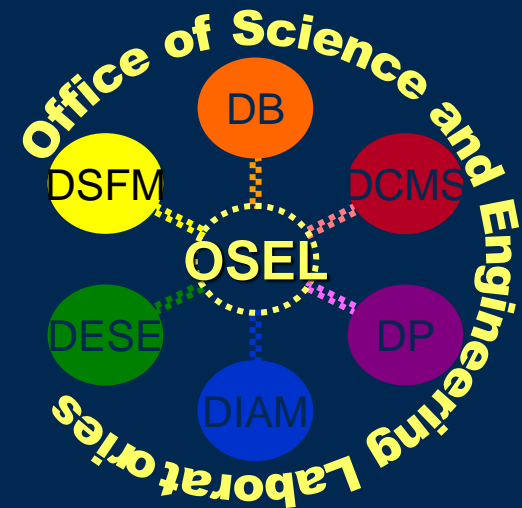


# Assessment of computer-aided image analysis devices: Experience gained from CAD in radiology



Berkman Sahiner, PhD

USFDA/CDRH/OSEL

Division of Imaging and Applied Mathematics

# OUTLINE

- **What are CAD devices?**
  - CADe
  - CADx
- **Types of CAD assessment**
  - Standalone
  - With clinicians
    - Sequential
    - Concurrent
    - Interactive
- **Datasets**
- **Reference standard**
- **Scoring / Labeling**
- **Performance measures**
  - Sensitivity / specificity
  - Area under the ROC curve
- **Summary**

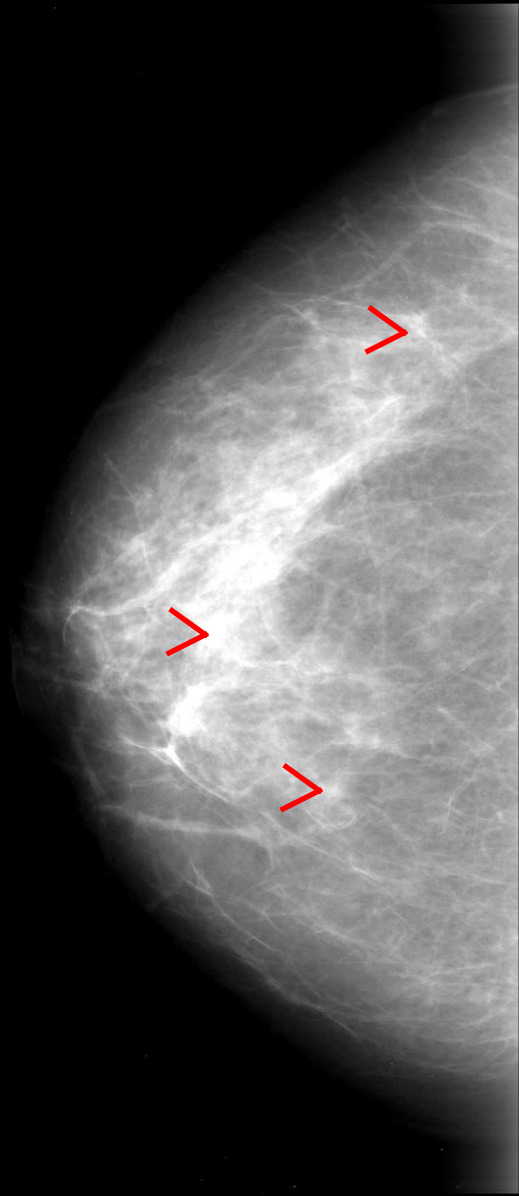
# WHAT ARE CAD DEVICES?

- **Computer-aided detection (CADe) and computer-aided diagnosis (CADx)**

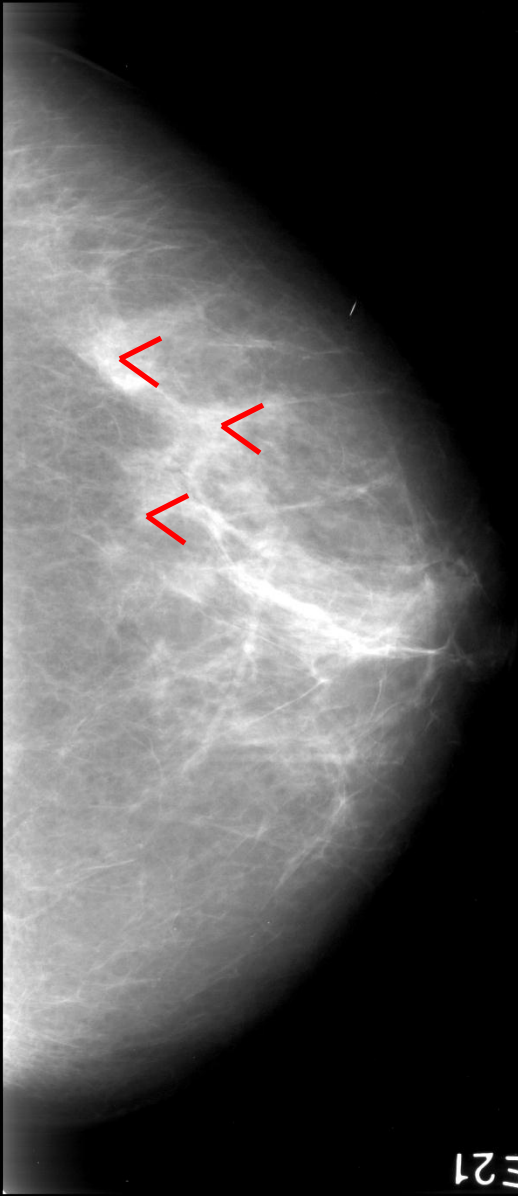
# CADe DEVICE IN RADIOLOGY

- **A computerized system that**
  - identifies portions of an image in order to reveal abnormalities during interpretation of patient radiology images by the clinician
  - Detection of breast cancer on mammograms
  - Detection of lung nodules on thoracic CT images
- **Commercial devices have been around since 1998**

RCC



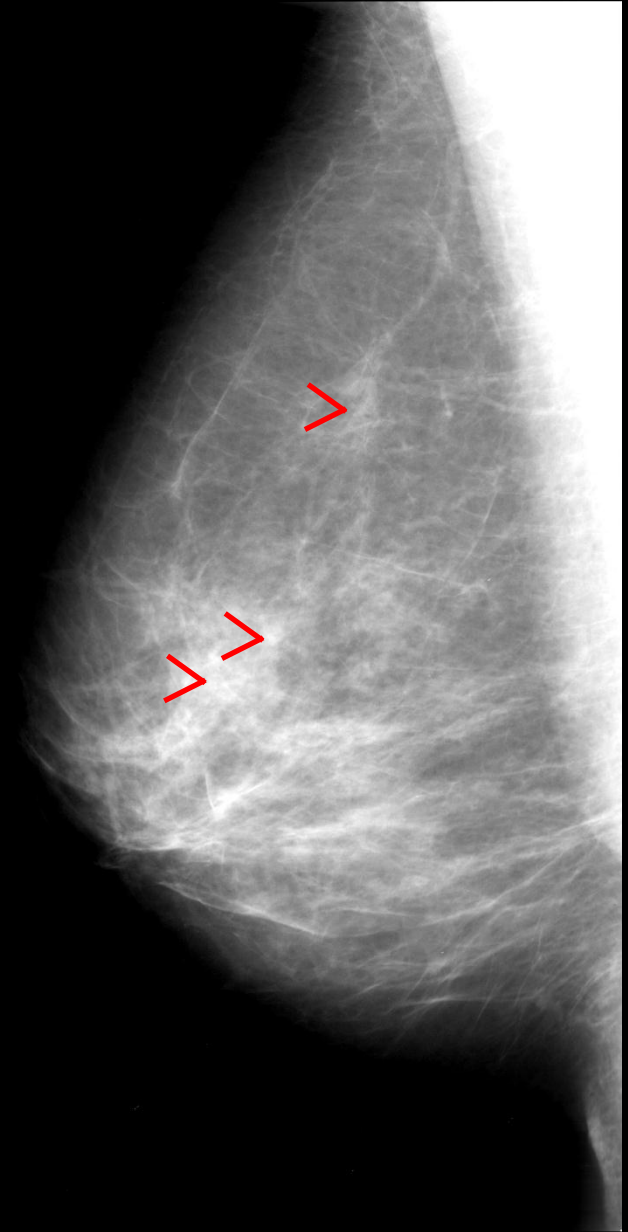
LCC



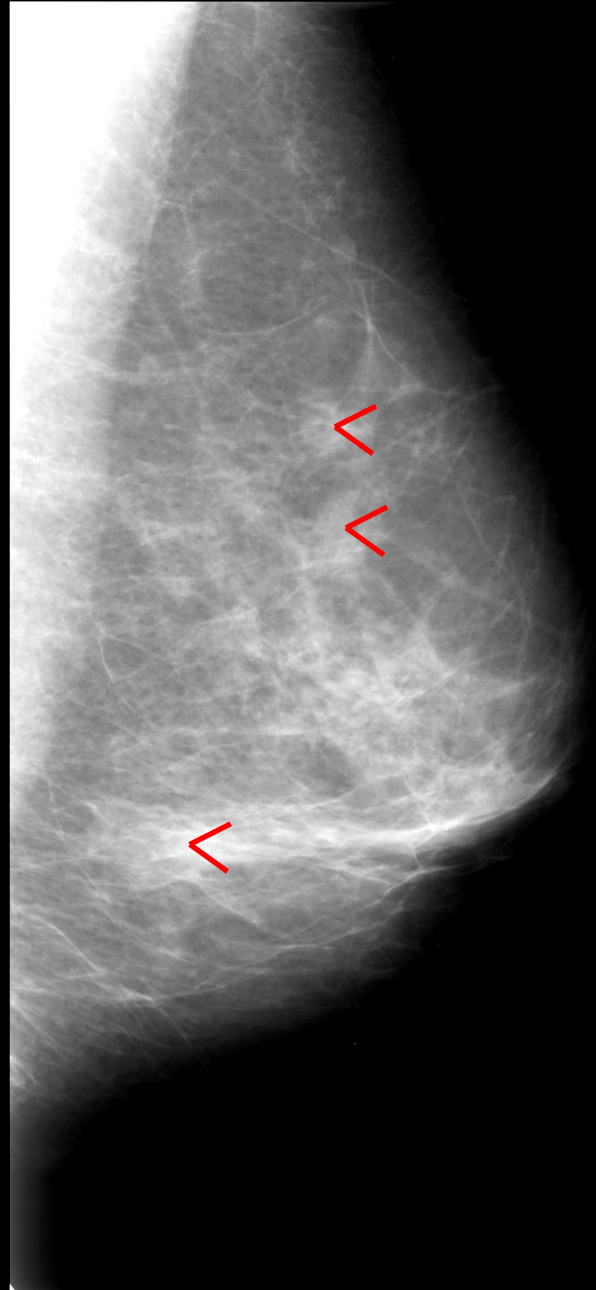
E21

RMLO

E13



LMLO



E30



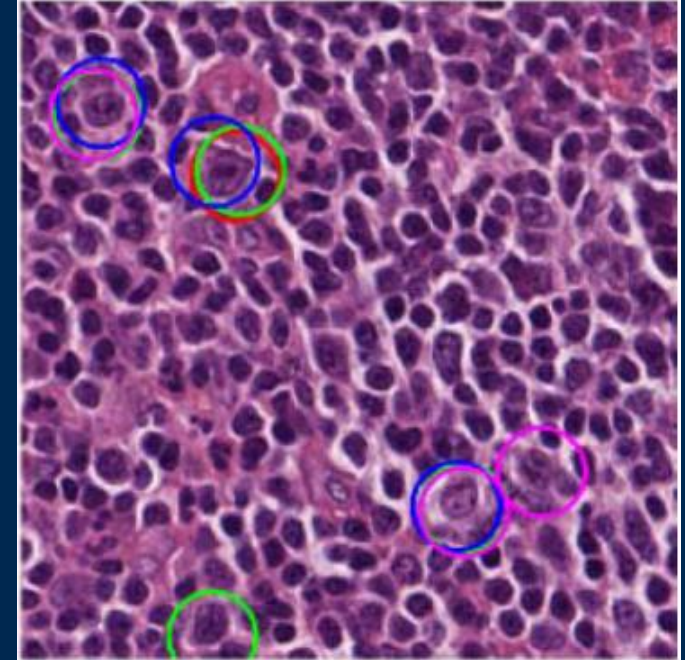
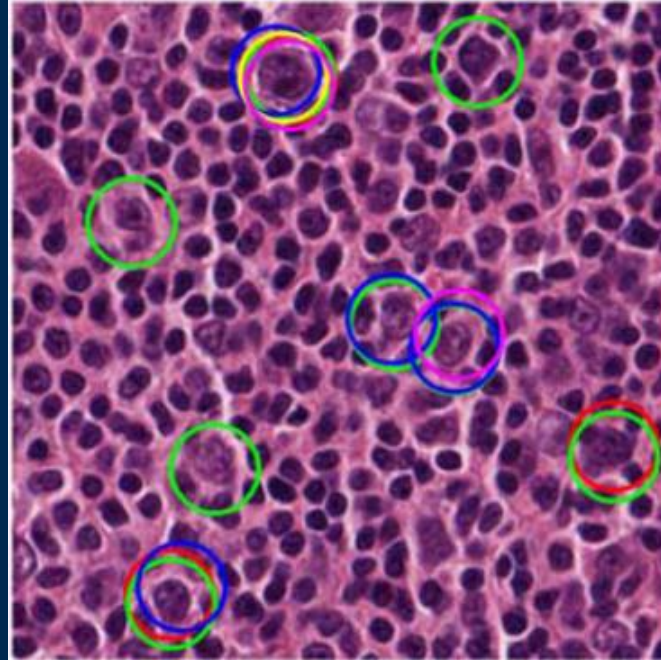
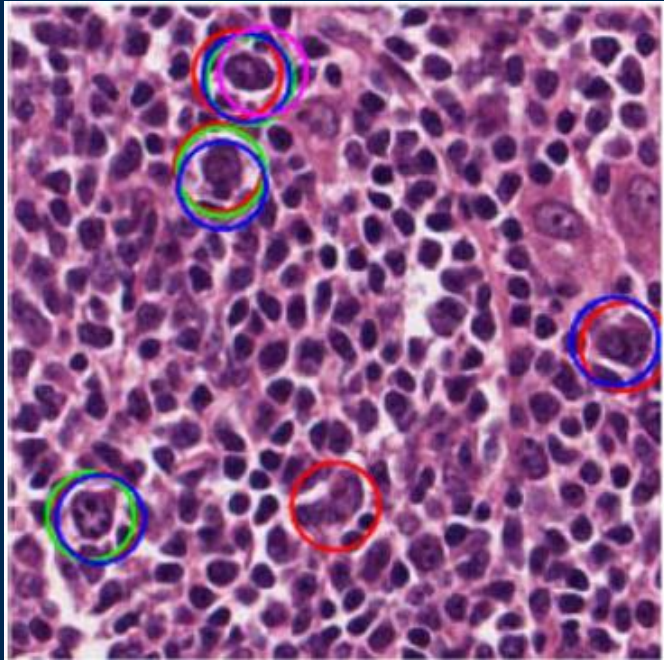
Zoom In

Zoom Out

Prev Image

Next Image

# DETECTION OF CENTROBLASTS FOR FOLLICULAR LYMPHOMA GRADING



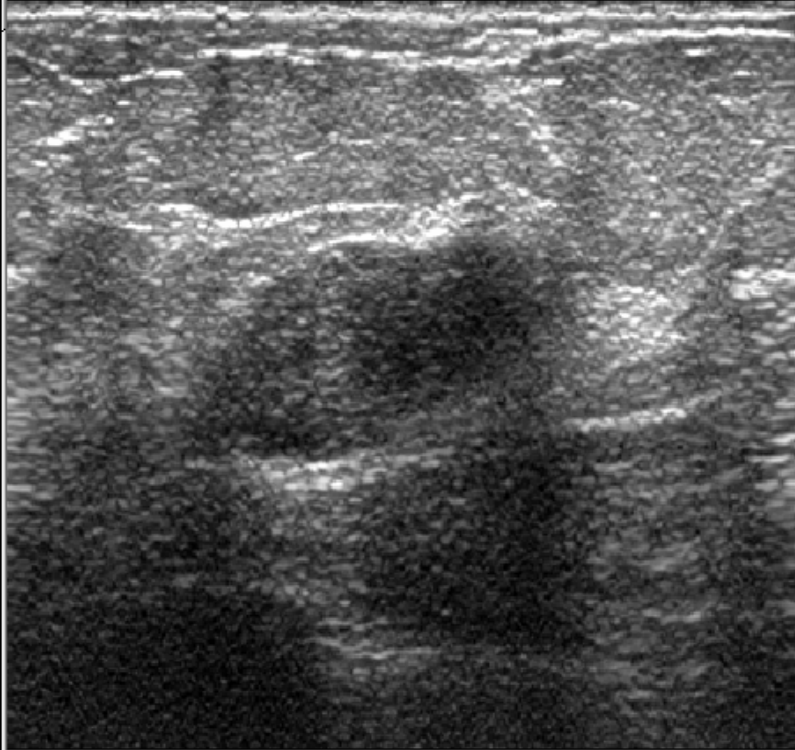
**Reference standard by five expert hematopathologists**

**Sertel et al., "Computer-Aided Detection of Centroblasts for Follicular Lymphoma Grading Using Adaptive Likelihood-Based Cell Segmentation," IEEE Trans. Biomed. Eng., 2010.**

# CADx DEVICE IN RADIOLOGY

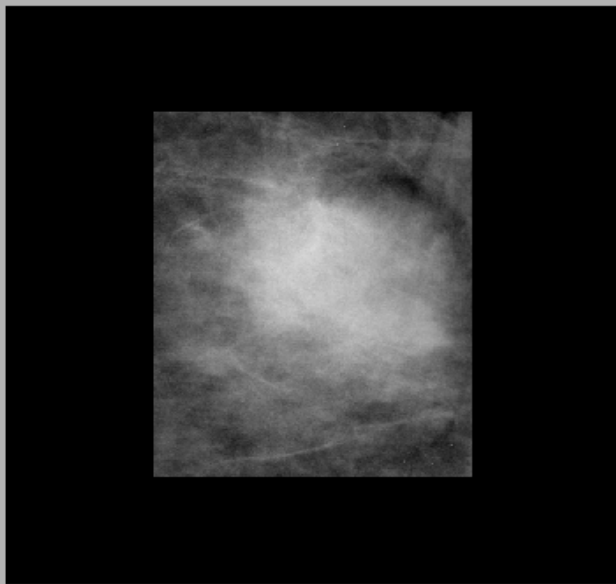
- **A computerized system that may**
  - provide an assessment of disease
  - specify disease type
  - specify severity, stage, or intervention recommended

for the clinician, based on radiology images

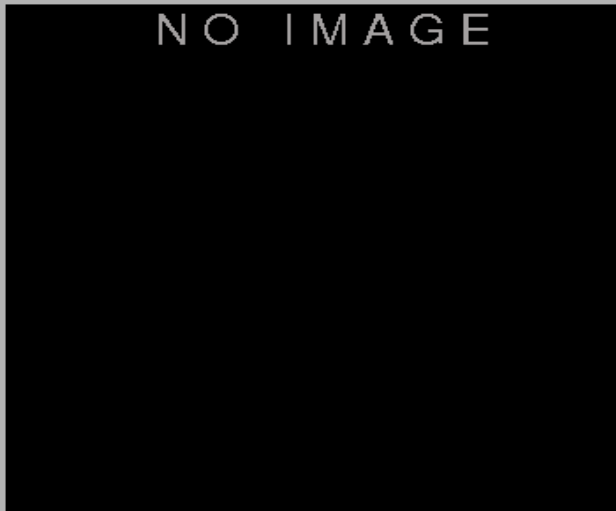
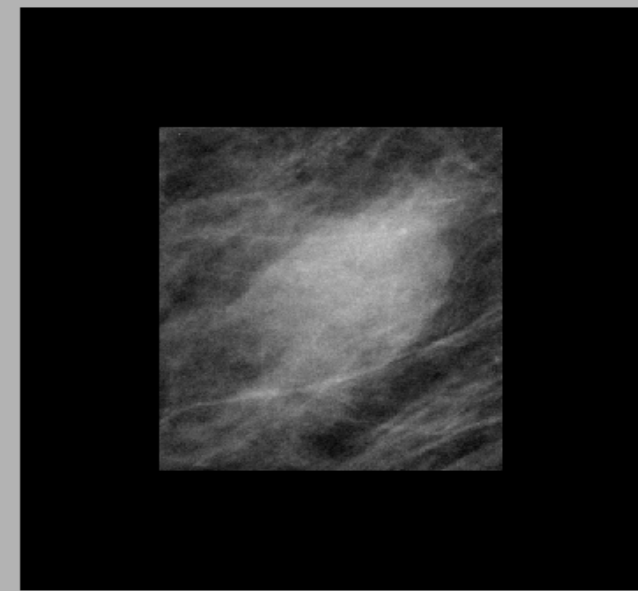


Case: m0911 Image 68

CC view of the left breast of m0911

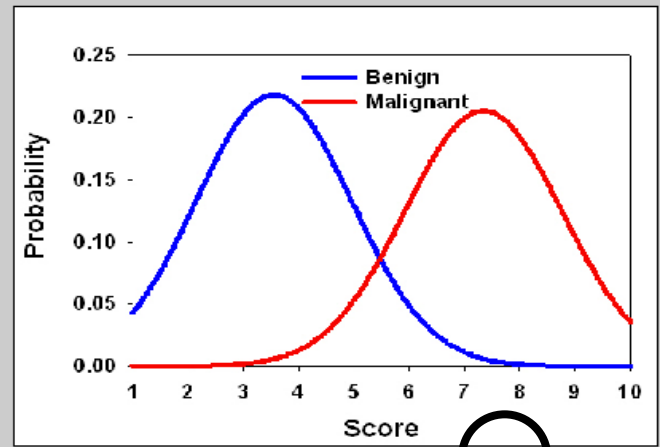


MLO view of the left breast of m0911



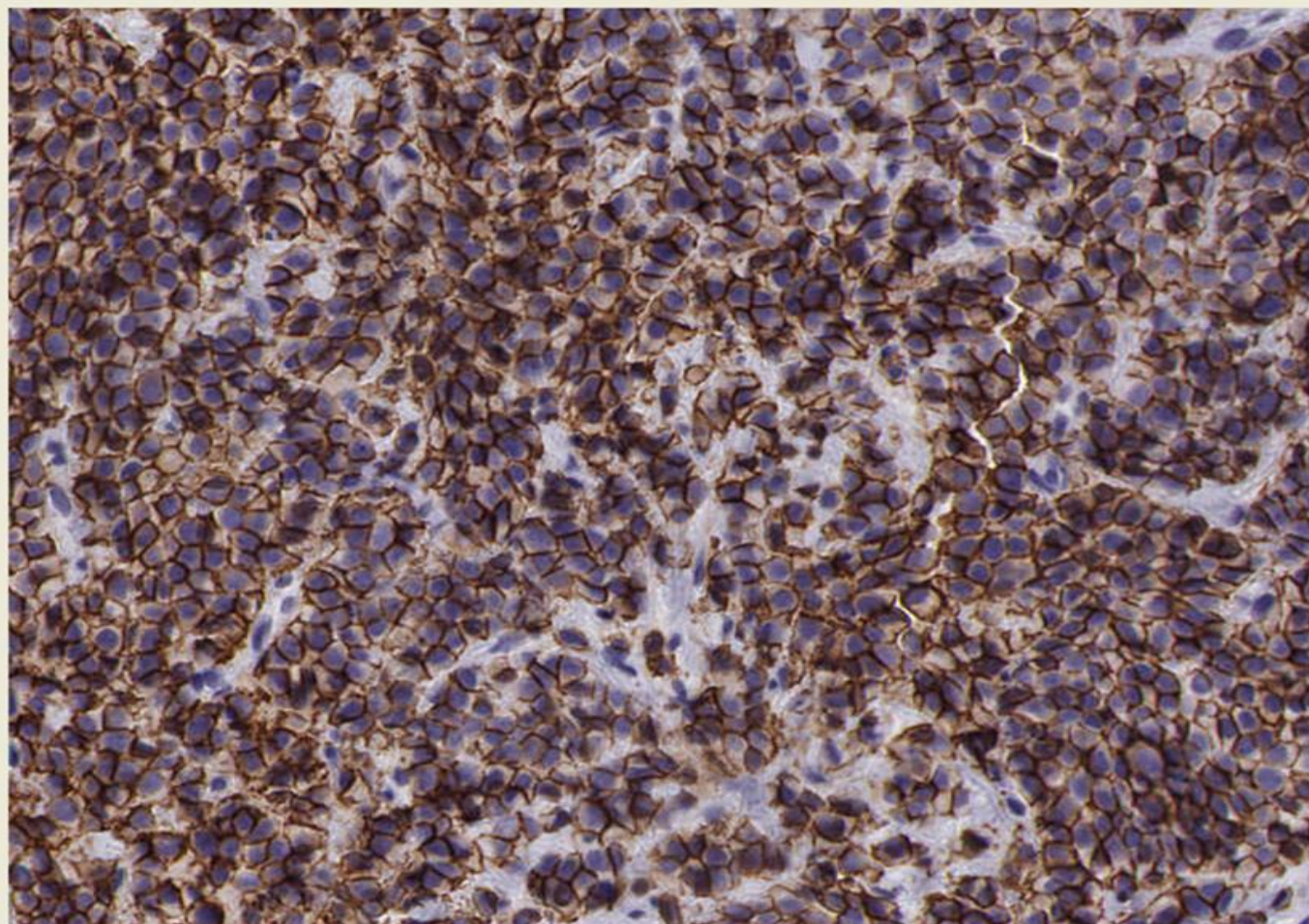
<< < || > >>

Classifier Score Distribution for Current Case Set

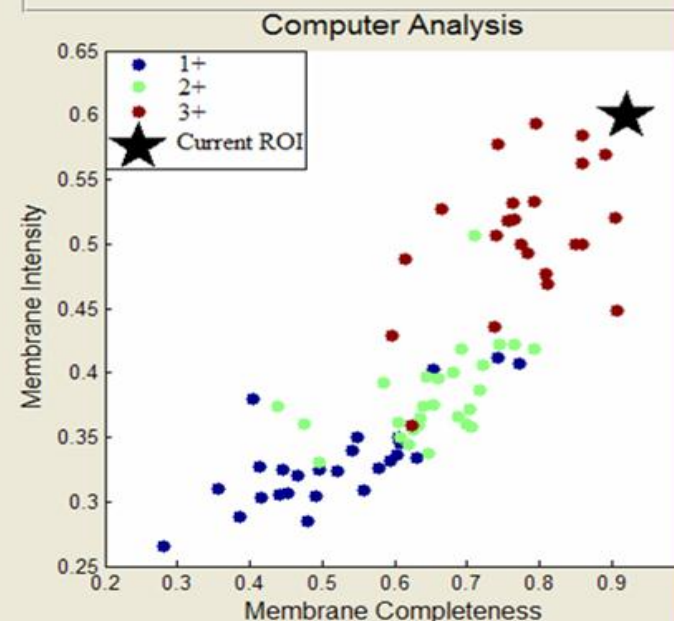


Classifier Score m0911 -> 3

# HER2 Assessment: computer-aided



ROI 1/500



The graph depicts computer score of slides and where the ROI viewed lays. Slides of the same score tend to group, however, the score is based on a pathologist score and as such, is not always accurate

1+

2+

3+

Score

3

Rating

97

Pause

Next

# COMPUTER TRIAGE DEVICE

- **Intended to reduce or eliminate any aspect of clinical care provided by a clinician**
  - e.g, the output indicates that the patient is normal and does not require clinician's interpretation
- **Cytological image analysis**
  - Computer-assisted pap smear analysis

# CAD SYSTEM ASSESSMENT

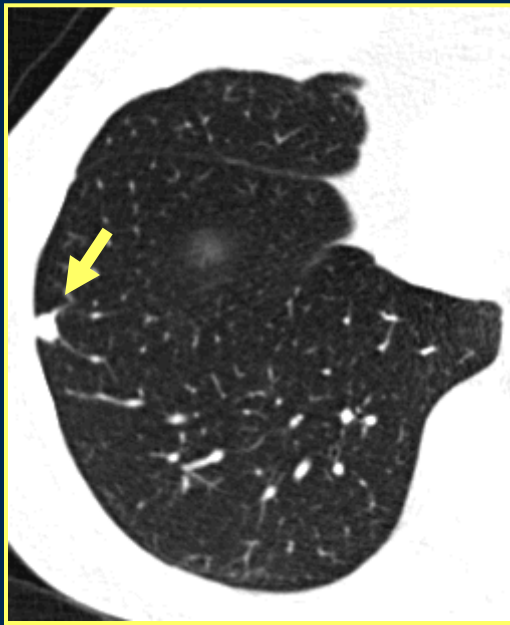
- **Measure the performance of your system**
  - Informs your peers, users, regulators, scientific community, and yourself
- **If you can't assess it, you will not know how to improve it**

# TYPES OF ASSESSMENT

- **Standalone**
  - Evaluate the performance of the computer system only
- **The effect on clinicians**
  - In a controlled environment:
    - Laboratory observer study
  - In practical, daily use
    - Clinical use study

# BOTH STANDALONE ASSESSMENT AND EFFECT ON CLINICIANS ARE IMPORTANT

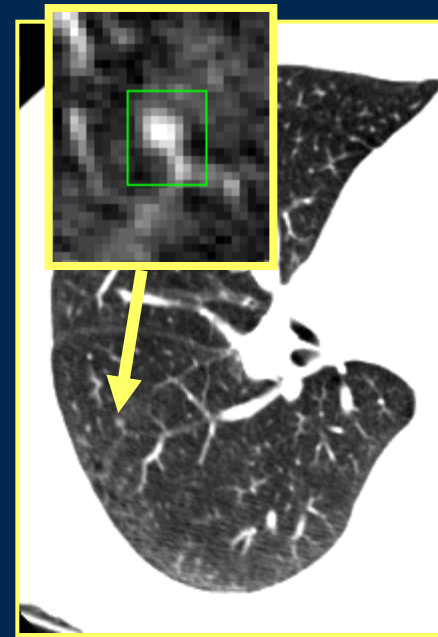
- Why is standalone performance by itself inadequate?
  - Not all TP and FP marks are equal



**CAD1: Detect**  
**CAD2: Miss**



**CAD1: Detect**  
**CAD2: Miss**



**CAD1: Miss**  
**CAD2: Detect**

# EFFECT OF CAD ON CLINICIANS

- **Differences between laboratory observer study and clinical use study**
  - Difference in mindset
  - Availability of other clinical data and images
  - Prevalence
  - Case difficulty spectrum

# EFFECT OF CAD ON CLINICIANS

- **Clinical use studies are**
  - Ideally representative of the true performance
  - Difficult to conduct, esp. before dissemination
  - Premature before a CAD device finds its niche
  - Costly
- **Earlier phase**
  - Laboratory observer study
- **Advanced phase**
  - Clinical use study

# INTERPRETATION PARADIGMS

- **Sequential**
  - Clinician interprets first, followed by CAD results or prompts
- **Concurrent**
  - CAD results are displayed when the clinician starts interpretation
- **Interactive**
  - CAD results are displayed only for cases or locations indicated by clinician

# SEQUENTIAL INTERPRETATION W/O CAD

Name: 0159  
Registration Number: 159  
Case Number: 159  
Exam Date: 20021011  
Nodule Number: 1

Viewing image 54 of 241 · p0159\_20021011\_s2\_054.dcm

Disable ROI Nodule Numbers  Hide ROI Boxes

Nodule Navigation

Nodule Information  
Check if necessary:  Non-Nodule  < 3mm Nodule  Nodule Not Seen

Likelihood of Being a True Nodule (%)  
0 10 20 30 40 50 60 70 80 90 100%

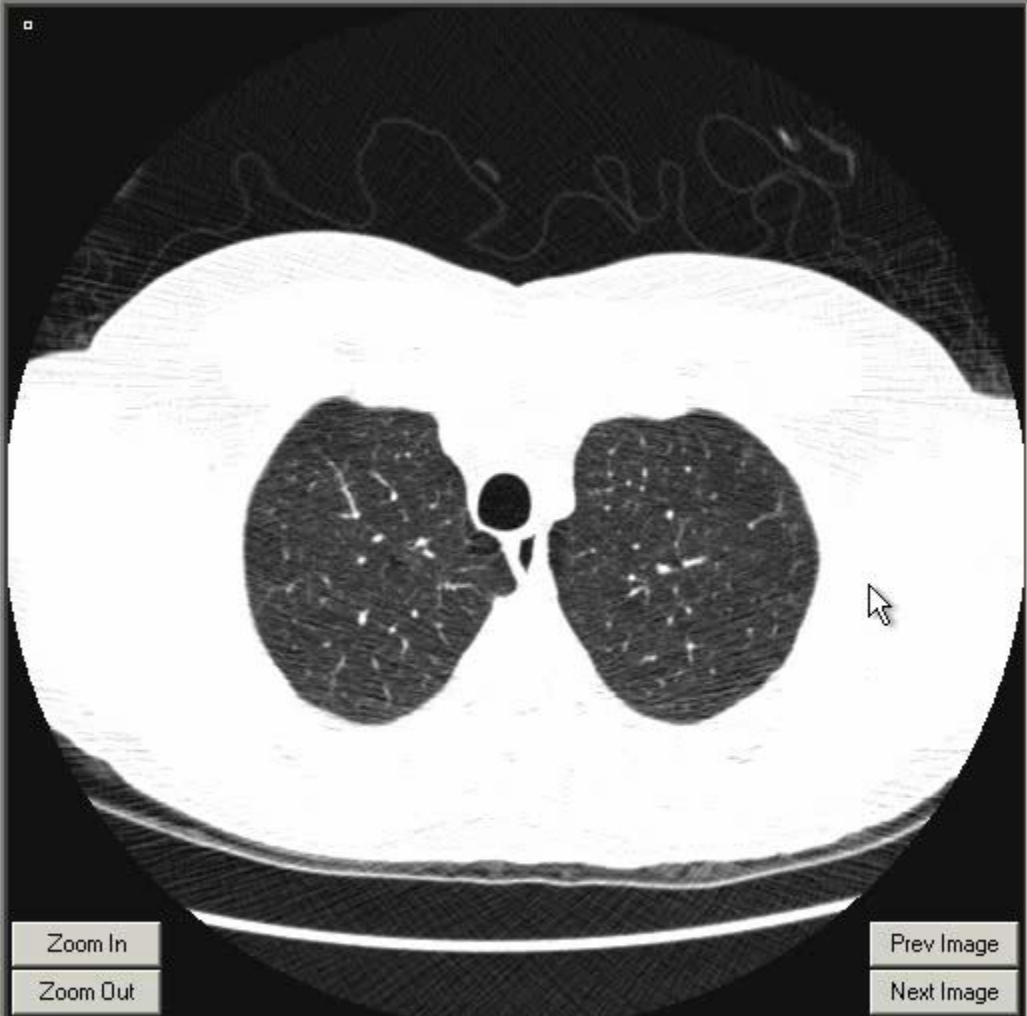
Likelihood of Malignancy (%)  
0 10 20 30 40 50 60 70 80 90 100%

Nodule Subtlety (5 = Most Subtle)  
 1  2  3  4  5

User Detected:  Yes  No

CAD Detection Coincides:

Additional Comments:



1 Image Series Number: 2 Image Slice Thickness (mm): 1.250

# SEQUENTIAL INTERPRETATION WITH CAD

Name: 0159  
Registration Number: 159  
Case Number: 159  
Exam Date: 20021011  
Nodule Number: 1

Viewing image 219 of 241 - p0159\_20021011\_s2\_219.dcm

Disable ROI Nodule Numbers  Hide ROI Boxes

Nodule Navigation

Nodule Information  
Check if necessary:  Non-Nodule  < 3mm Nodule  Nodule Not Seen

Likelihood of Being a True Nodule (%)  
0 10 20 30 40 50 60 70 80 90 100%  
80


Likelihood of Malignancy (%)  
0 10 20 30 40 50 60 70 80 90 100%  
20

Nodule Subtlety (5 = Most Subtle)  
 1  2  3  4  5

User Detected:  Yes  No

CAD Detection Coincides:

Additional Comments:



1 Image Series Number: 2 Image Slice Thickness (mm): 1.250

# TEST DATASET

- **Must be independent of the training data set used at any stage of development**
- **Should include the range of abnormalities seen in practice desired for CAD assistance**
- **Should include a set of cases free of the disease of interest**
- **Should be large enough for adequate statistical power to demonstrate study objectives**

# ENRICHMENT

- **Low prevalence disease**
  - Enhance with cases containing disease
    - Will not affect sensitivity, specificity, area under the ROC curve
    - In an observer study, may affect the clinician's behavior

# SPECTRUM OF DIFFICULTY

- **Spectrum of difficulty for test cases versus spectrum of difficulty for true population:**
  - If different, test results may be biased
- **Bias may be acceptable if**
  - Comparing two modalities
    - e.g., clinicians' performance with and without CAD
  - Both modalities are affected similarly by spectrum bias

# STRESS TEST

- **Study differences between competing modalities using cases selected to challenge those differences**
- **Example in CADe**
  - Exclude obvious cases because they will be detected both with and without CAD

# REFERENCE STANDARD

- **Ideally, independent of the modality being studied**
- **Example:**
  - Mammography CAD
    - Cancer cases: Biopsy
    - Normal cases: 2-yr follow-up or biopsy
- **Sometimes, this is not possible**
  - Lung nodule detection on thoracic CT scans
  - Pulmonary embolism detection on CT

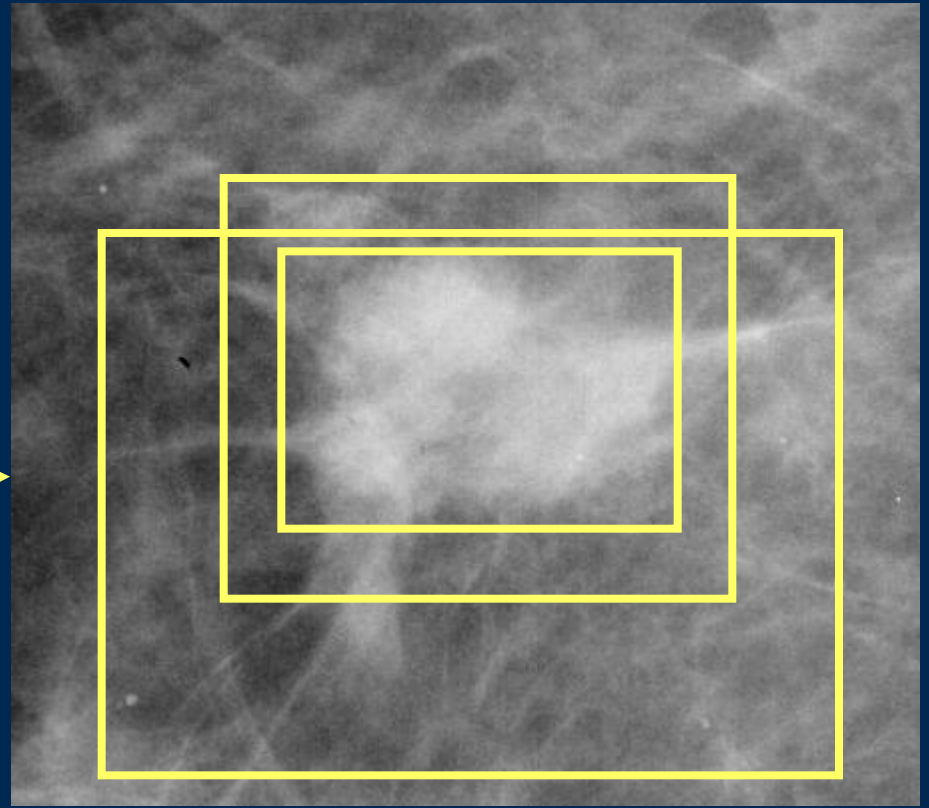
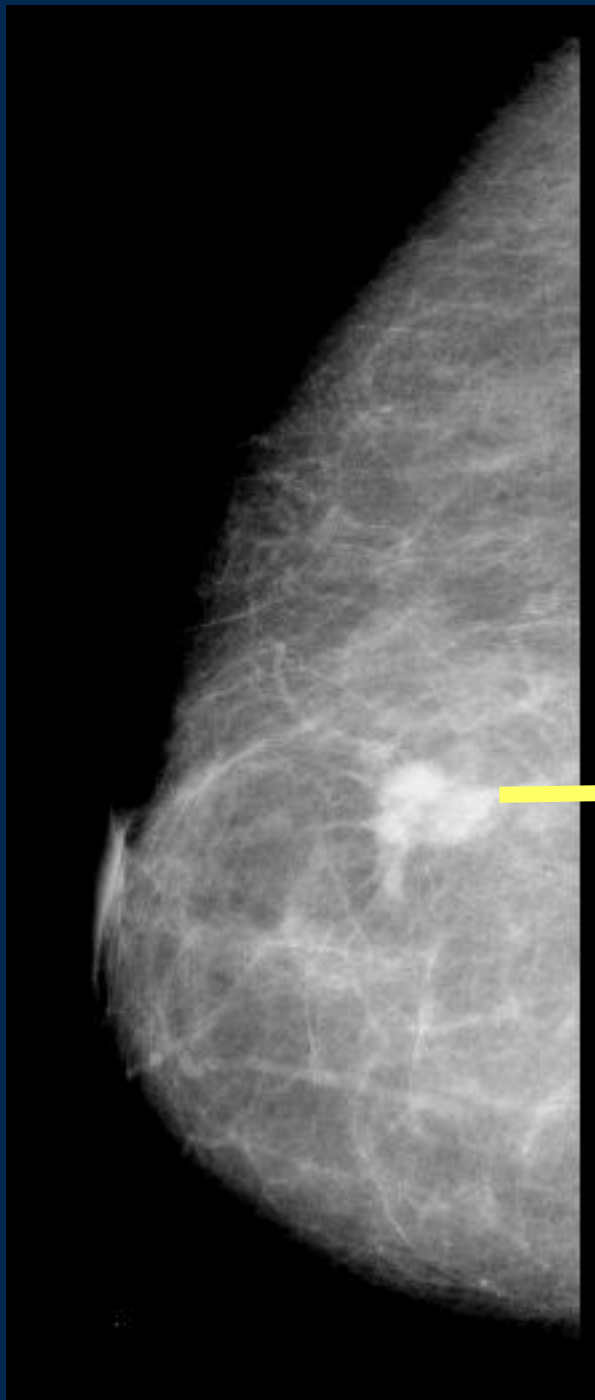
# REFERENCE STANDARD

- **CADx**

- Type of disease
- Diseased/non-diseased

- **CADe**

- Location/extent of disease
- Variability may exist even when using independent source to establish the reference standard
- Important to document



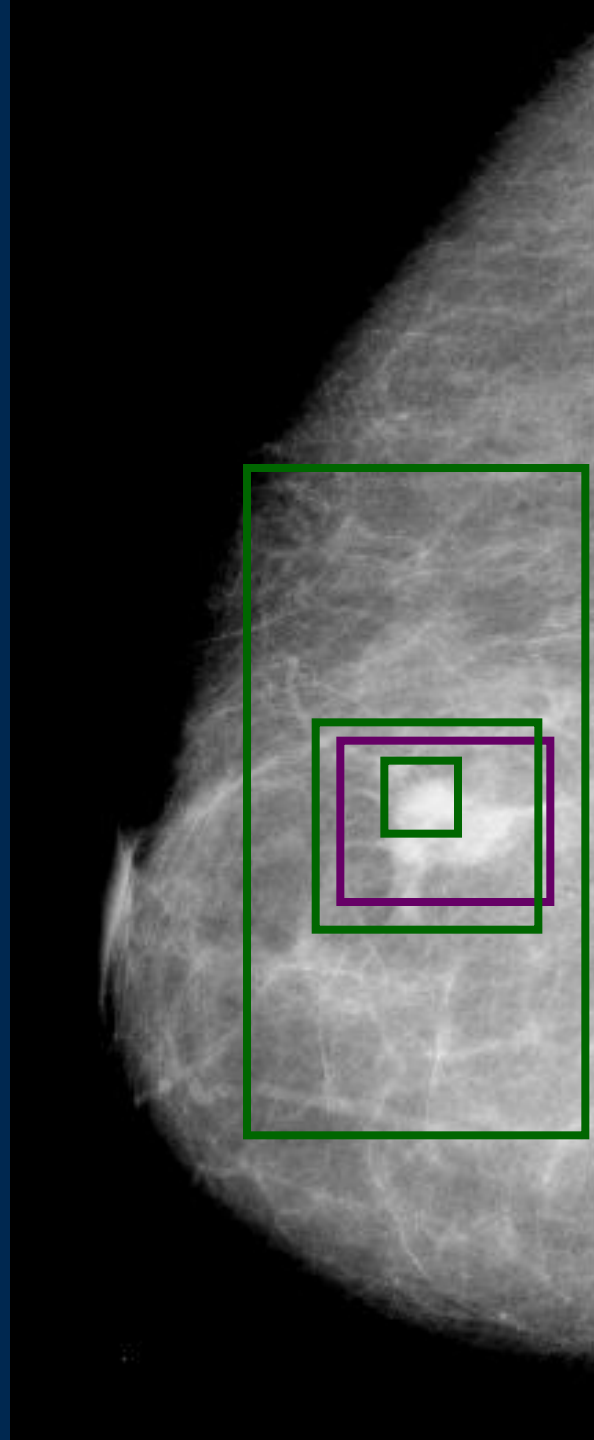
# SCORING / LABELING

- **Correspondence between CAD output (or study clinician's interpretation) and the ground truth**
- **Scoring (labeling) for CADe**
  - Rules for declaring a mark (by a clinician or CAD) as a true-positive or false-positive

# SCORING

- **By a human**
- **Automated:**
  - **Compare computer (or clinician) mark to reference standard mark using an automated rule**
    - **Overlap area divided by ground truth area**
    - **Overlap area divided by union area**
    - **Distance of centroids**

# SCORING



# PERFORMANCE MEASURES: STANDALONE CAD

- **Consider a CAD system intended to classify cases as negative or positive**
- **Such CAD systems often include a classifier which provides an ordinal output**
  - Decision variable
- **Intuitive analysis method:**
  - Threshold classifier output, compute sensitivity and specificity

# SENSITIVITY AND SPECIFICITY

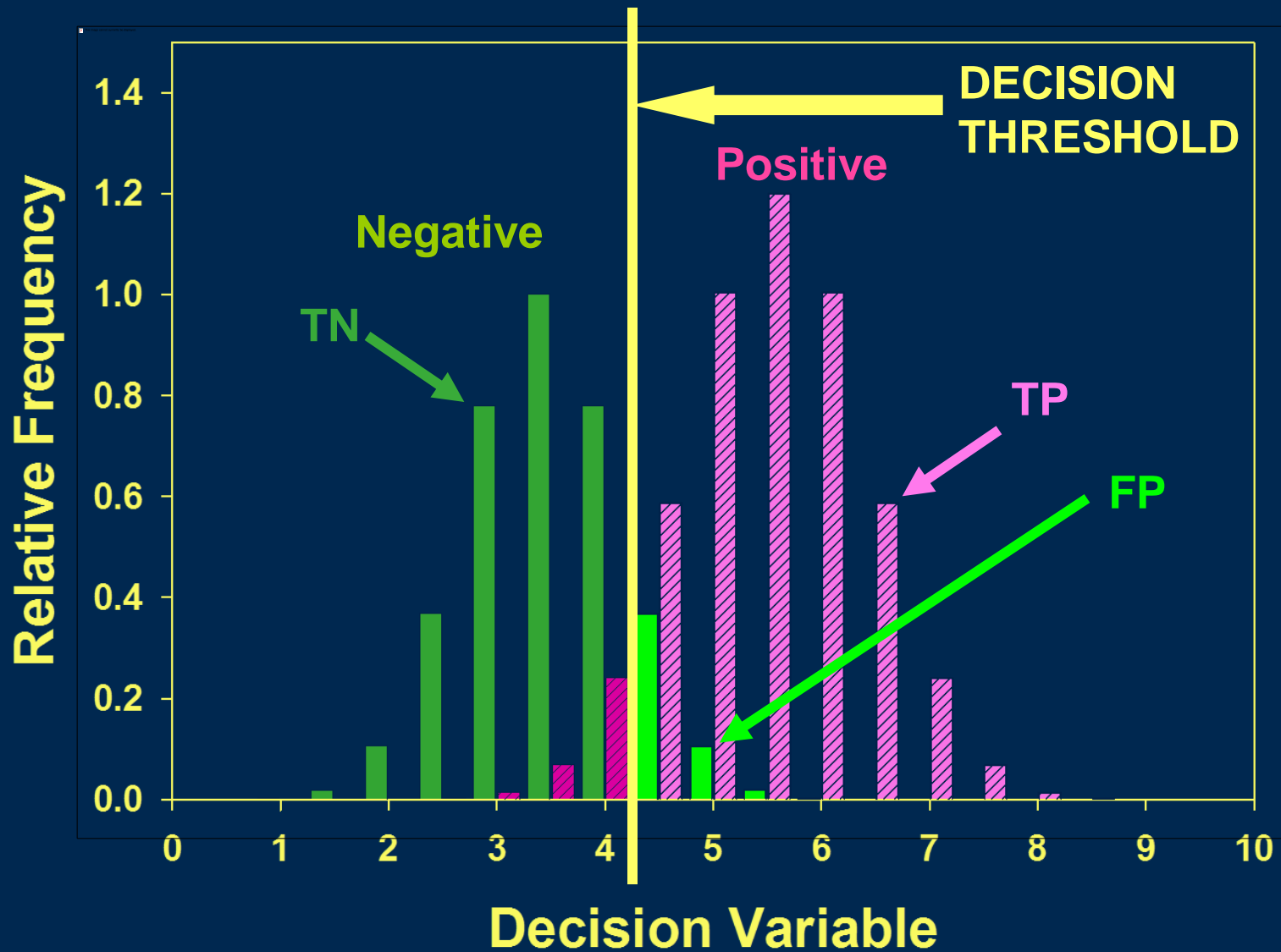
$$\text{Sensitivity} = \frac{\text{Number of cases correctly called positive}}{\text{Total number of positive cases}}$$

$$\text{Specificity} = \frac{\text{Number of cases correctly called negative}}{\text{Total number of negative cases}}$$

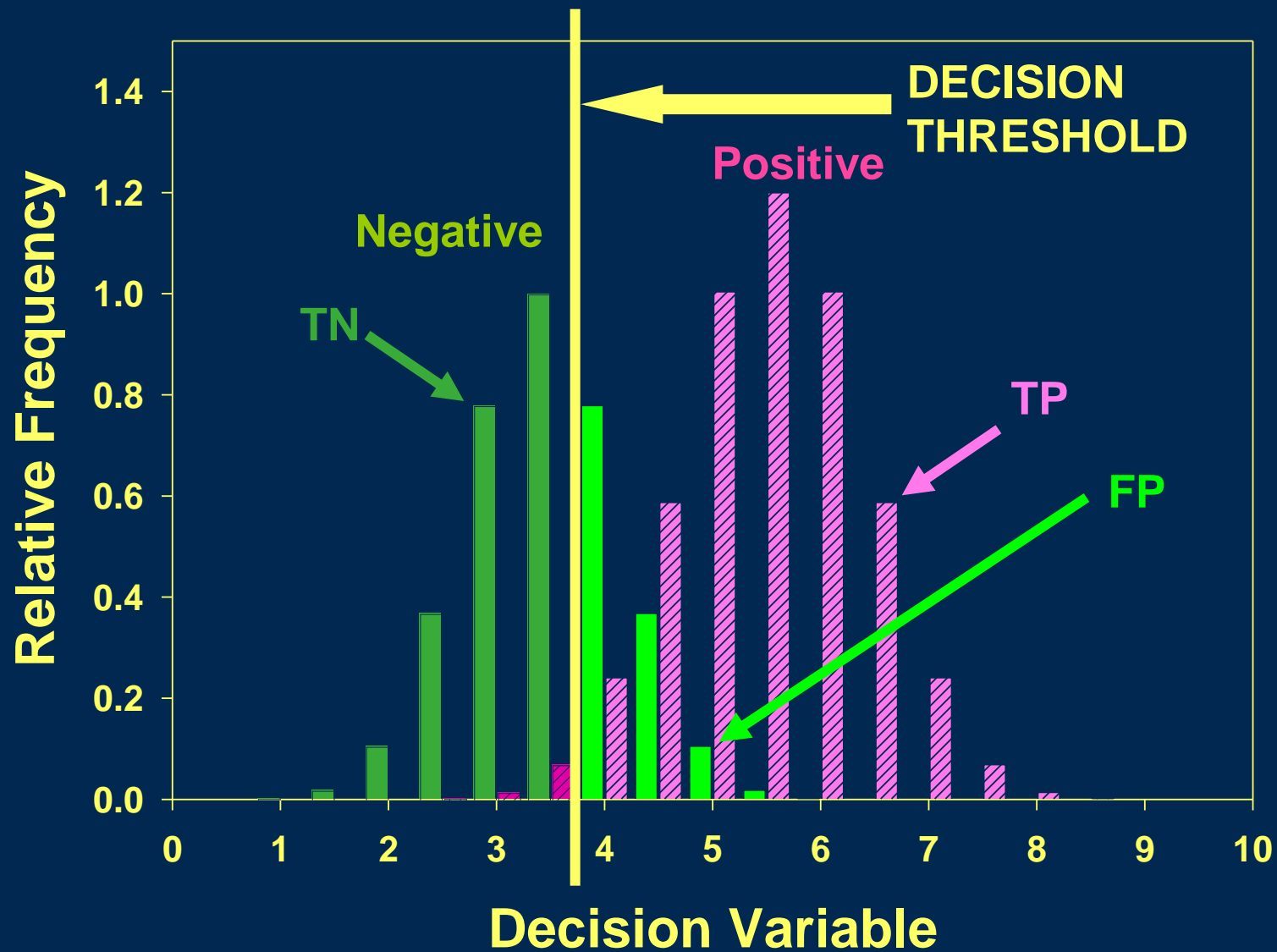
# WHO HAS THE BETTER CAD SYSTEM?

- **Two CAD systems A and B**
- **A: Sensitivity = 95%, Specificity = 60%**
- **B: Sensitivity = 75%, Specificity = 80%**

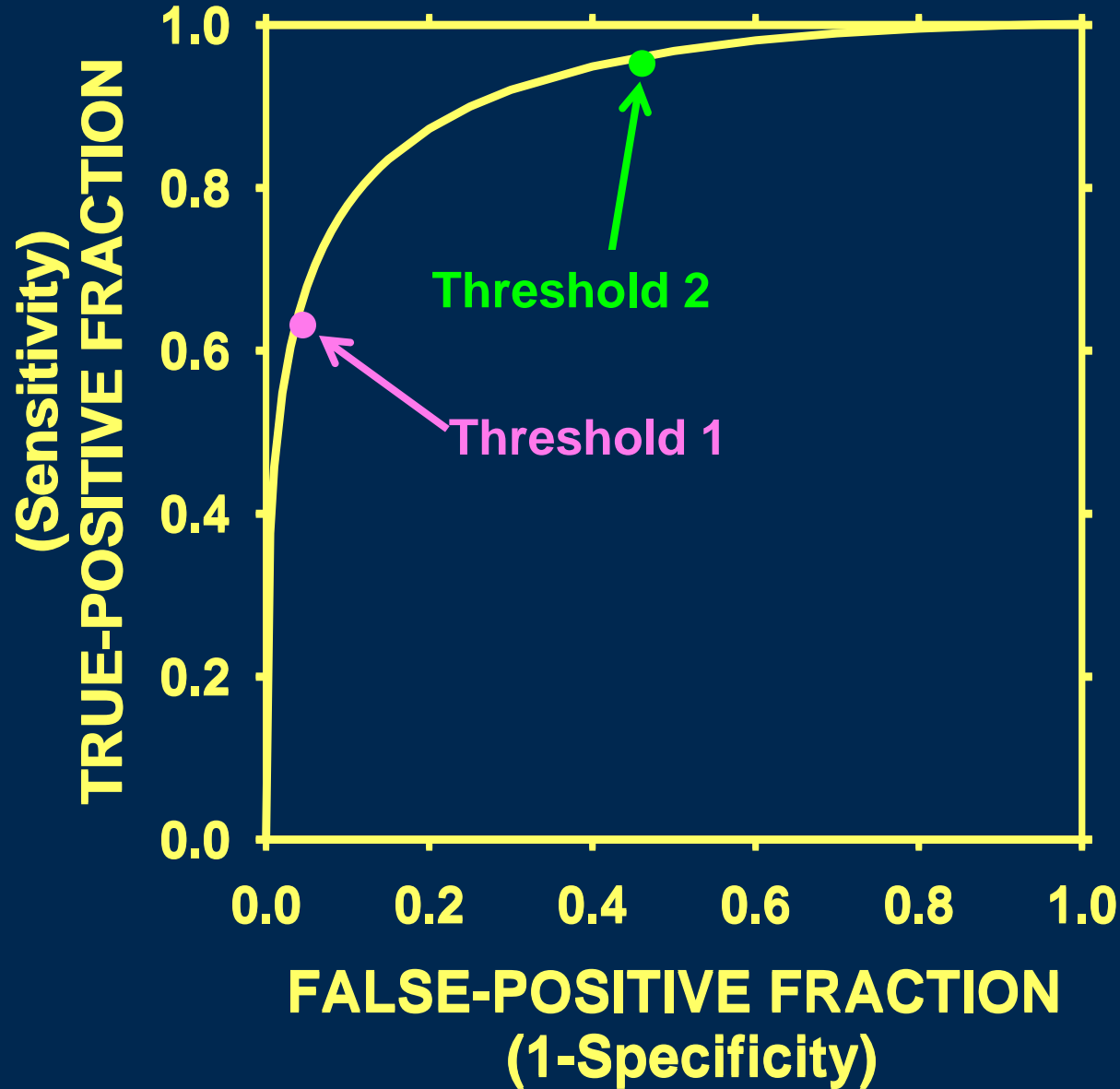
# THRESHOLDING



# THRESHOLDING



# RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



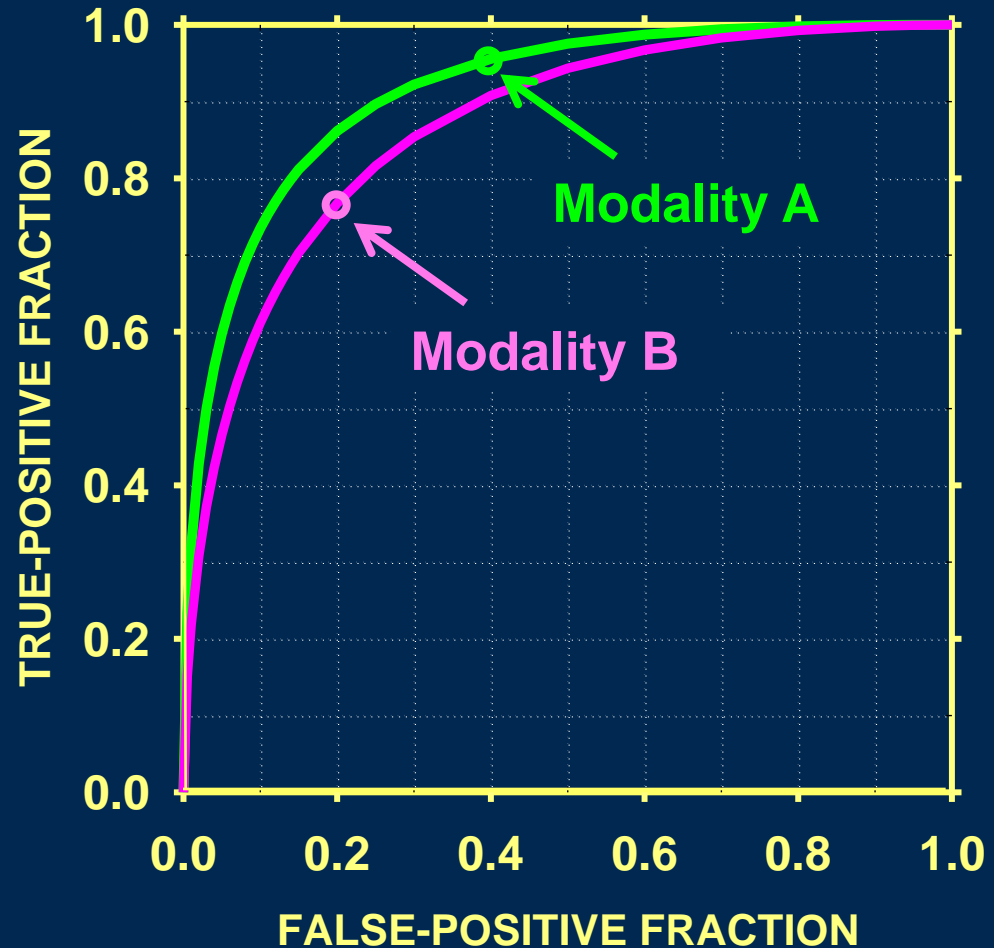
# ROC ANALYSIS

- **Evaluation of the performance of a diagnostic system**
  - Sensitivity as a function of (1-specificity) as the decision threshold varies
  - TPF vs. FPF

# WHO HAS THE BETTER CAD?

**A: Sensitivity = 95%**  
**Specificity = 60%**

**B: Sensitivity = 75%**  
**Specificity = 80%**



# PERFORMANCE MEASURE

- **Area under the curve (AUC)**
  - Average sensitivity over all specificities
  - A measure of the separation between positives and negatives
  - Entirely avoids the use of thresholds

# PERFORMANCE MEASURES: OBSERVER STUDY

- **Theoretical model for binary decision:**
  - Rank the case for likelihood of positive (diseased)
  - Decide if the rating is high enough to call the case positive
- **When this model applies**
  - All the more reason to avoid thresholds in observer performance studies

# THRESHOLDS VARY!

## SENSITIVITY-SPECIFICITY VARY!

- **Clinician to clinician (inter-reader)**
- **Same clinician interpreting the same case twice (intra-reader)**
- **In time:**
  - e.g., with availability of treatments

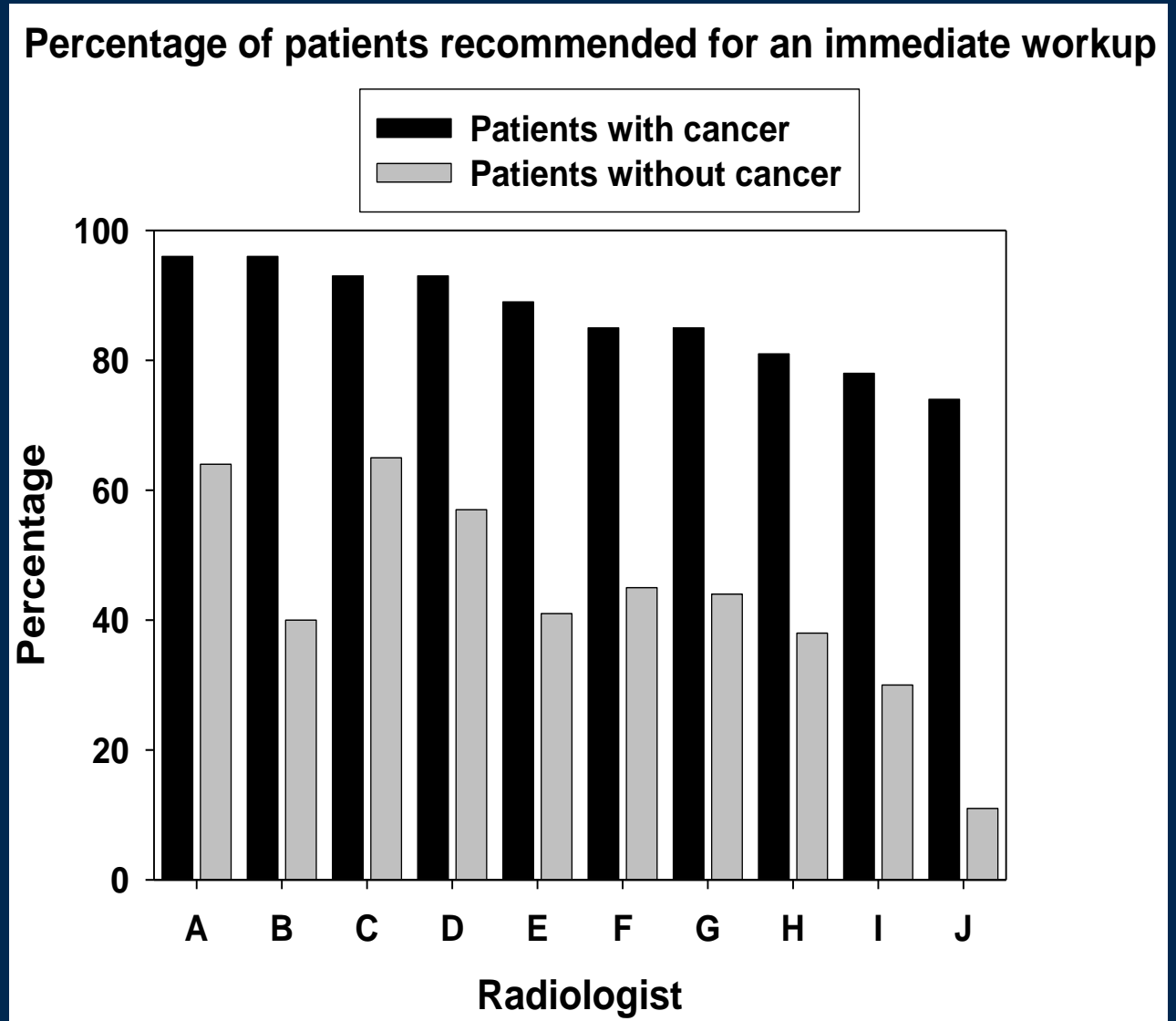
# INTER-CLINICIAN VARIABILITY

- **Elmore et al. investigated the inter-radiologist variability in mammogram interpretation**
  - 150 mammograms selected using stratified random sampling
  - 123 non-cancer
  - 27 cancer
  - 10 radiologists

\*Elmore *et al.*, "Variability in Radiologists' Interpretations of Mammograms," NEJM, 1994

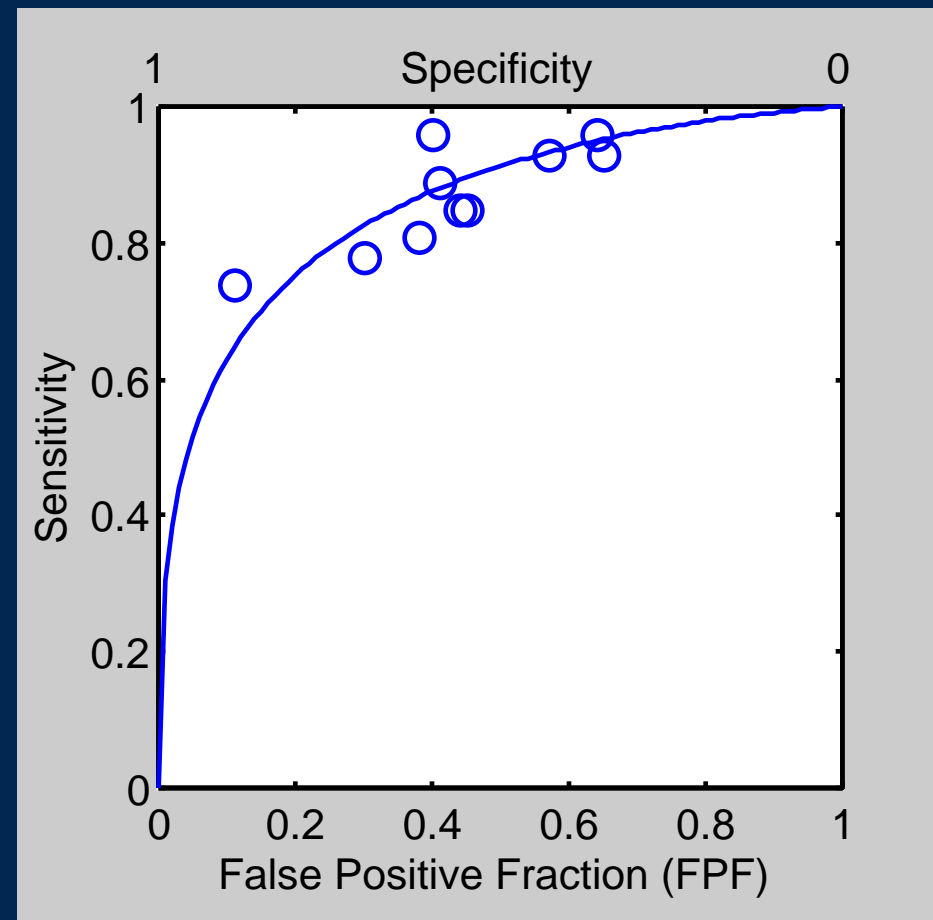
# INTER-CLINICIAN VARIABILITY

- Large variability among clinicians
- Each clinician has their own operating point



# INTER-CLINICIAN VARIABILITY

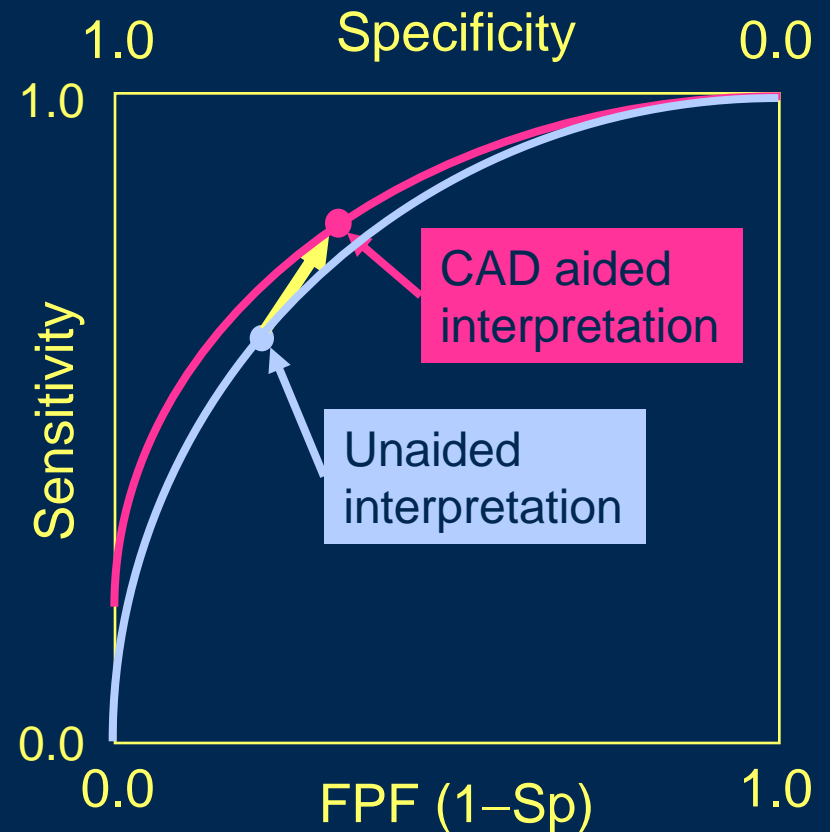
- **What happens if we plot in ROC space?**
  - Data fit well by an ROC curve
- **For different readers:**
  - Similar skill levels
  - Different threshold / operating points



\*Elmore *et al.*, "Variability in Radiologists' Interpretations of Mammograms," NEJM, 1994

# OBSERVER STUDY WITH CAD

- **Collect binary decisions from clinicians**
- **Collect ratings from clinicians and analyze with ROC**



# STATISTICAL INFERENCE

- **To conclude that interpretation with CAD is better than without CAD**
  - Performance measure difference
  - Standard deviation of performance measure difference
- **AUC reduces variability by**
  - Avoiding thresholds
  - Averaging
- **Use of AUC helps increase study power**

# I WISH I HAD TIME TO COVER

- **Location-specific ROC (FROC, LROC)**
- **Sources of variability**
- **MRMC studies**
- **Agreement measures**

# SUMMARY

- **Both standalone and observer studies are essential to assess novel CAD systems**
- **Study design and data analysis for CAD system assessment**
  - Important components:
  - Training and test databases, reference standard, scoring, observer study
- **ROC is the preferred assessment methodology to collect and analyze data**