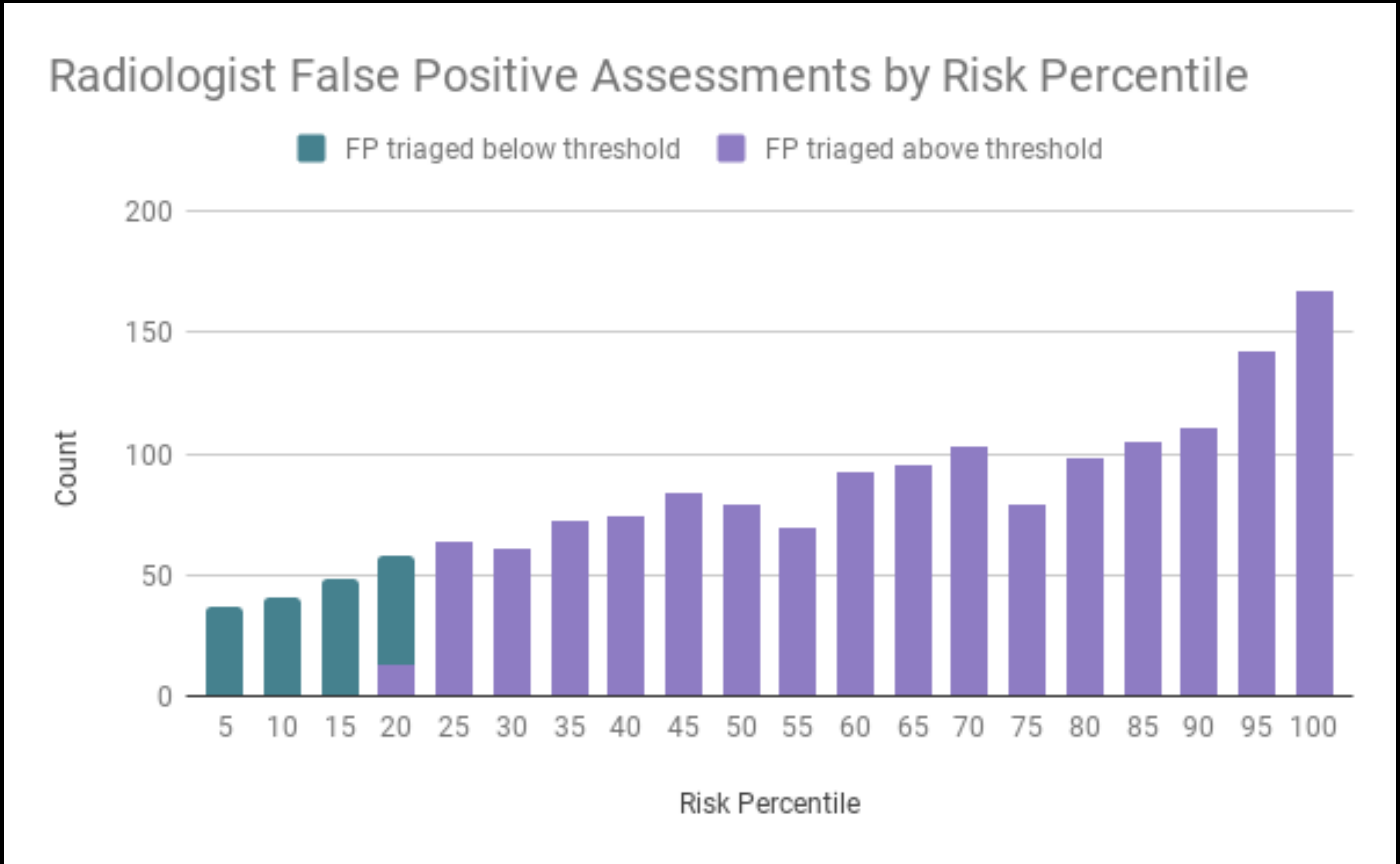
Overall AUC: **0.82 (95%CI .80, .85)**



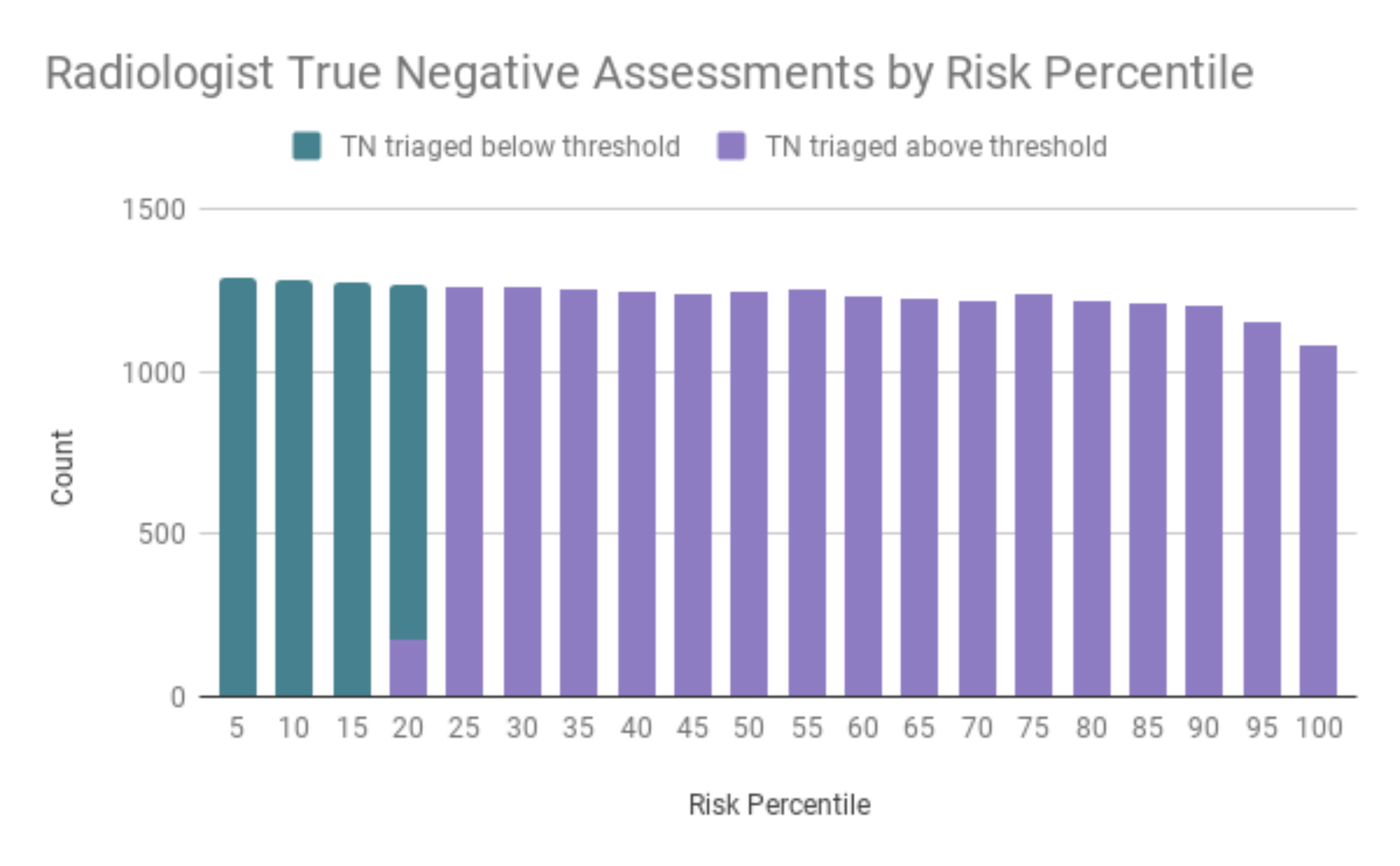
Analysis: Comparison to radioligists



Analysis: Comparison to radioligists



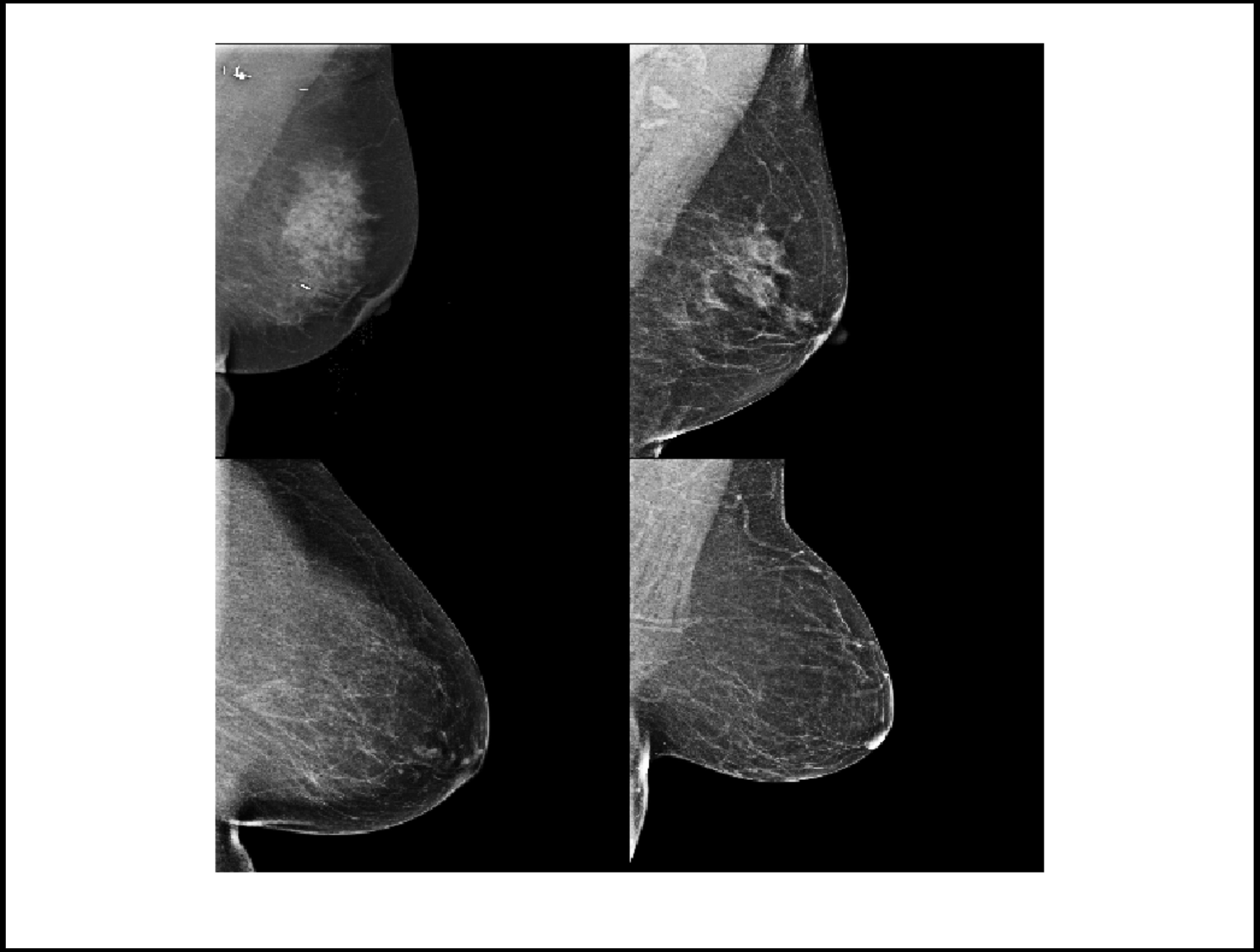
Analysis: Comparison to radioligists

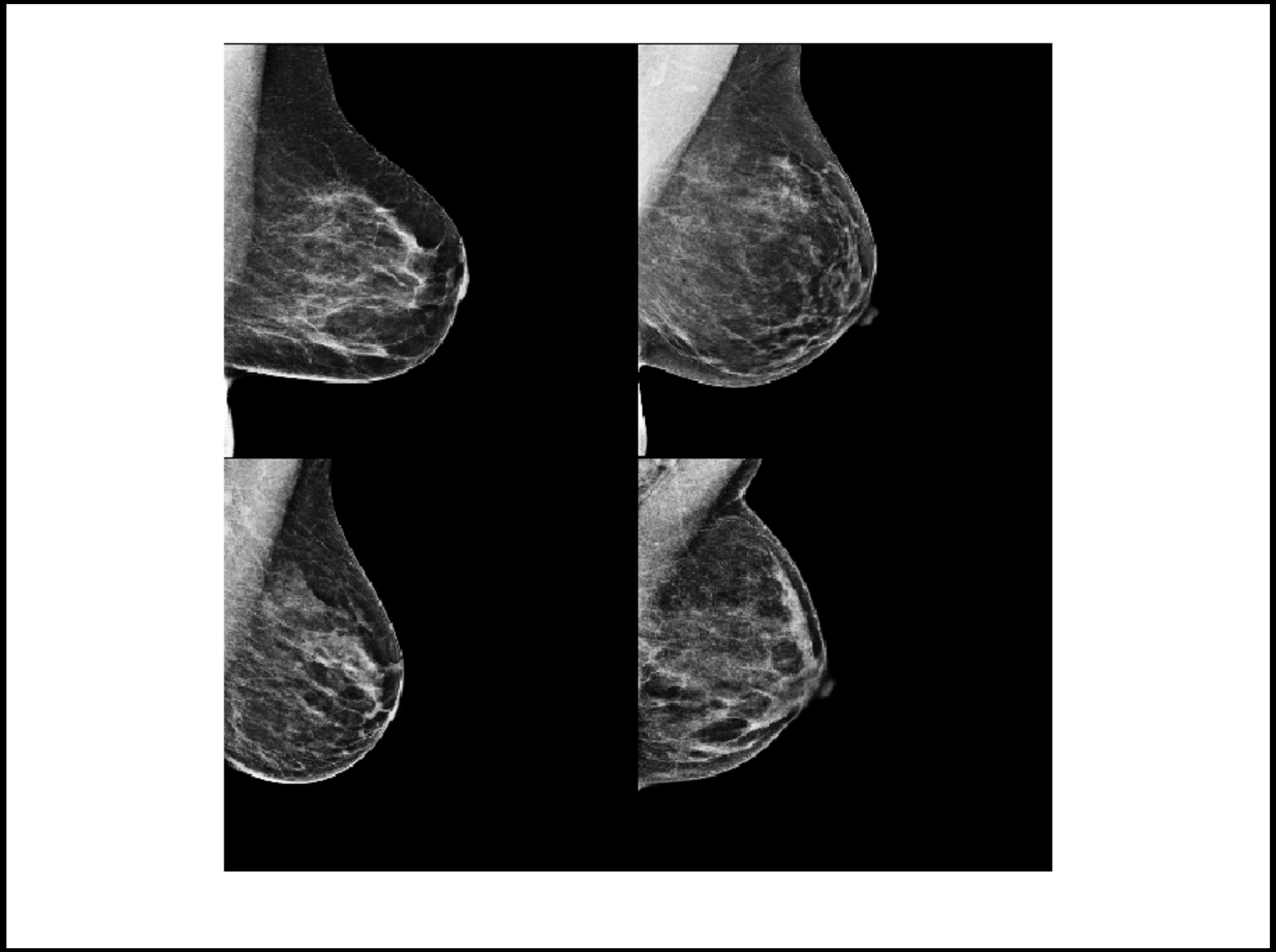


Analysis: Simulating Impact

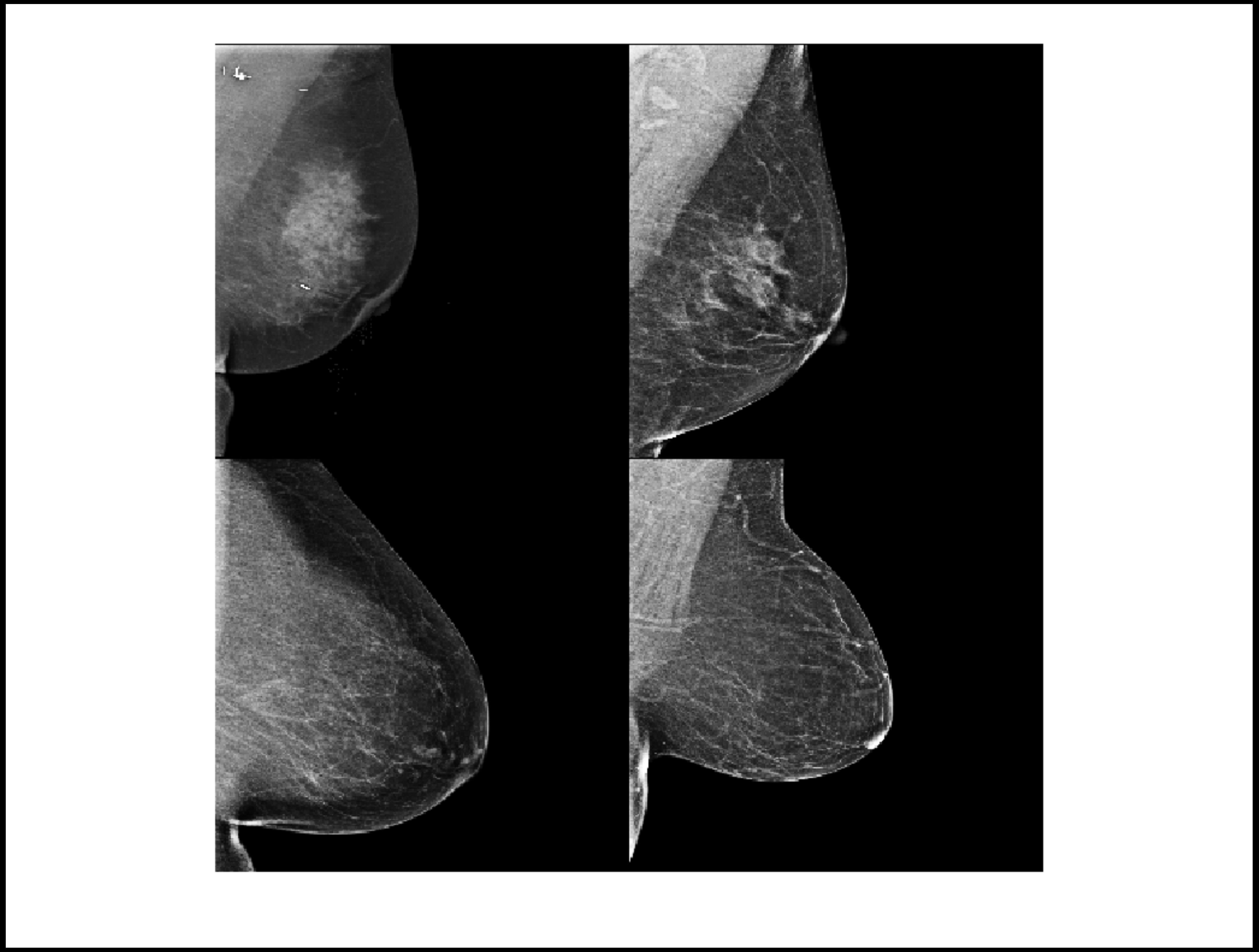
Setting	Sensitivity (95% CI)	Specificity (95% CI)	% Mammograms Read (95% CI)
Original Interpreting Radiologist	90.6% (86.7, 94.8)	93.0% (92.7, 93.3)	100% (100, 100)
Original Interpreting Radiologist + Triage	**90.1% (86.1, 94.5)**	**93.7% (93.0, 94.4)**	**80.7% (80.0, 81.5)**

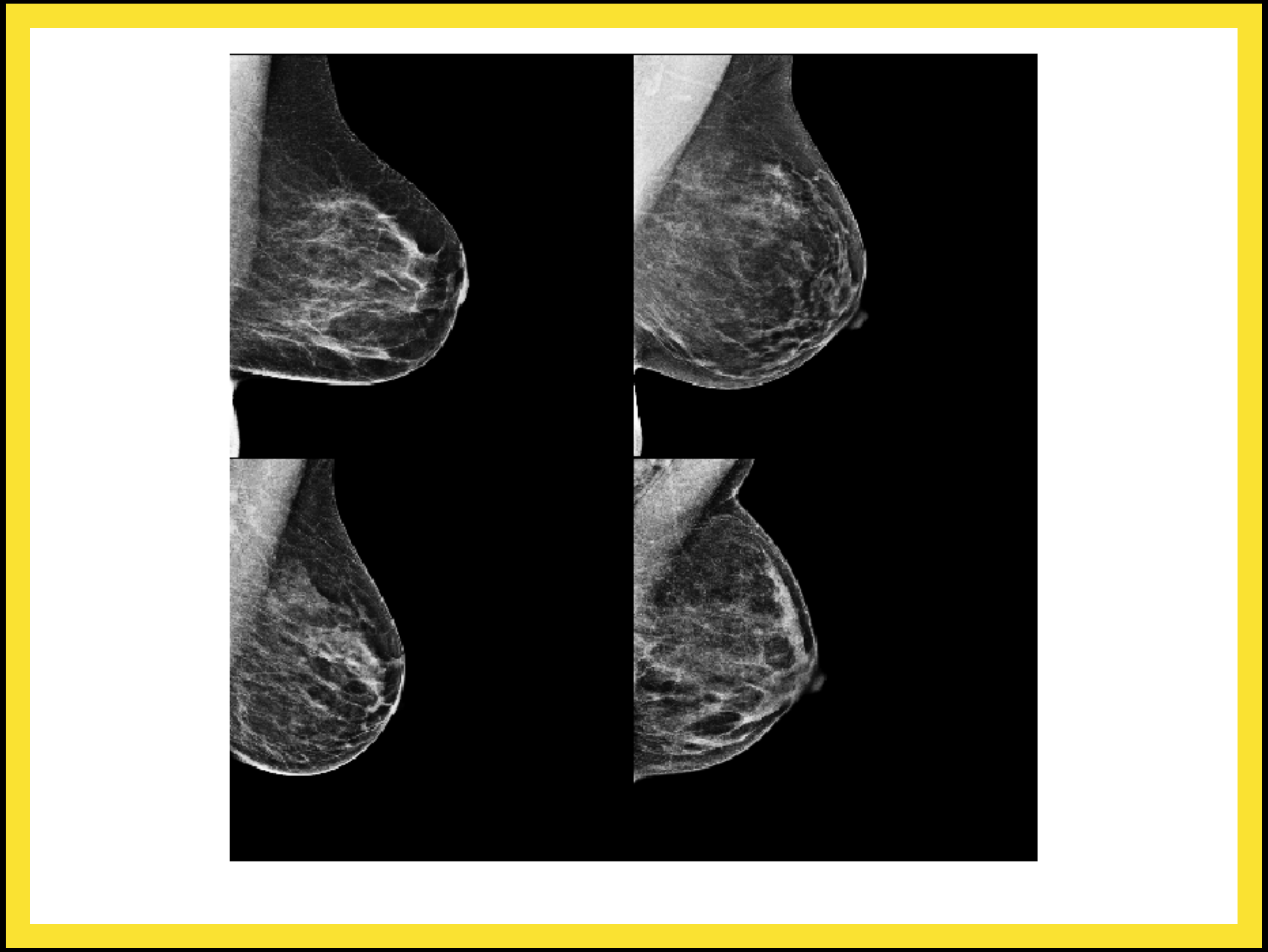
Example: Which were triaged?





Example: Which were triaged as cancer-free?





Next Step: Clinical Implementation

Agenda

- Interpreting Mammograms
  - Cancer Detection and Triage
- **Assessing Breast Cancer Risk**
- How to Mess up
- How to Deploy

Classical Risk Models: BCSC

Age
Family History
Prior Breast Procedure
Breast Density

Risk

AUC: **0.631**

AUC: 0.607 without Density

J Natl Cancer Inst. 2006 Sep 6;98(17):1204-14.

Prospective breast cancer risk prediction model for women undergoing screening mammography.

Barlow WE[1], White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DS, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K.

Assessing Breast Cancer Risk

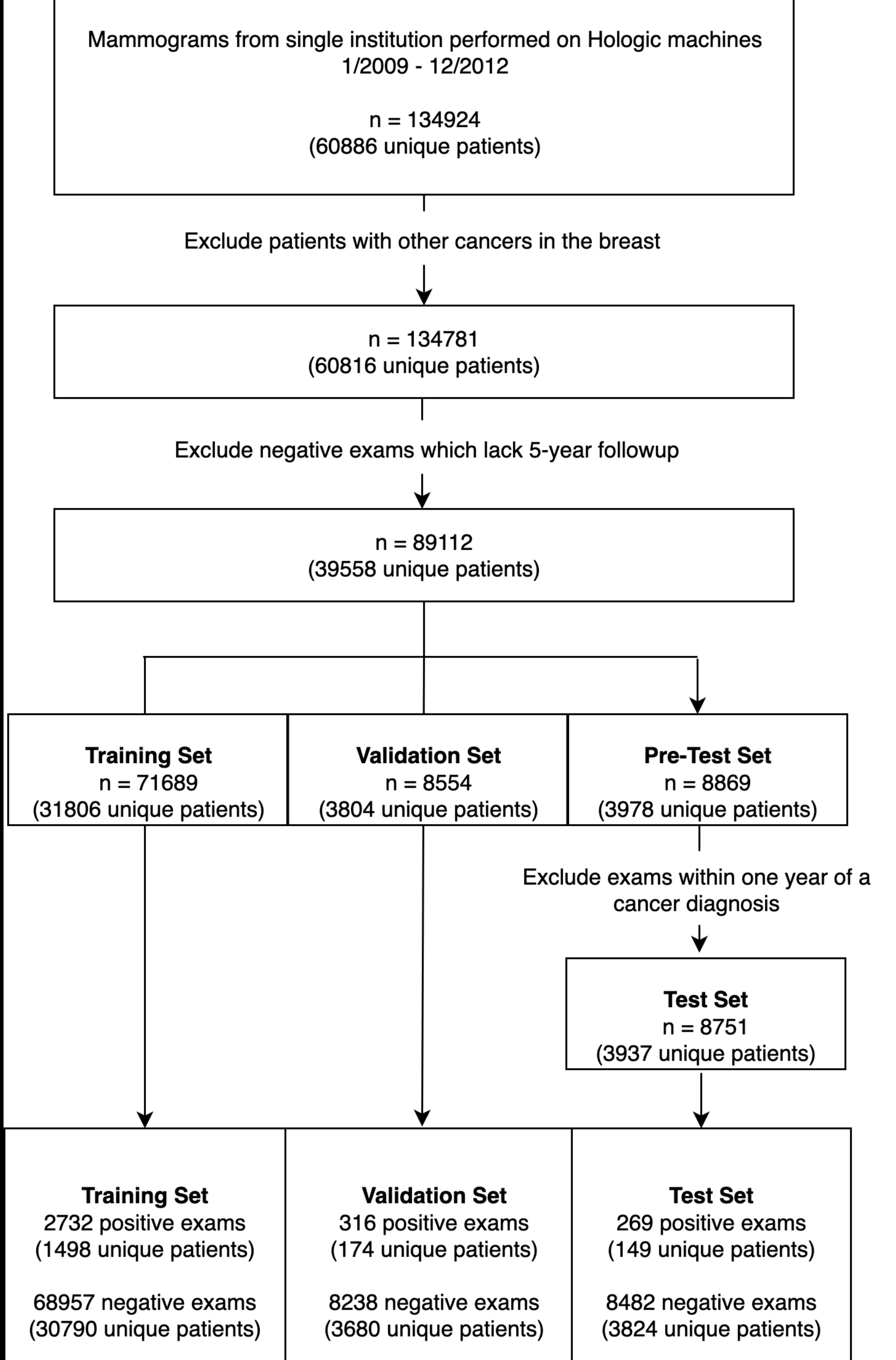
- The plan
    - **Dataset Collection**
    - Modeling
    - Analysis

Dataset Collection

- Consecutive Screening Mammograms
  - 2009-2012
- Outcomes from Radiology EHR, and Partners

5 Hospital Registry

- No exclusions based on race, implants etc.
- Exclude for followup for negatives
- Split into Train/Dev/Test by Patient

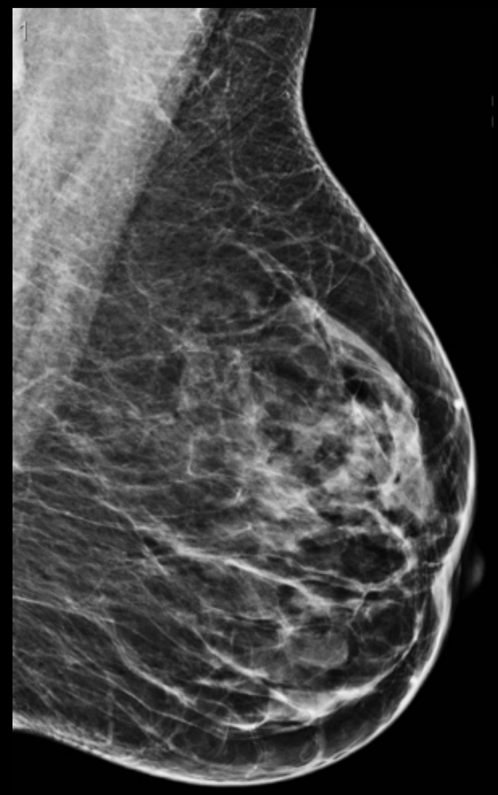


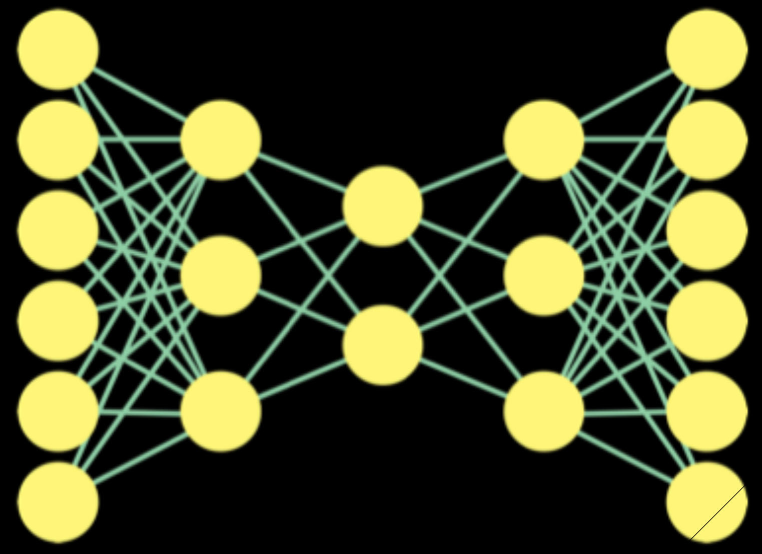
Modeling

- **ImageOnly**: Same model setup as for Triage
- **Image+RF :** ImageOnly + traditional Risk Factors at last layer trained jointly

Analysis: Objectives

- Is the model discriminative across all populations?
  - Subgroup Analysis by **Race**, **Menopause Status, Family History**
- How does this relate to classical approaches?





5 Year Breast Cancer Risk

Training Set:

Patients: 30,790
Exams: 71,689

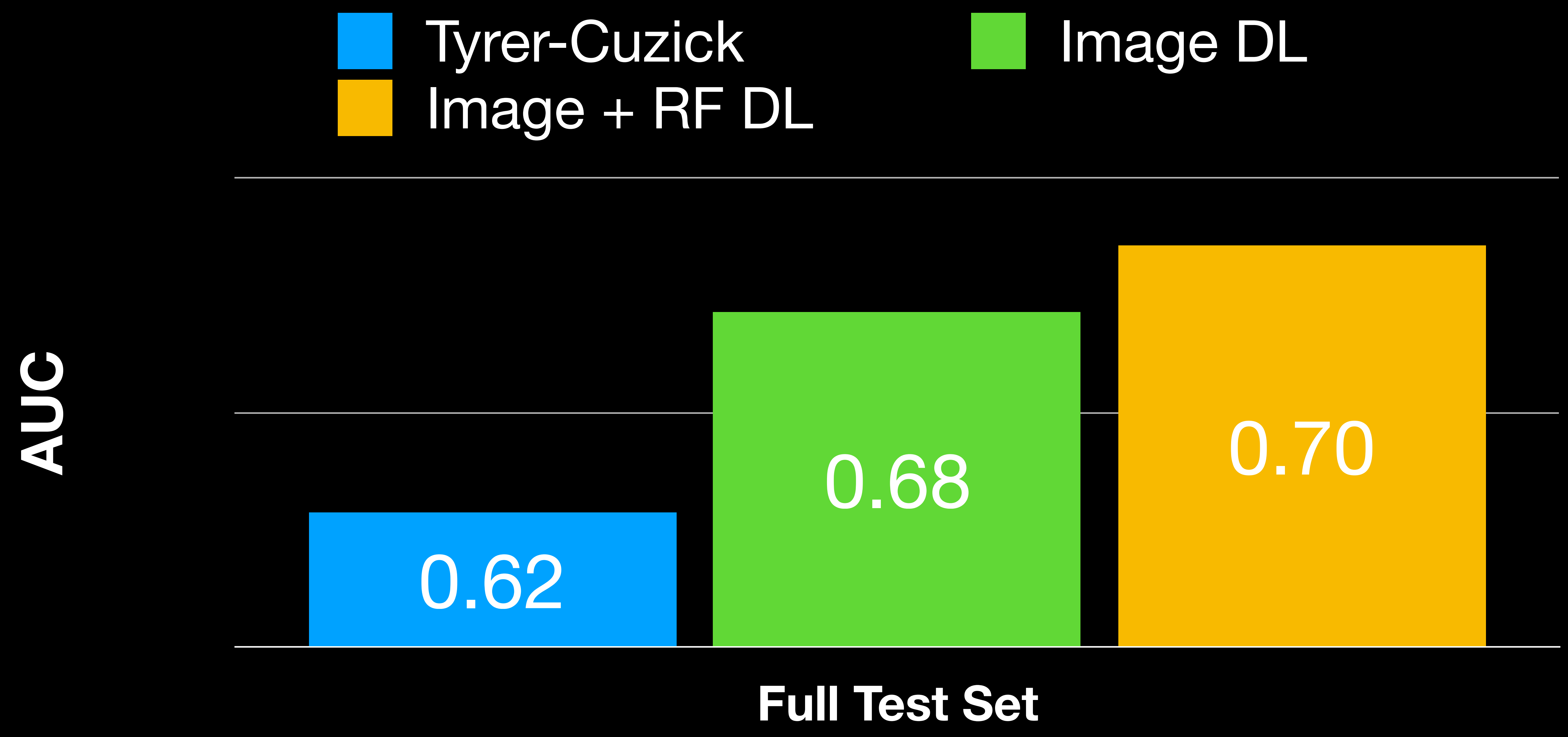
No Exclusions

Testing Set:

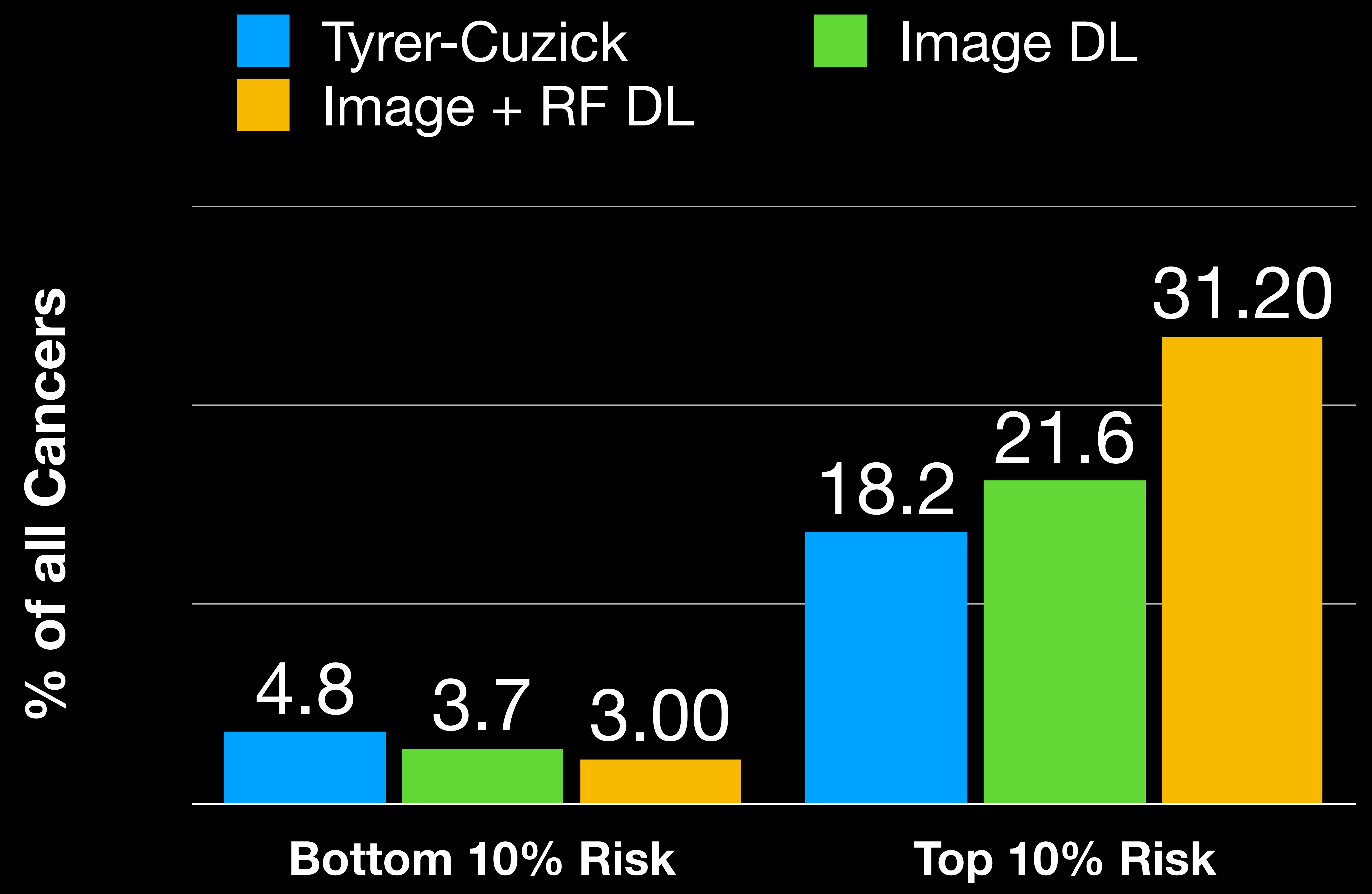
Patients: 3,937
Exams: 8,751

Exclude Cancers within 1 Year of mammogram

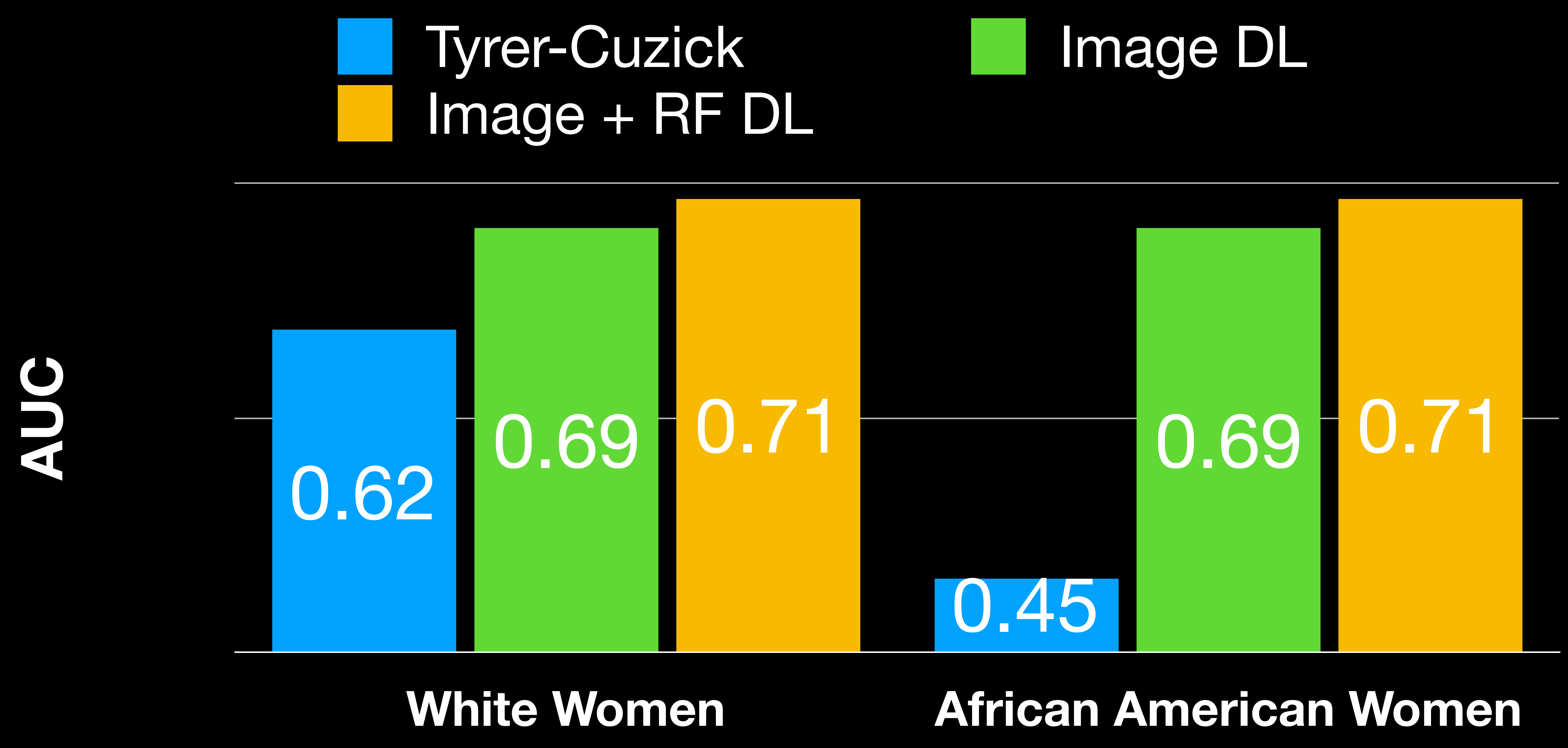
Performance



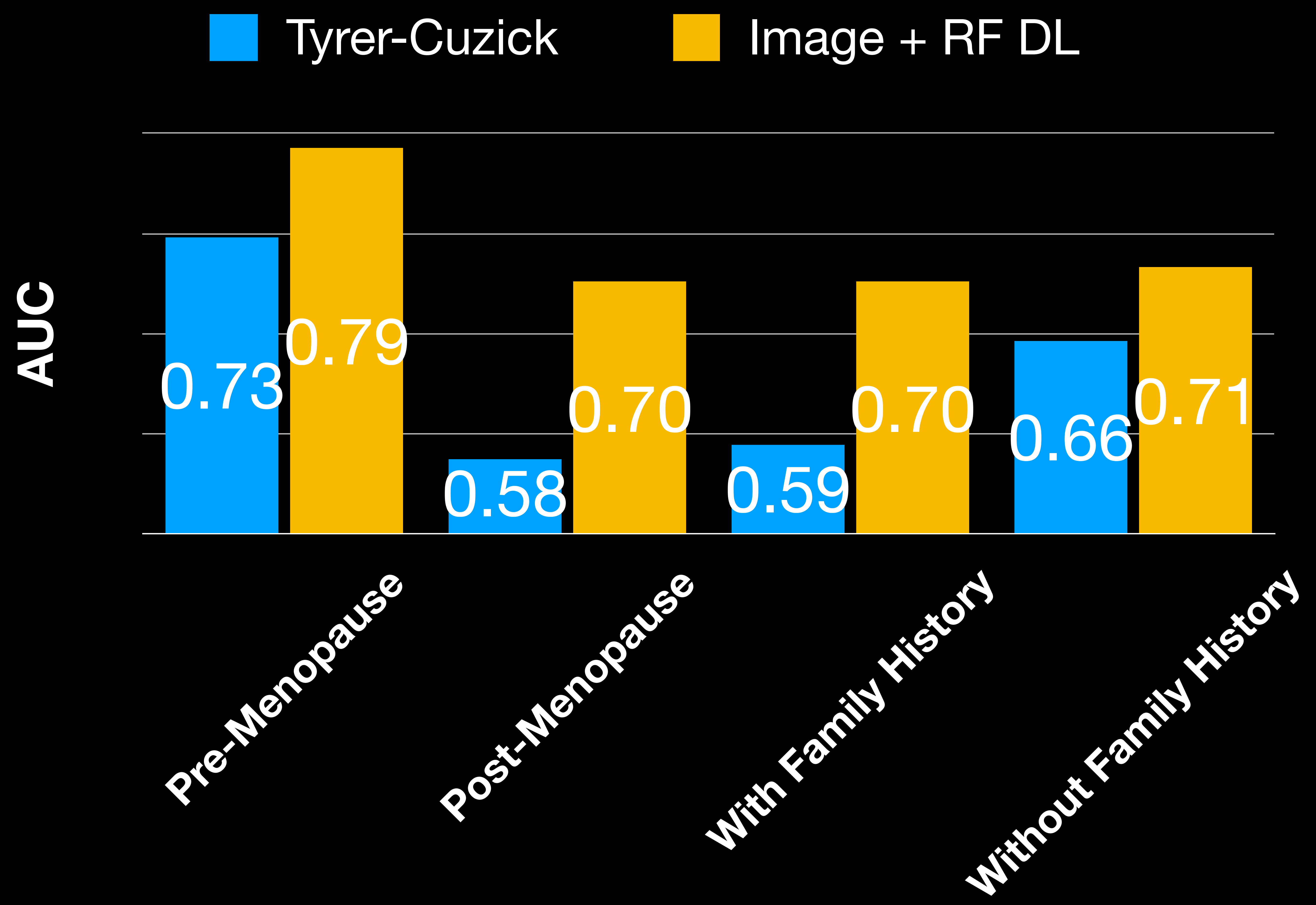
Performance



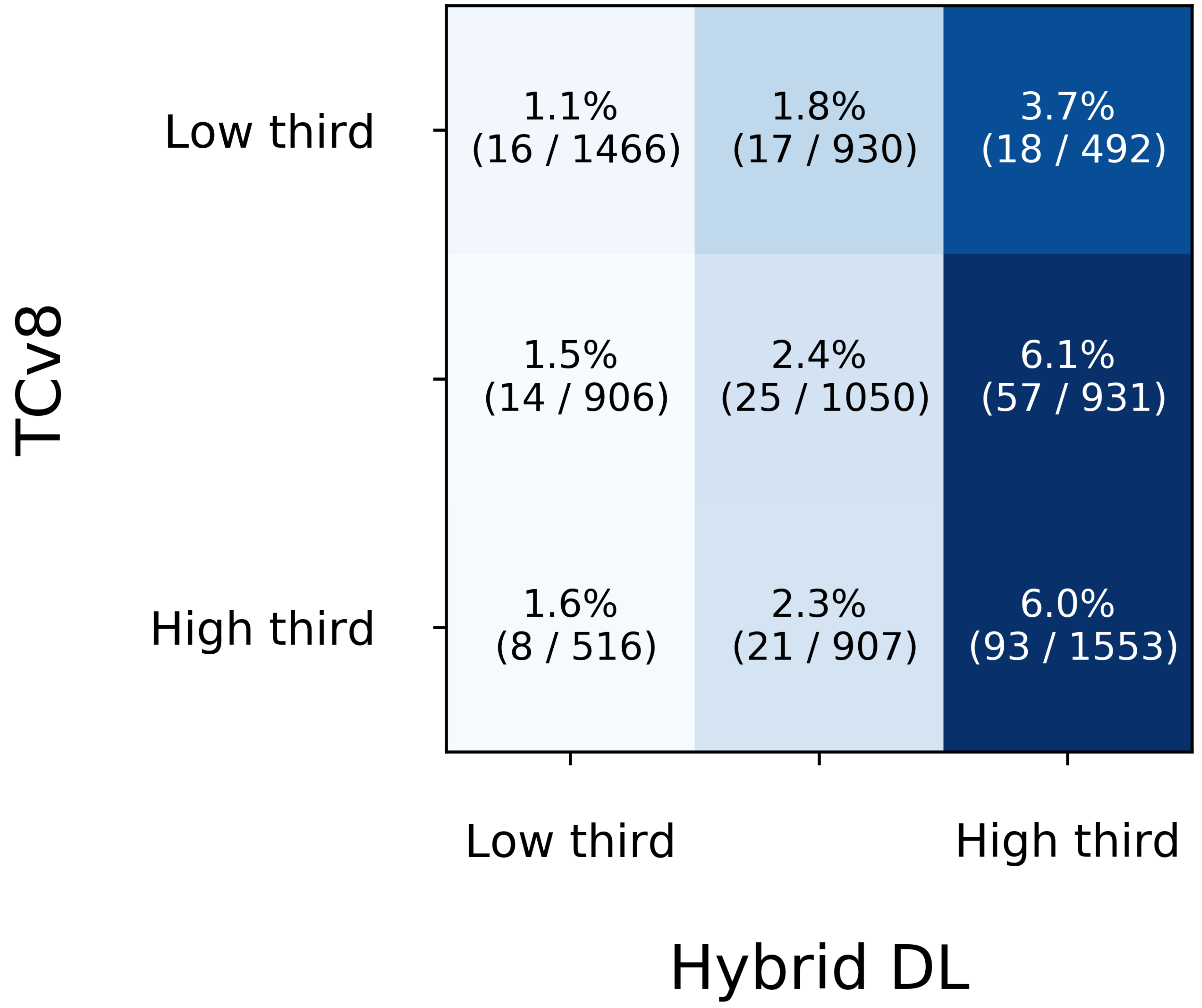
Performance

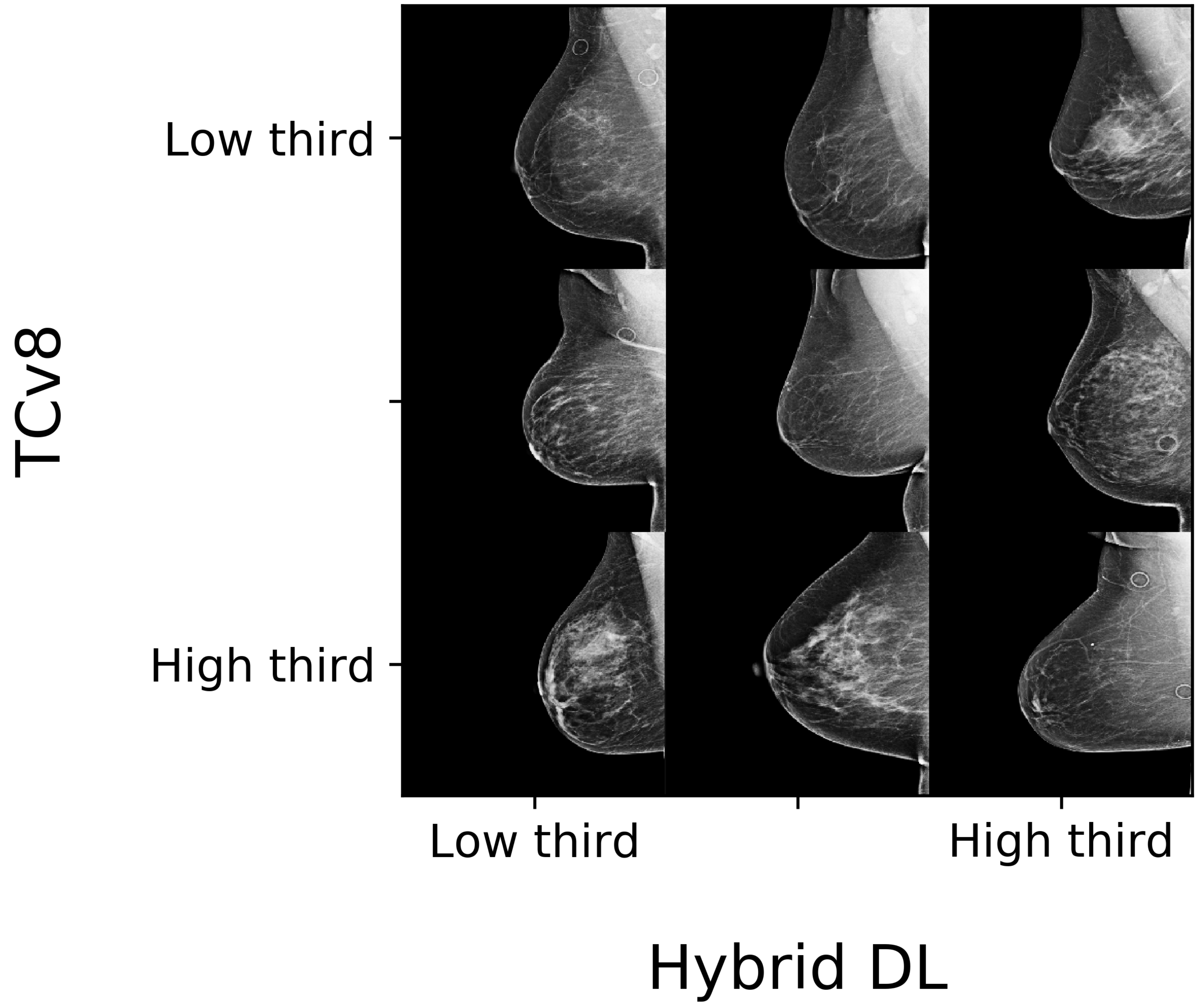


Performance

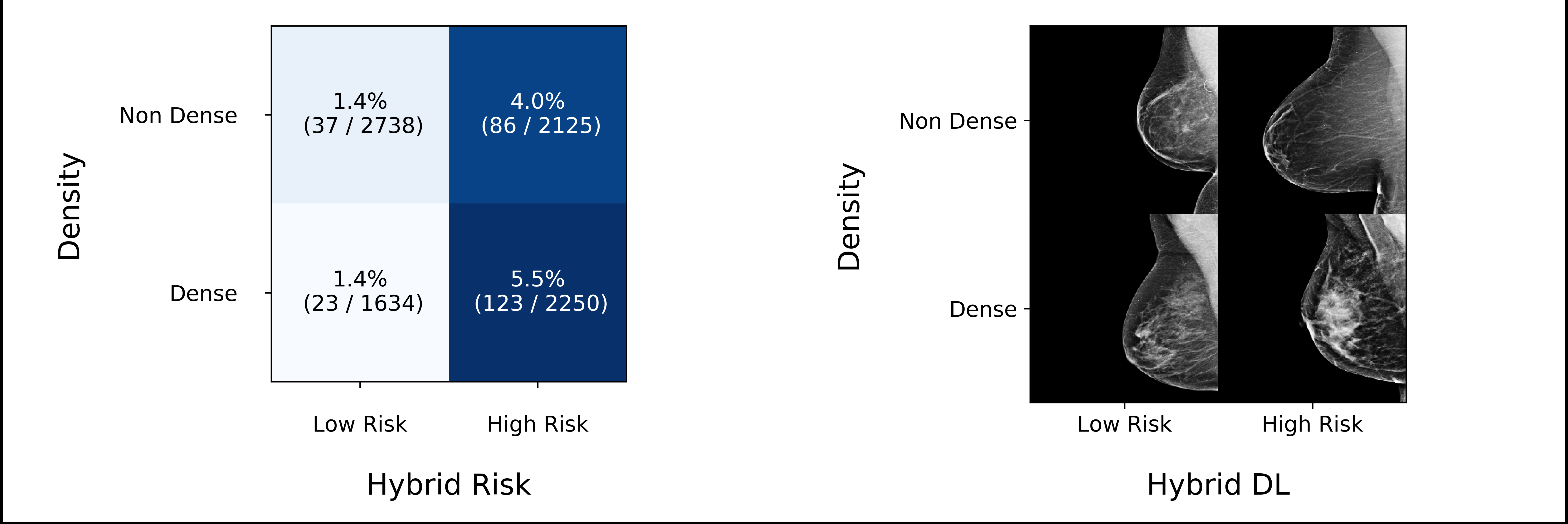


Performance





Performance



Next Step: Clinical Implementation

Agenda

- Interpreting Mammograms
  - Cancer Detection and Triage
  - Assessing Breast Density
- Assessing Breast Cancer Risk
- **How to Mess up**
- How to Deploy

How to Mess Up

- The many ways this can go wrong:
  - **Dataset Collection**
  - Modeling
  - Analysis

How to Mess Up: **Dataset Collection**

- Enriched Datasets contain nasty biases
    - **Story**: Emotional Rollercoaster in Shanghai
        - Dataset with all Cancers collected first.
        - Negatives collected consecutively from 2009-2016
- Use old images (Film mammography) or datasets with huge tumors.
- Use a dataset without tumor registry linking.
- Is your dataset reflective of your actual use-case?

How to Mess Up: **Modeling**

- Assume the model will be Mammography Machine invariant
    - Now exploring conditional-adversarial training…

How to Mess Up: **Analysis**

- Only Test your model on White women and exclude *inconvenient* cases
    - Common standard in classical risk models; can’t assume model will transfer.
- Assume ***reader study = clinical implementation***

Agenda

- Interpreting Mammograms
  - Cancer Detection and Triage
  - Assessing Breast Density
- Assessing Breast Cancer Risk
- How to Mess up
- **How to Deploy**

How to Deploy?