Session 3: Identifying Radiologists Who Might Benefit from Intervention

Patricia A. Carney, PhD
Oregon Health & Science University
Knight Cancer Institute
Portland, OR

Background

- Significant variability noted in interpretive acumen of practicing radiologists:
  - 75% to 95% for ***sensitivity***
  - 83% to 98.5% for ***specificity***
- Only factor consistently associated with improved acumen is ***fellowship training*** (Miglioretti, et al, Elmore, et al)
- Desirable indices have been published as ***target goals***:
  - E.g., 1994 Agency for Health Care Policy and Research "desirable goal" of 85% for Sensitivity
- We need ***cut points for low performers*** to identify and encourage them to attain additional training.

Background

- Angoff Method - Process Approach for Setting Cut-Point Criteria for Low Performers
- Developed in 1970s - Applied in International and National Board Certification & Licensing Exams in Medicine (Both knowledge and skill based exams!! (e.g., USMLE-CX)
- Purpose is to Increase “Accountability” for Meeting a Proficiency Standard ***derived by those in the field***
- Most Commonly Used Method to Set Educational Performance Standards Today
- Our goal was to come to consensus on cut-points for interpretive performance for both screening and diagnostic mammography

Held Two Angoff Meetings

Meeting 1 – Seattle, WA to address Screening mammography (January 2009)

Included: 10 experts with Eligibility Criteria:
1) Devoted ≥ 75% time to breast imaging,
2) Been interpreting mammograms for at least 10 years, and
3) Completed fellowship training in breast imaging (such training programs began around 1985) or had more than 15 years of experience in interpreting mammograms.

Meeting 2 – Seattle, WA to address Diagnostic Mammography (**September, 2011**)
Included: 11 Experts (Same criteria)

Modified Angoff Methods

Phase I: Consider a hypothetical group of 100 radiologists who are "minimally" capable performers (those who you think might benefit from additional training):

- ***Working independently***, what performance cut-point would you set for ***sensitivity***, where falling below the cut-point would hypothetically result in recommending additional training

- Screening Mammography
  - Sensitivity
  - Specificity
  - Recall
  - PPV1
  - CA Detection Rate

Diagnostic Mammography

- Sensitivity
- Specificity
- Abnormal Interpretation
- PPV2 and PPV 3
- CA Diagnosis Rate

Screening Mammography – Definitions

- **Screening Mammogram** – Bilateral mammogram done for asymptomatic women
- **Sensitivity** - Ability to find a cancer when it is present [TP/(TP+FN)]
- **Specificity** - Ability of the test to determine that a disease is absent when a patient is disease-free [TN/(TN+FP)]
- **Recall Rate** - Proportion of all women undergoing screening mammography who are given a positive interpretation (Category 0, 4, 5)
- **PPV1** - Proportion of women with positive screening examinations (Category 0, 4, 5) who are diagnosed with breast cancer [TP/(TP +FP1)].
- **PPV2** - Proportion of all women with positive screening examinations and a recommendation for biopsy at the end of imaging work-up (BI-RADS category 4 or 5) who are diagnosed with breast cancer [TP 2 /(TP 2 + FP 2)
- **CA Detection** - Number of women found with breast cancer per 1,000 women screened.

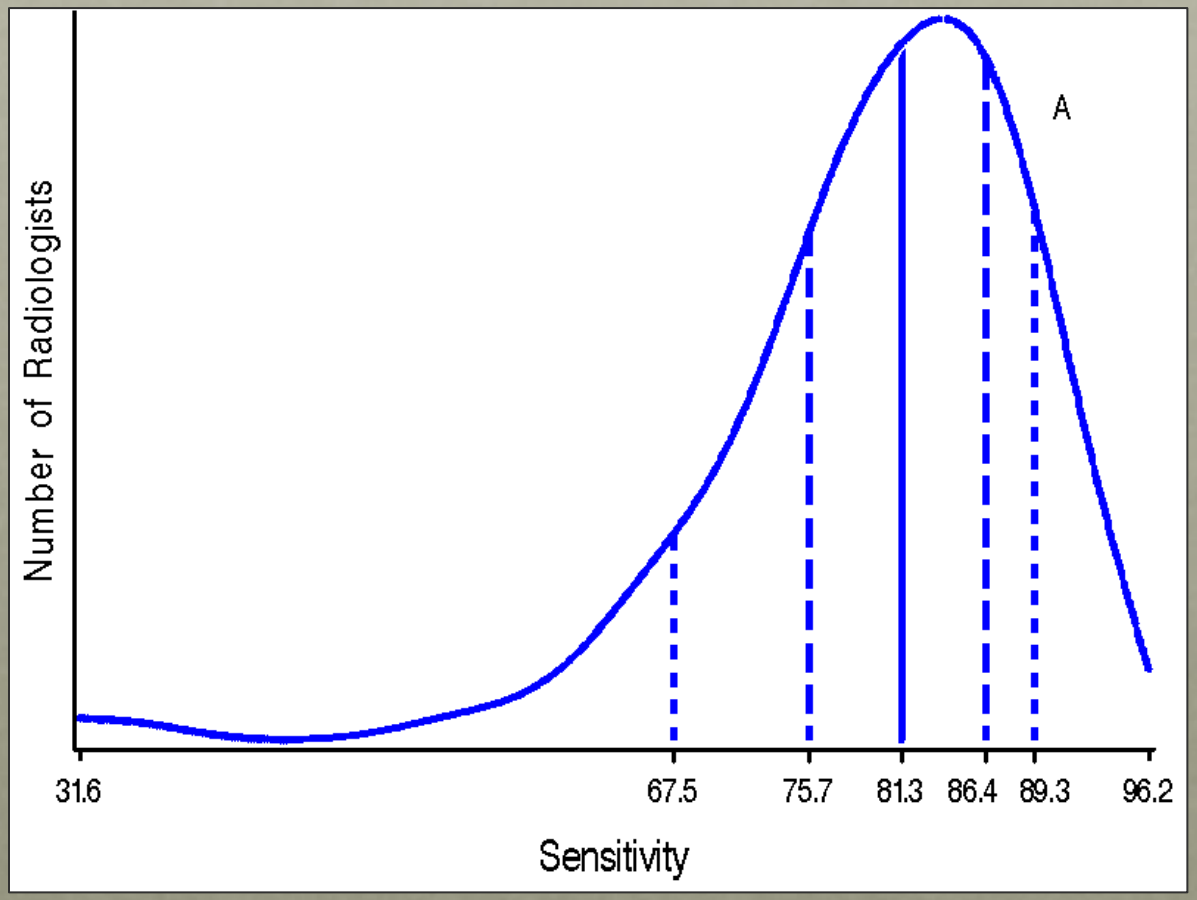
Diagnostic Mammography – Definitions

- **Diagnostic Mammogram (1)** – For work-up of prior abnormal screening mammograms
- **Diagnostic Mammogram (2)** – For work-up breast lump
- **Sensitivity** - Ability to find a cancer when it is present [TP/(TP+FN)]
- **Specificity** - Ability of the test to determine that a disease is absent when a patient is disease-free [TN/(TN+FP)]
- **Abnormal Interpretation Rate** - Proportion of all women undergoing diagnostic mammography who are given a positive final assessment (Category 4, 5)
- **PPV2** - Proportion of all women *recommended for biopsy* after diagnostic mammography (Category 4, 5) who are diagnosed with breast cancer [TP/(TP +FP2)].
- **PPV3 -** Proportion of all women *who received a biopsy* after diagnostic mammography (Category 4, 5) who are diagnosed with breast cancer [TP/(TP +FP2)].
- **CA Diagnosis** - umber of women found with breast cancer per 1,000 women receiving diagnostic mammography.

Screening Mammography

- **Phase II: Normative Data for Sensitivity**
- Open Discussion of Working Cut-Points

Smoothed Plots of Sensitivity for 16,324 Cancers on Screening Mammography (Among Radiologists interpreting 30 or More Cancer cases), 1996 - 2005

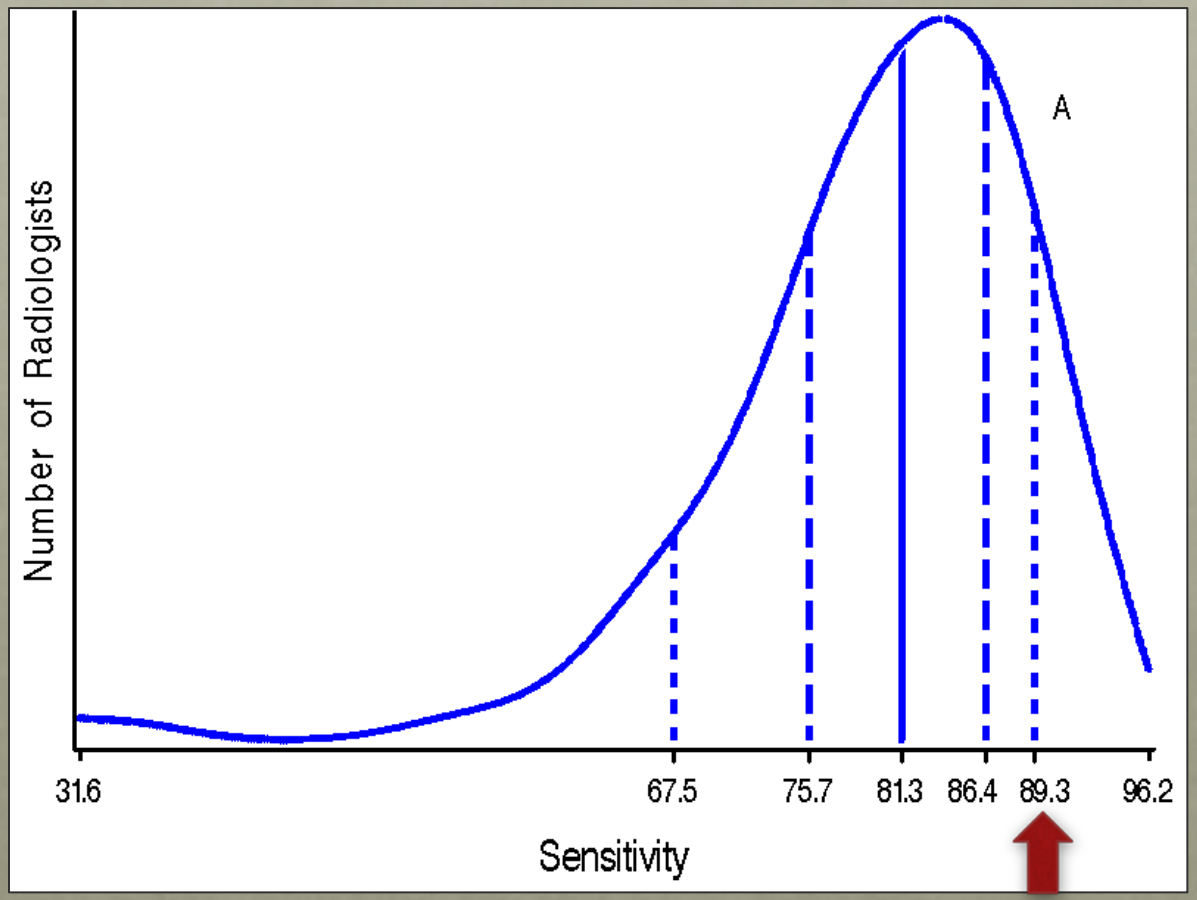


An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Screening Mammography

- **Phase II: Normative Data for Sensitivity**
- Open Discussion of Working Cut-Points

Smoothed Plots of Sensitivity for 16,324 Cancers on Screening Mammography (Among Radiologists interpreting 30 or More Cancer cases), 1996 - 2005

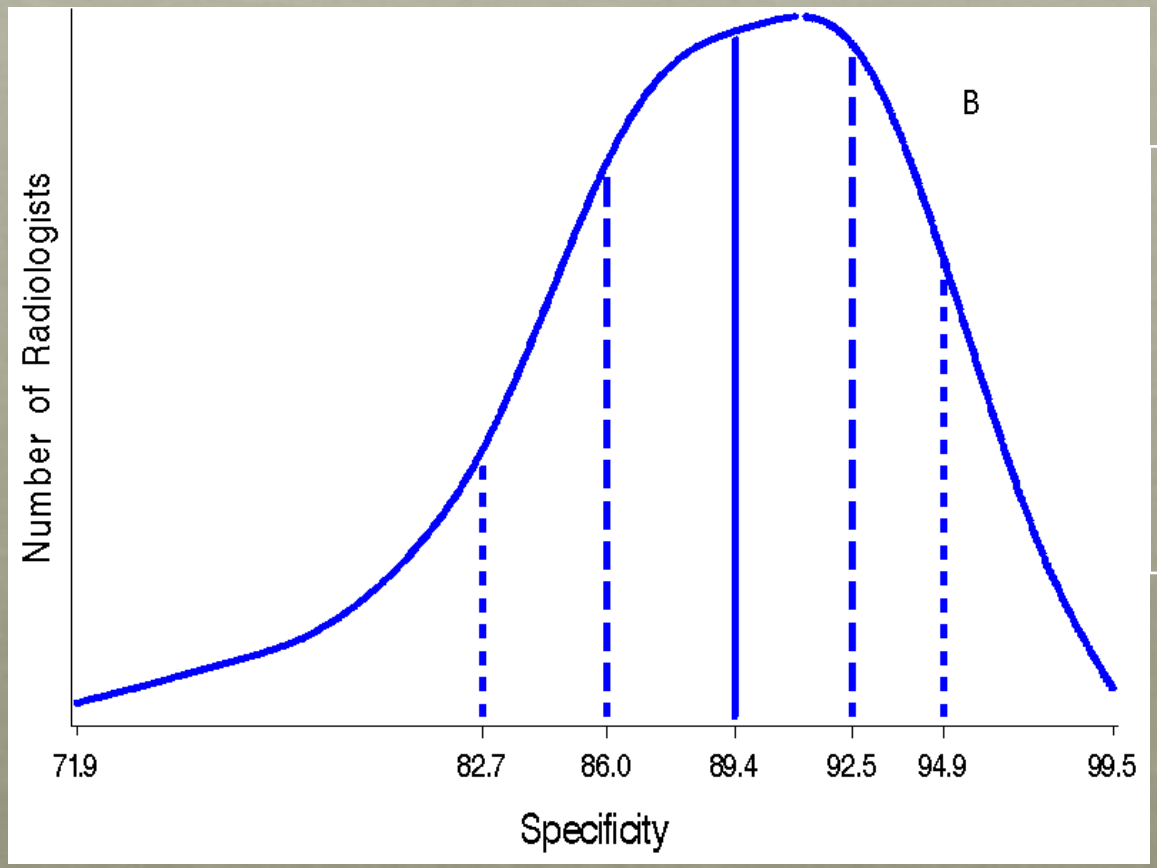


An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Screening Mammography

- **Phase II: Normative Data for *Specificity***
- Open Discussion of Working Cut-Points

Smoothed Plots of Specificity for 3,275,015 Non-cancers (Among Radiologists interpreting 1000 or More Non-Cancers), 1996 - 2005

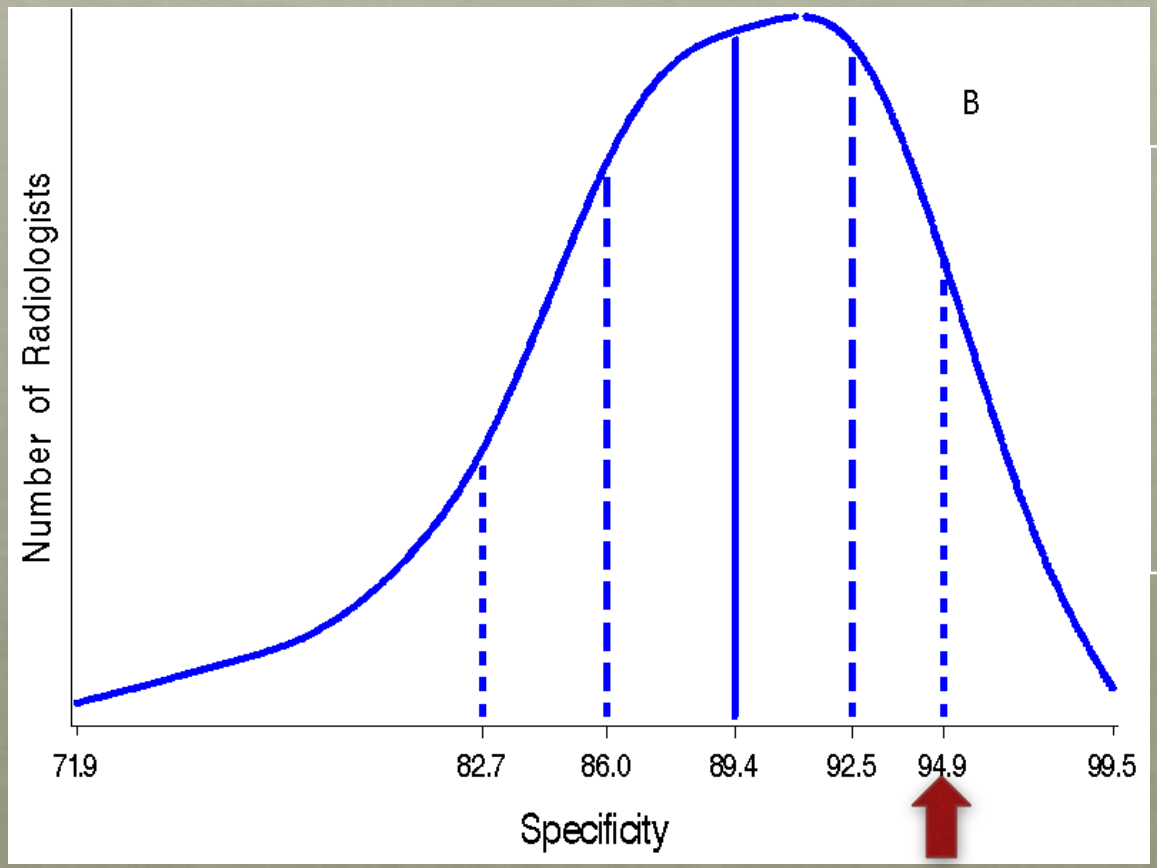


An overlaid solid line indicates 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Screening Mammography

- **Phase II: Normative Data for *Specificity***
- Open Discussion of Working Cut-Points

Smoothed Plots of Specificity for 3,275,015 Non-cancers (Among Radiologists interpreting 1000 or More Non-Cancers), 1996 - 2005

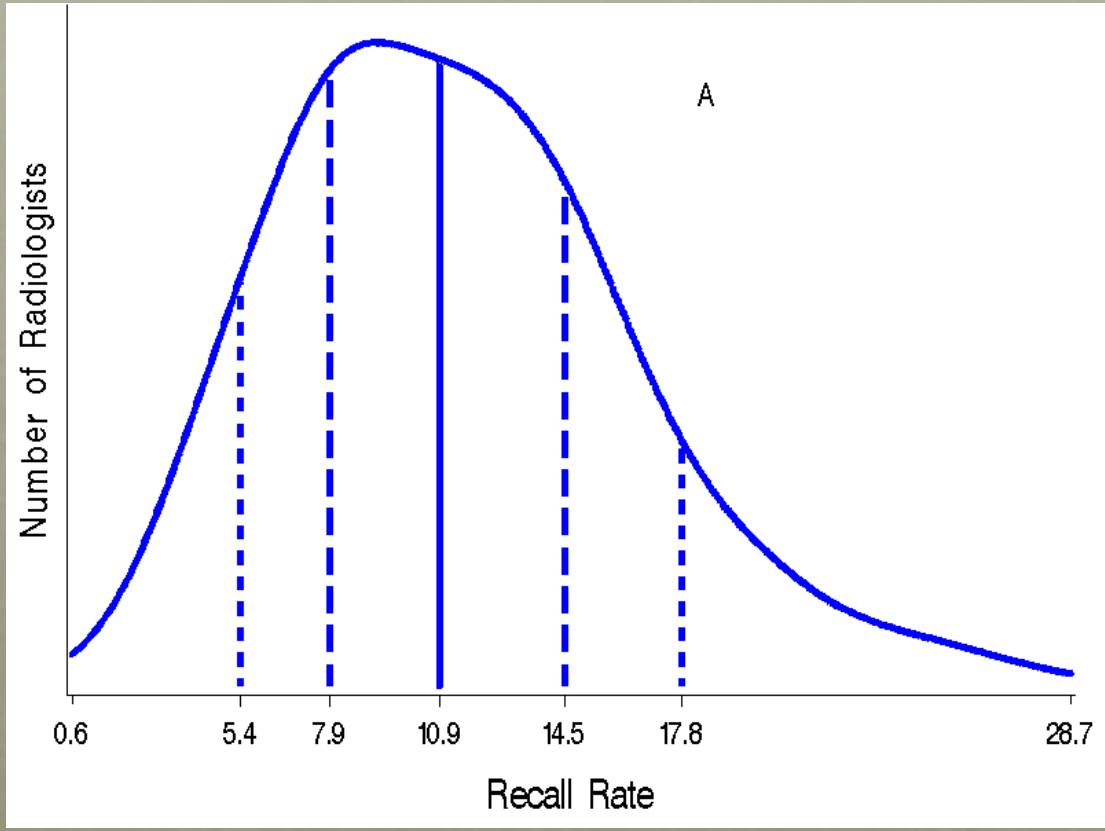


An overlaid solid line indicates 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Screening Mammography

- **Phase II: Normative Data for *Recall***
- Open Discussion of Working Cut-Points

Smoothed Plots of Frequency Distributions of Recall Rates for 3,294,680 Screening Mammography Examinations (Among Radiologists with 1000 or More Examinations), 1996 - 2005

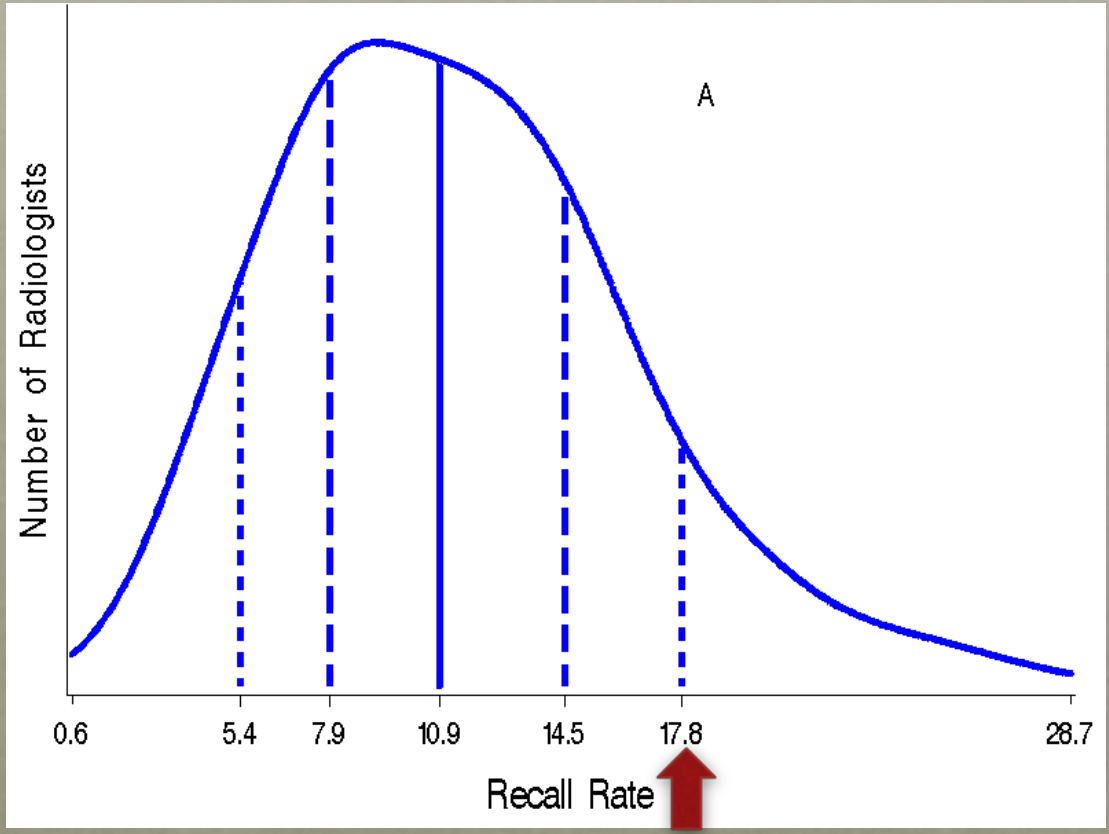


An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Screening Mammography

- **Phase II: Normative Data for *Recall***
- Open Discussion of Working Cut-Points

Smoothed Plots of Frequency Distributions of Recall Rates for 3,294,680 Screening Mammography Examinations (Among Radiologists with 1000 or More Examinations), 1996 - 2005



An overlaid solid line indicates the 50th percentile (median), paired dashed lines indicate the 25th and 75th percentiles, and paired dotted lines indicate the 10th and 90th percentiles.

Final Cut Points for Screening Mammography

Measure	Low Performance Range	Percent of BCSC Radiologists in Low Performance Range
Sensitivity	<75	18.0 %
Specificity	<88 or >95	47.7 %
Recall Rate	<5 or >12	49.1 %
PPV1	<3 or >8	38.4 %
PPV2	<20 or >40	34.0 %
CDR*	<2.5/1,000	28.4 %

* CDR = Cancer Detection Rate

Final Cut Points for Diagnostic Mammography (1)

Diagnostic Mammograms to Work-up Prior Abnormal Screening Exams		
Measure	**Low Performance Range**	**Percent of BCSC Radiologists in Low Performance Range**
Sensitivity	<80	21.5 %
Specificity	<80 or >95	25.1 %
Abnormal Interpretation Rate	<8 or >25	25.7 %
PPV_2	<15 or >40	21.8 %
PPV_3	<20 or >45	27.6 %
CDR*	<20/1,000	23.2 %

* CDR = Cancer Diagnosis Rate

Final Cut Points for Diagnostic Mammography (2)

Diagnostic Mammograms to Work-up a Breast Lump		
Measure	**Low Performance Range**	**Percent of BCSC Radiologists in Low Performance Range**
Sensitivity	<85	31.6 %
Specificity	<83 or >95	24.0 %
Recall Rate	<10 or >25	20.5 %
PPV_2	<25 or >50	32.3 %
PPV_3	<30 or >55	46.3 %
CDR*	<40/1,000	19.7 %

* CDR = Cancer Diagnosis Rate

Screening Simulations

Simulated a cohort of 1 million women and a cancer status for each woman based on a prevalence of ~ five cases per 1000 women in the BCSC to investigate the potential impact of moving the lower-performing physicians' performance measures into the acceptable range on the basis of the BCSC normative data.

- 1,000,000 women with 4,834 having breast cancer
- # of cancers correctly recalled if performance improved increased from 4,078 to 4,216
- # of false positives would decrease from 91,454 to 82,621

Screening Simulations

If underperforming physicians moved into the acceptable range, we would expect:

- Detection of an additional 14 cancers per 100,000 women screened
- Reduction in the number of false-positive examinations by 880/100,000 women screened

Diagnostic Simulations

Diagnostic Mammography: If underperforming physicians moved into the acceptable range after remedial training, the expected result would be:

Work-up after abnormal screening:

- Diagnosis of an additional 86 cancers per 100,000 women
- Reduction in the number of false-positive examinations by 1,067 per 100,000 women undergoing this workup

Work-up of a breast lump:

- Diagnosis of an additional 335 cancers per 100,000 women
- Reduction in the number of false-positive examinations by 634 per 100,000 women

Limitations

- We examined performances measures independently of each other, but they are very inter-related…
- For the normative data, we required at least 30 cancer interpretations for sensitivity and 1000 interpretations for the other performance measures – However these numbers may be too small to provide stable estimates…

Limitations

- Single measure of sensitivity does not discriminate between interpreting physicians given that tumor size varies.
- Typically not possible to accurately calculate some of these key measures (e.g., sensitivity and specificity) in actual clinical practice.
- Experts taking part in Angoff process may not be representative of all expert mammographers in U.S.
- Educating those who fall below cut points identified may not improve their performance – this requires further study.

Areas for Future Research

- Can Mini-fellowships, Areas of Concentration or 'Selectives' done during Residency Improve interpretative performance or is a full breast imaging fellowship needed?
- Stakes for Continuing Professional Development Programs Should be Higher - but ***not*** until evidence of their effectiveness can be determined.
- Rather than creating physician life-long learners, we need *master adaptive learners* who can adjust rapidly to new technologies, new health systems changes and emerging information on patient risk.

Thank You!