Session 2: US and International Variation in Volume and Performance Measures

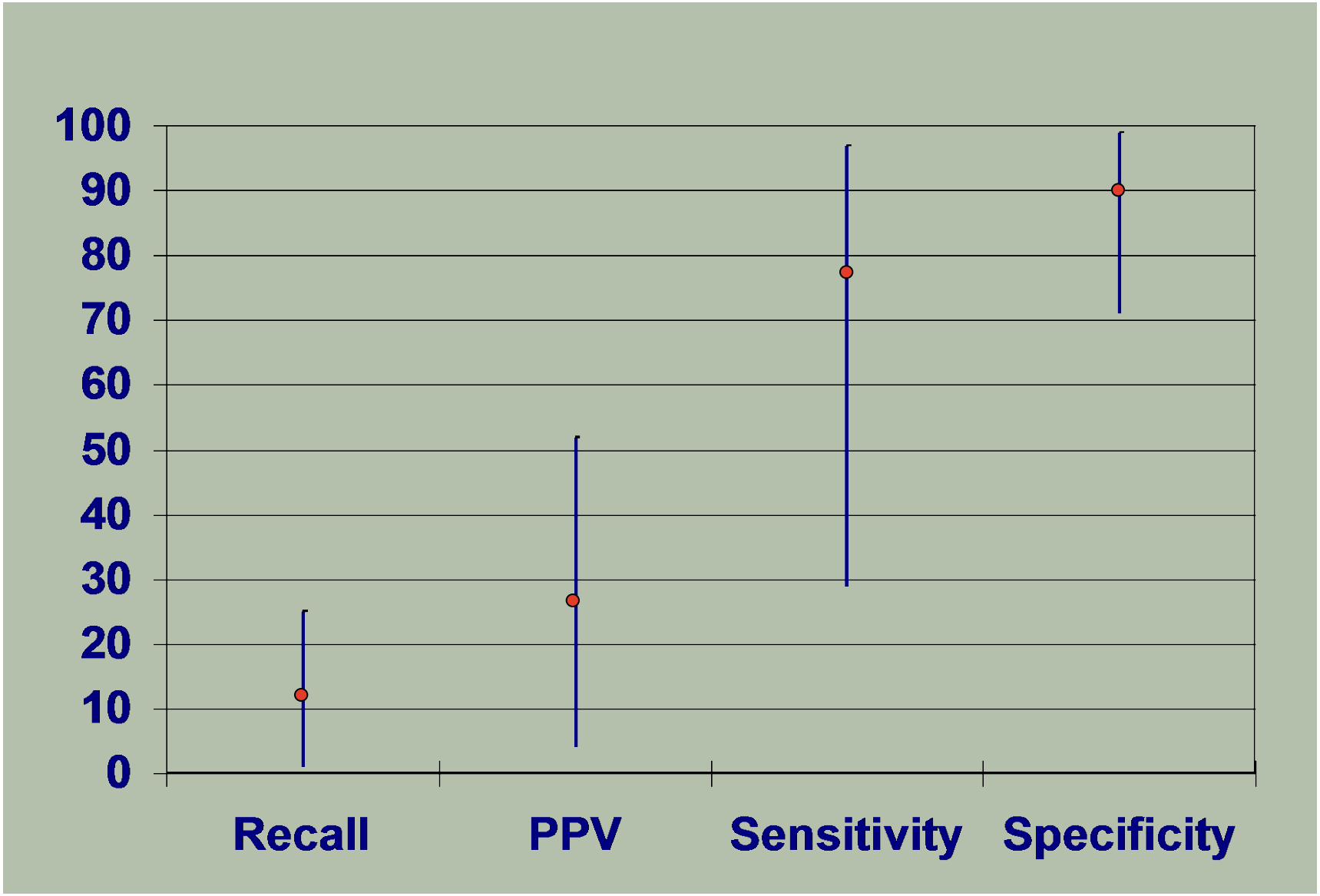
Diana SM Buist, PhD, MPH

Senior Investigator, Group Health Research Institute

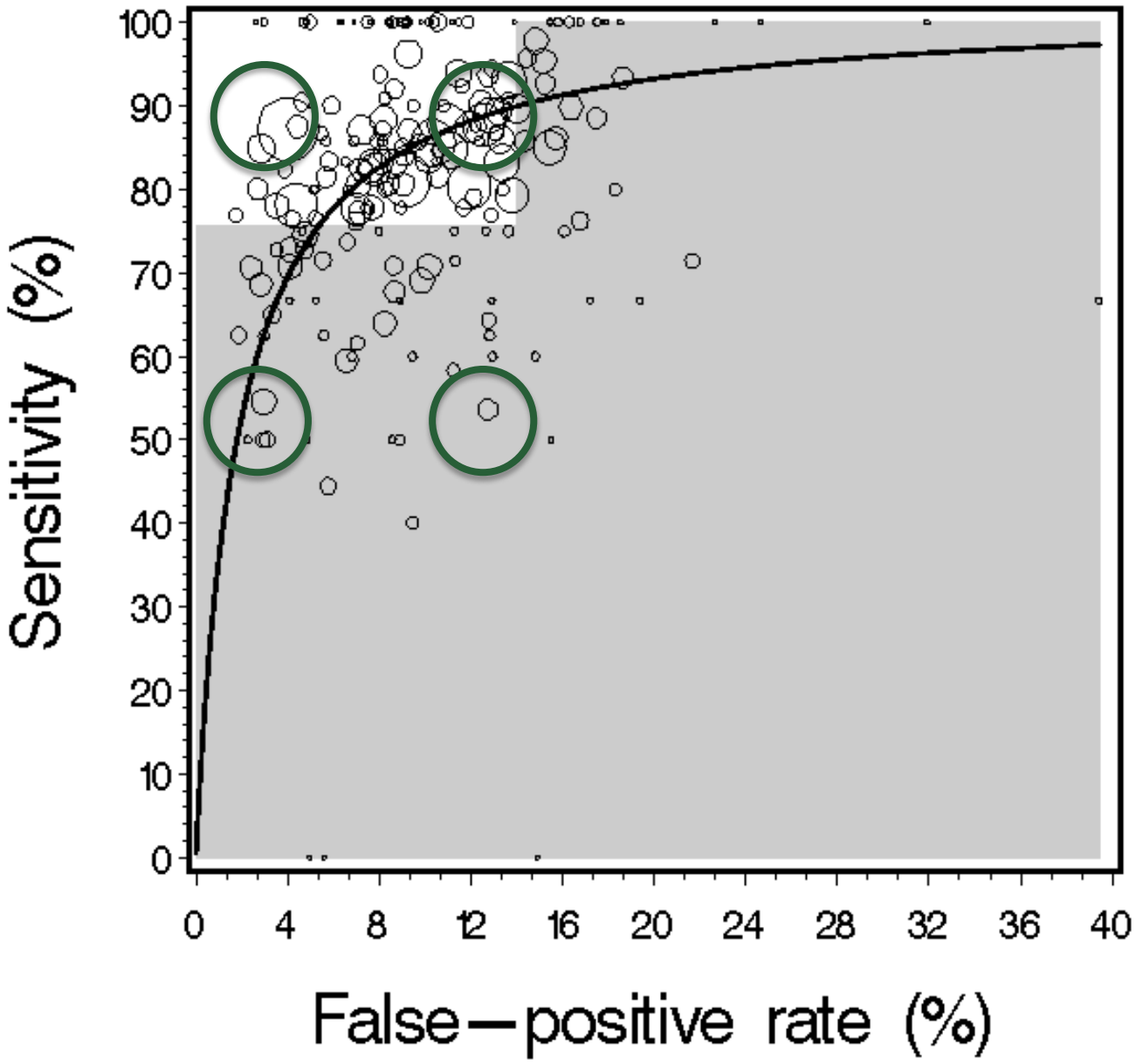
Affiliate Member, Fred Hutchinson Cancer Research Center

Affiliate Professor, University of Washington School of Public Health, Department of Health Services, Department of Epidemiology

Known: wide interpretive variability in US



Known: high performers on one measure not necessarily high on others



Breast Cancer Surveillance Consortium
BCSC
Working together to advance breast cancer research

2005 IOM Report Conclusions:

Extensive variability in mammography interpretation exists among radiologists in the United States.

Interest in understanding reasons for this variability

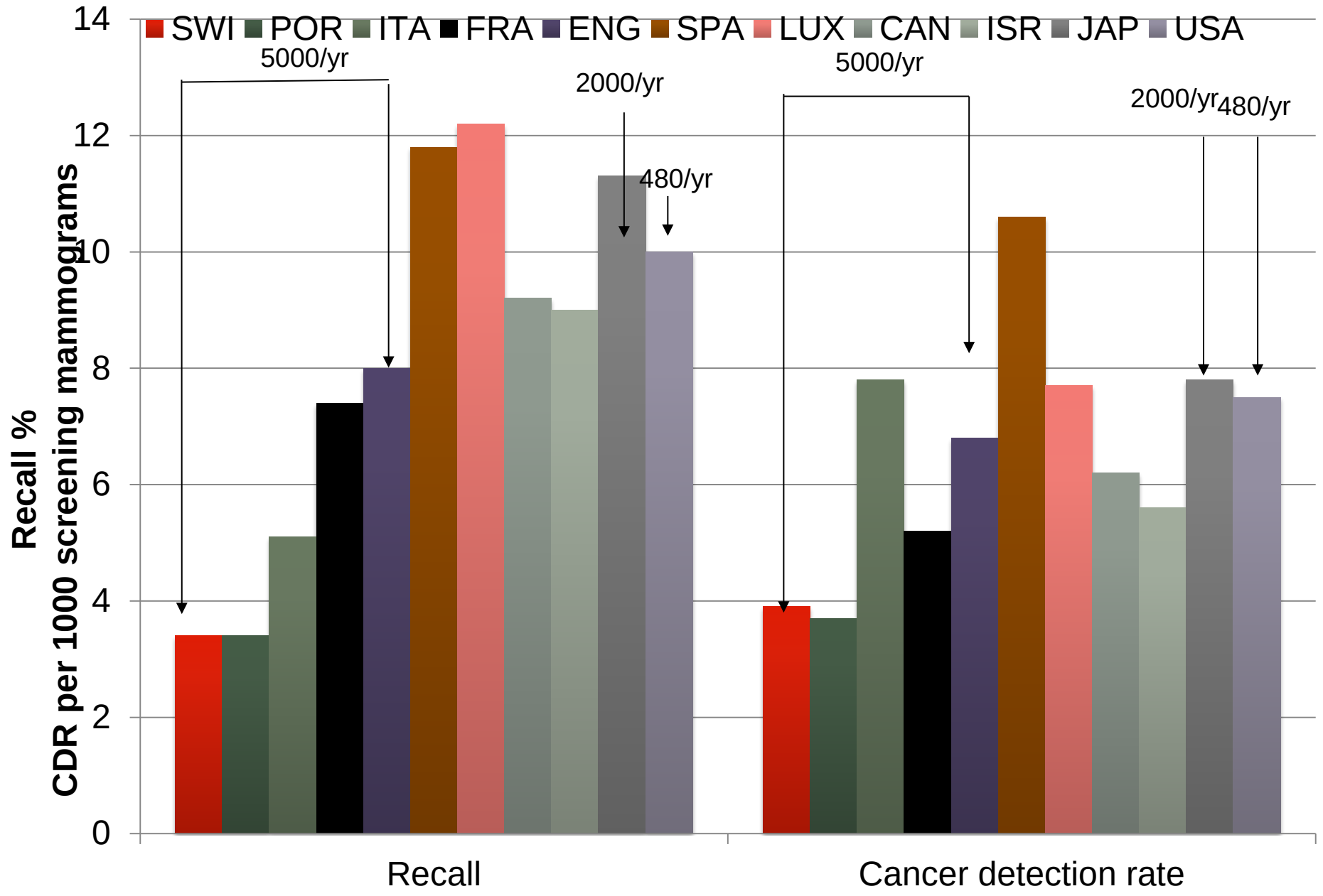
- Patient factors
- Practice and facility characteristics
- Radiologist characteristics
  - Years of experience
  - Training
  - Specialty
  - **Annual interpretive volume**

MQSA interpretation requirements

960 mammograms in last 24 months

What can we learn from other countries?

International mammography performance variability by minimum annual volume



Possible reasons for international variation

- US:
  - Generalized not organized screening programs
  - Shorter screening interval
  - Screening starts younger & goes to older ages
  - Legal differences
  - Radiologist's incentives
    - >66% of radiologists do some breast imaging
    - Not most lucrative
- Different reading/interpretation practices
  - Double reading
  - Centralized reading
  - Higher volume – more specialized
- Different volume requirements and quality standards
  - Under-performing programs and MDs reviewed
  - Physicians take biennial exams – voluntary

Interpretive volume and sensitivity

Study	Country	Years	Volume categories (min, max)	
			Total Volume	Screening Volume
Théberge 2014	Canada	2000-2006	< 250, ≥ 2 500	500-999, ≥ 4 000
Buist 2011	USA	2002-2006	< 480, ≥ 5 000	< 480 , ≥ 3 000
Duncan 2011	Scotland	2006-2009		< 5 000, > 8 333
Elmore 2009	USA	1998-2005	≤ 1 000, > 2 000	
Woodard 2007	USA	1996-2001	≤ 1 000, > 2 000	
Smith-Bindman 2005	USA	1995-2000	480-750, > 4 000	
Barlow 2004	USA	1996-2001		≤ 1 000, > 2 000

Statistical significant association
Not statistical significant association

Interpretive volume and false-positive rates

Study	Country	Years	Volume Categories (min, max)	
			Total Volume	Screening Volume
Théberge 2014	Canada	2000-2006	< 250, ≥ 2 500	500-999, ≥ 4 000
Buist 2011	USA	2002-2006	< 480, ≥ 5 000	< 480 , ≥ 3 000
Cornford 2011	UK	2005-2008		< 5 000, ≥ 8 333
Duncan 2011	Scotland	2006-2009		< 5 000, > 8 333
Elmore 2009	USA	1998-2005	≤ 1 000, > 2 000	
Woodard 2007	USA	1996-2001	≤ 1 000, > 2 000	
Coldman 2006	Canada	1998-2000		480-699, ≥ 5 000
Tan 2006	USA	1998-1999	≤ 200, ≥ 560	
Smith-Bindman 2005	USA	1995-2000	480-750, > 4 000	
Théberge 2005	Canada	1998-2000		< 250, ≥ 1 500
Barlow 2004	USA	1996-2001		≤ 1 000, > 2 000
Kan 2000	Canada	1994-1997		< 2 000, 4 000-5 199

Statistical significant association: lower volume = higher FPR
Other statistical significant association
Not statistical significant association

5 / 12 | 2 / 12 | 5 / 12

Why conflicting study findings across countries & within?

- Volume measurements – self-report vs. comprehensive volume pooled across facilities
- Outcomes – what is optimal accuracy?
  - Carney *Radiology* 2013, Miglioretti *AJR* 2015
- Statistical modeling Miglioretti *Academic Radiology* 2009
  - Conditional/cluster-specific
  - Marginal/population-averaged

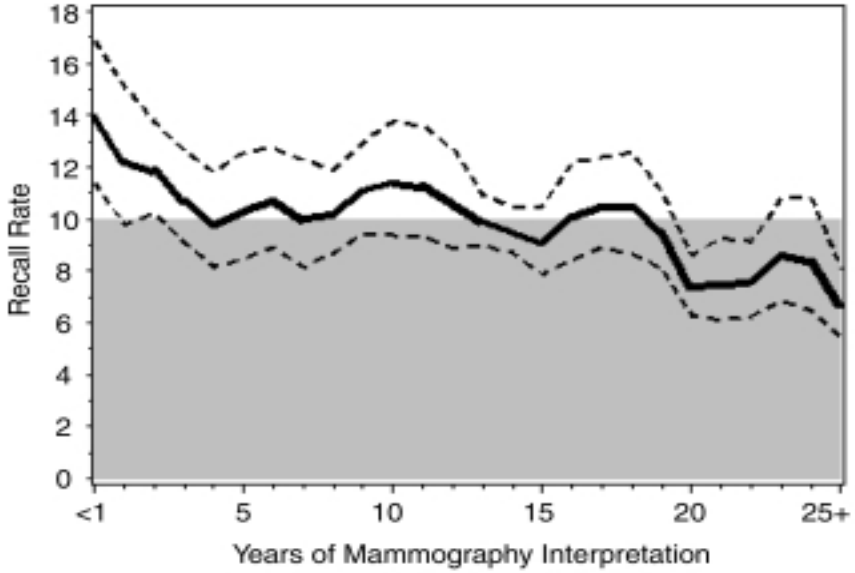
2005 IOM Report Conclusions:

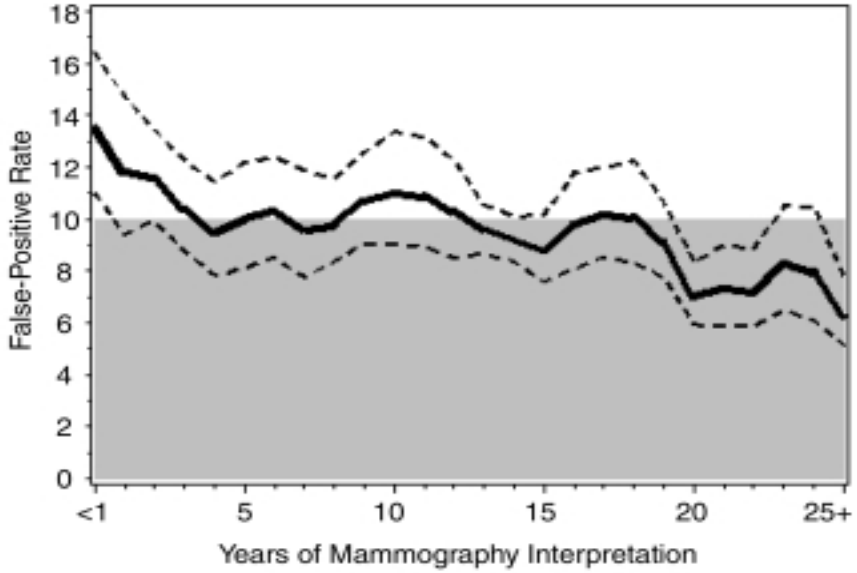
Extensive variability in mammography interpretation exists among radiologists in the United States.

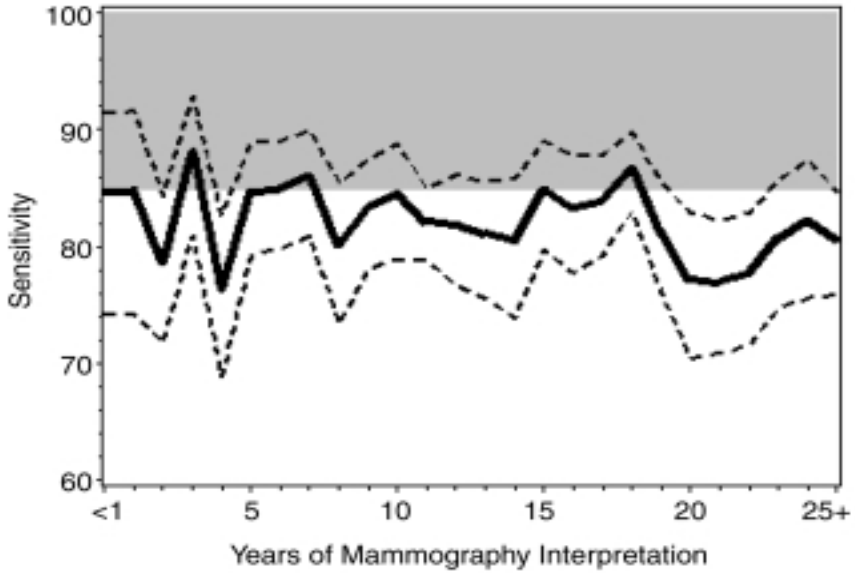
Interest in understanding reasons for this variability

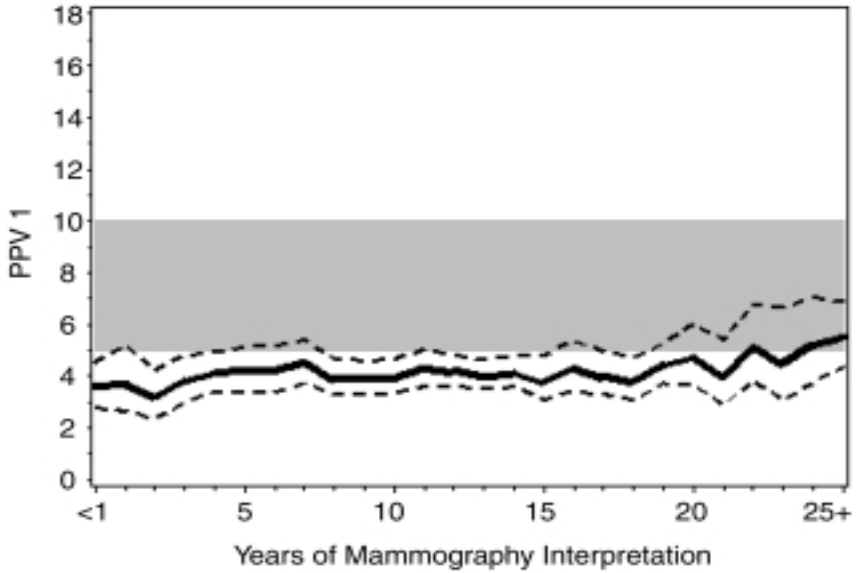
- Patient factors
- Practice and facility characteristics
- Radiologist characteristics
  - Years of experience
  - Training
  - Specialty
  - Annual interpretive volume

Experience matters: average performance by years of experience
AHRQ desirable goals for screening mammography (shaded)









Miglioretti *Radiology* 2009

INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES
Advising the nation • Improving health

Volume-Performance relationships within the Breast Cancer Surveillance Consortium

Volume measures:

- screening
- diagnostic
- total
- % screening/total

BCSC Radiologists

- 6 BCSC registries (San Francisco, North Carolina, New Hampshire, Vermont, Washington State, New Mexico)
- Radiologists who interpreted screening mammograms between 2005-2006 (N=214)
- Collected comprehensive volume measures from all facilities interpreted in between 2001-2005
- Final sample: 120 Radiologists (average 4 years of volume measures) for a total of 481 reader-years
  - 91% radiologists only interpreted in BCSC facilities

Volume definitions

Annual volume for each volume measure

Primary reader only

Screening volume

- Routine views only
- Screening + diagnostic views on same day by same reader

Diagnostic volume (SIFU, additional evaluation & symptomatic)

- Diagnostic views only
- Screening + diagnostic views on same day by same reader

Total

- Routine views
- Diagnostic view
- Screening + diagnostic views on same day by same reader

% screening/total

Outcomes measures

Linked annual volume measures to performance in the following year

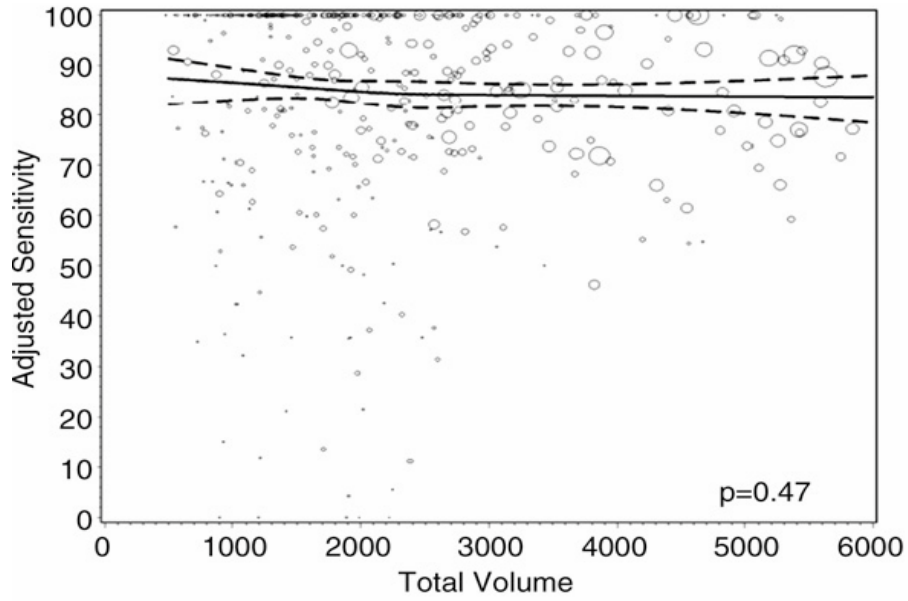
Performance measures:

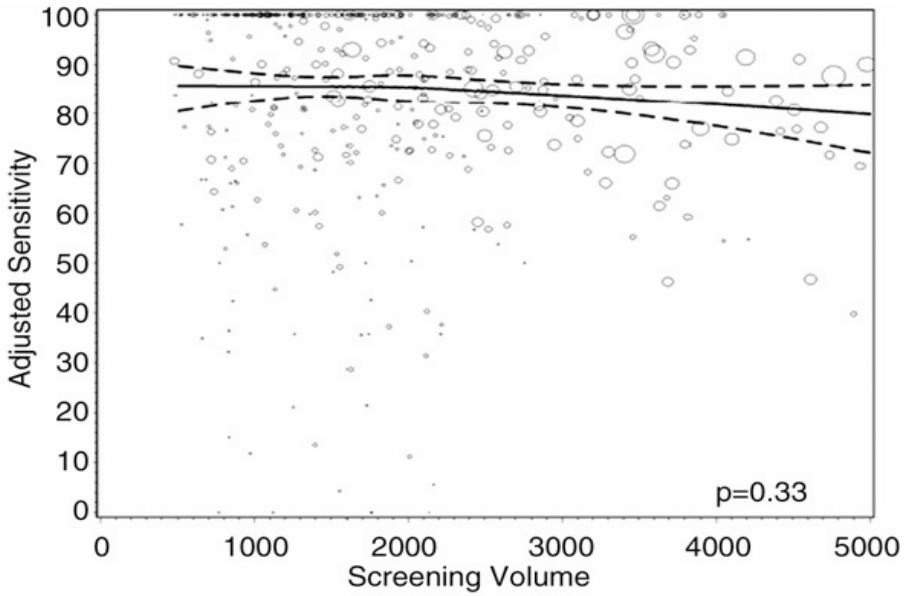
- Sensitivity
- False positive rate
- Cancer detection rate per 1000 screening mammograms
- Number of women recalled per cancer detected

All data shown adjusted for: age, time since last mammogram,

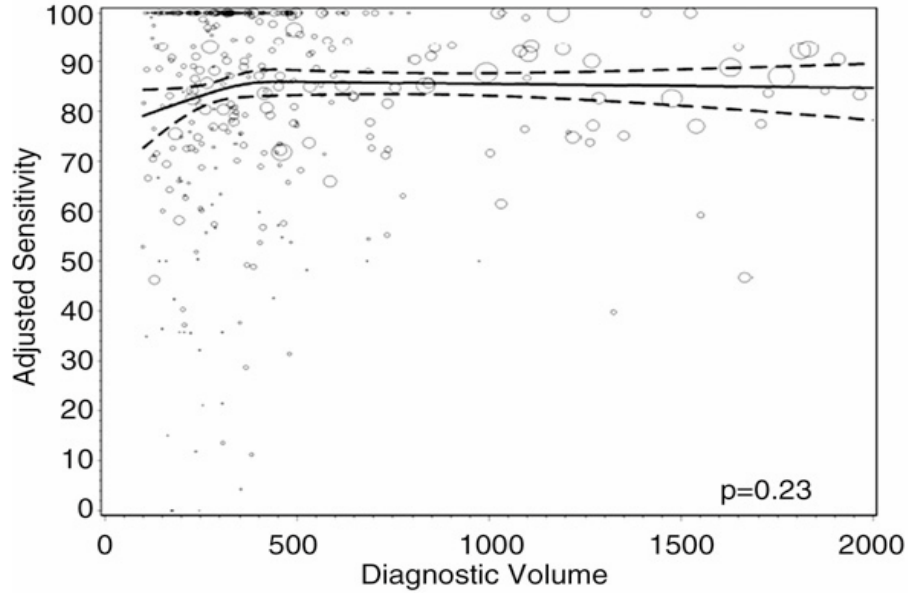
Buist *Radiology* 2011

Sensitivity

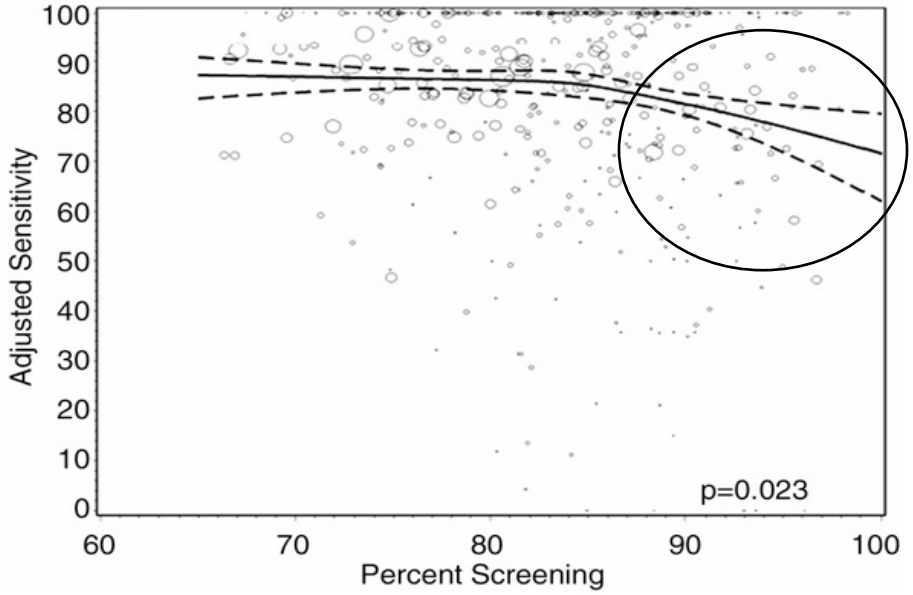




a.

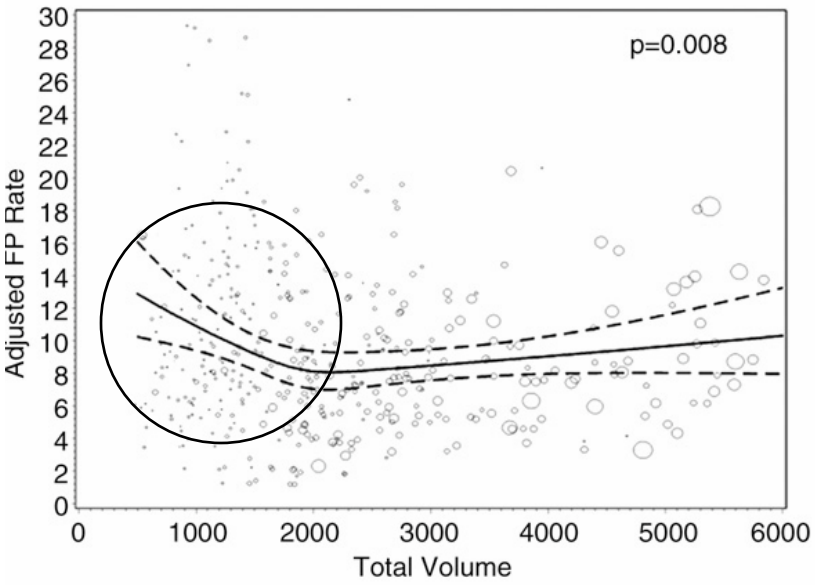


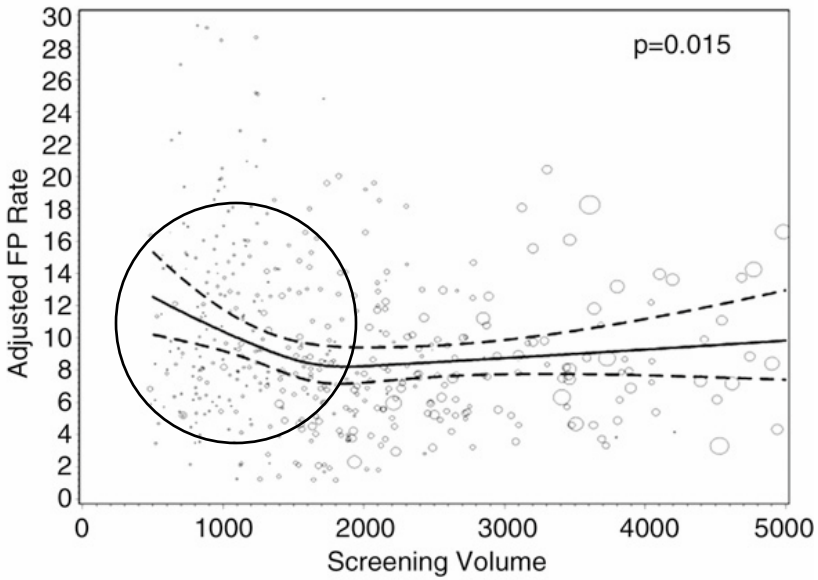
b.



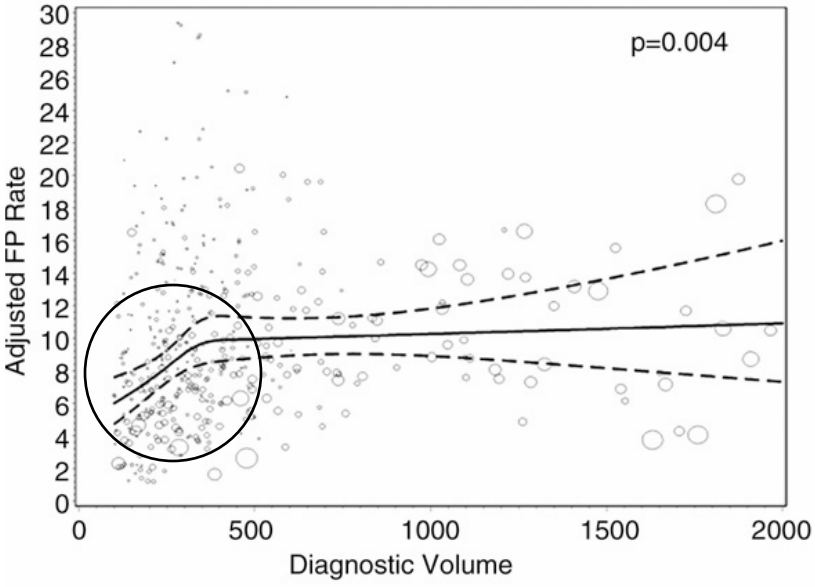
Buist *Radiology* 2011

False positive rate

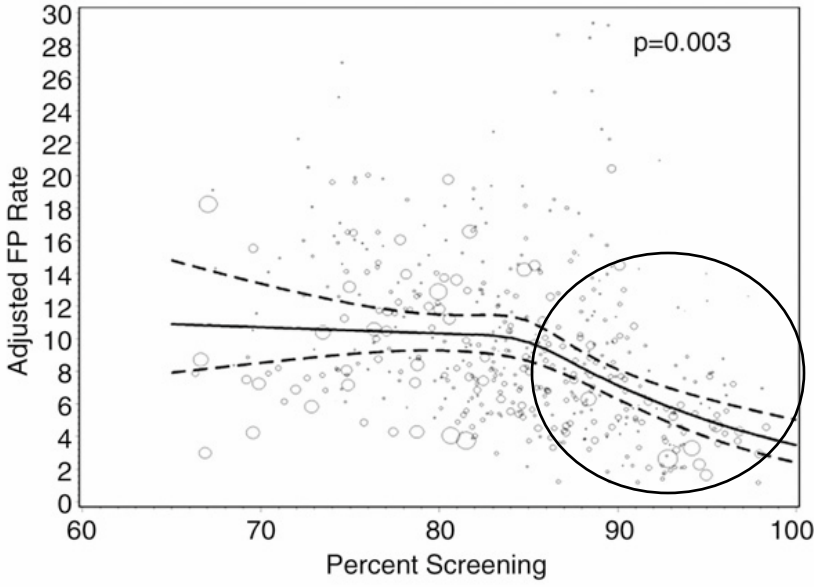




a.

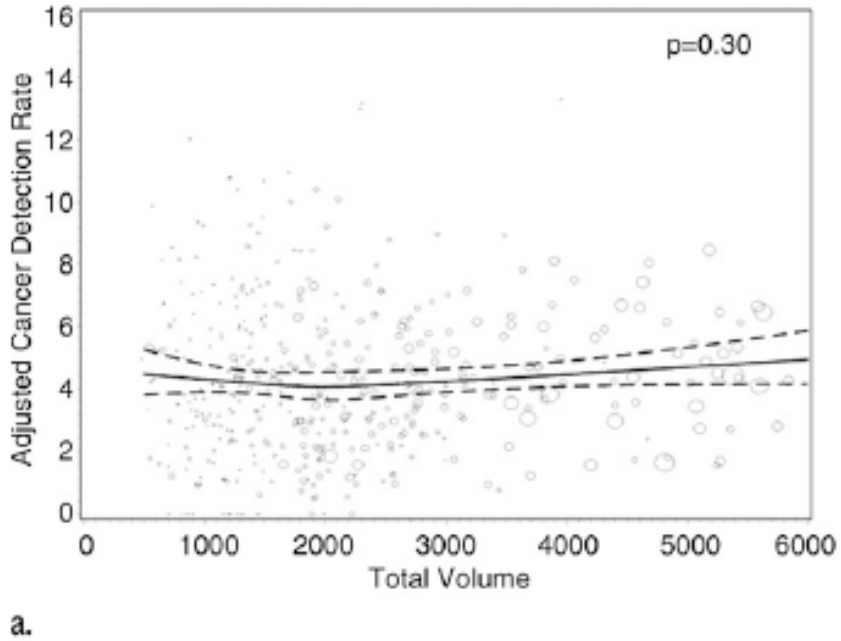


b.

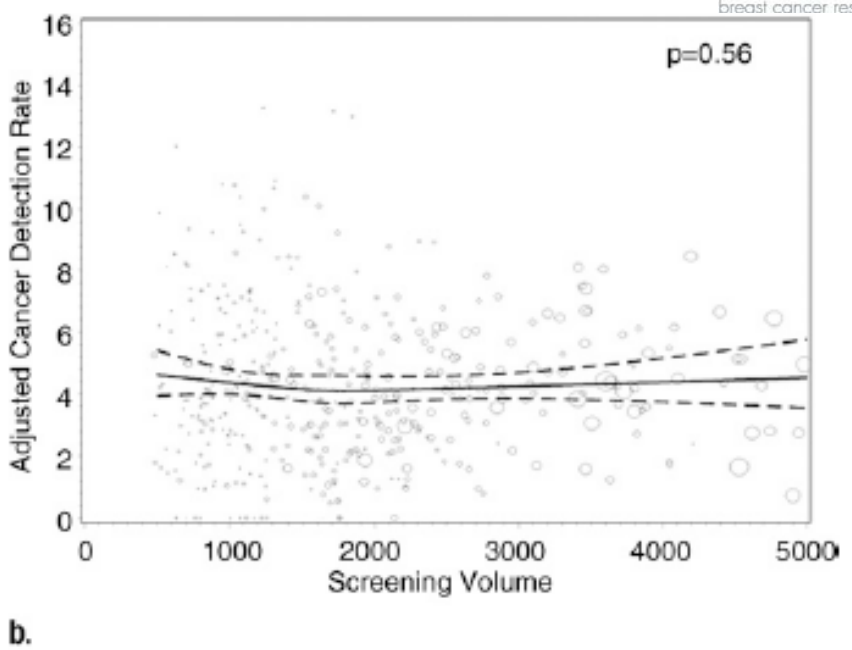


Buist *Radiology* 2011

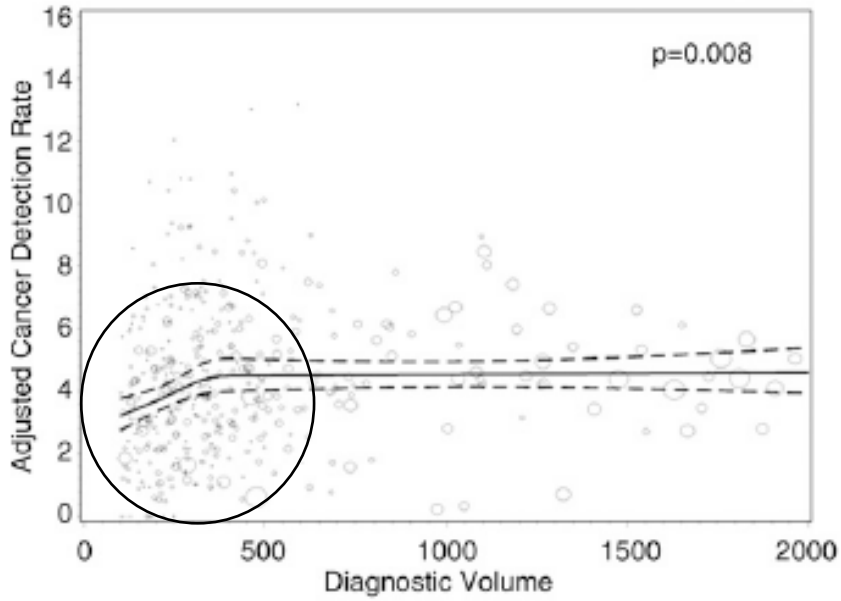
Cancer detection rate



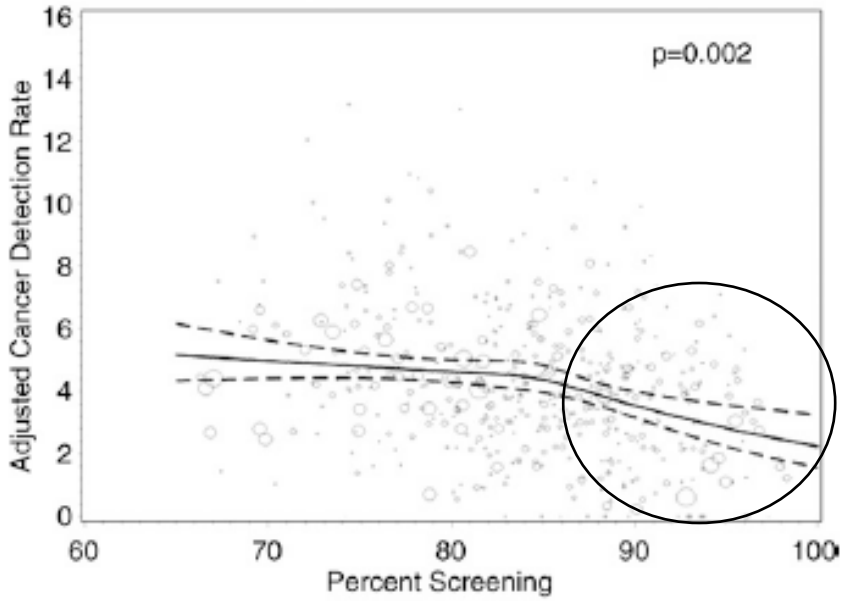
a.



b.



c.



d.

Buist *Radiology* 2011

Screening performance advances in knowledge

- Higher false positives with lower volumes (screening, diagnostic and total)
- Significantly lower cancer detection with low diagnostic volume and high % screening
- Significantly lower sensitivity with high % screening focus

Policy implications: Screening performance unlikely to be affected by volume alone balance in the interpreted examination composition

Diagnostic performance

Much less research on diagnostic performance

Want to weigh screening and diagnostic performance

- Subset of prior BCSC study
- 107 radiologists; 117,136 diagnostic mammograms (98,677 women)
  - 46,269 additional imaging
  - 70,767 symptomatic evaluation
- All analyses evaluated by exam indication

Diagnostic performance advances in knowledge

- No consistent association between total, screening and diagnostic volume and diagnostic performance
- False-positive rates were highest among radiologists who did interpreted <20% as diagnostic
- **Policy implications**: Diagnostic volume is a key determinant in developing thresholds for considering a diagnostic mammogram to be abnormal

Further exploration of composition of interpretive volume

- Volume did not explain much of the observed inter-radiologist variability in screening or diagnostic performance
- ***Composition of interpretive volume*** was the most important factor influencing screening and diagnostic performance

Does performance improve from working up your own recalled screening exams or just any diagnostic exams?

Breast Cancer Surveillance Consortium
BCSC
Working together to advance breast cancer research

What type of diagnostic work-up matters more?

Own

of radiologists' own recalled screening exams with diagnostic work up within 60 days

	annual ave	% distribution
Low	<25	38%
Med	26-50	24%
High	>50	39%

Any

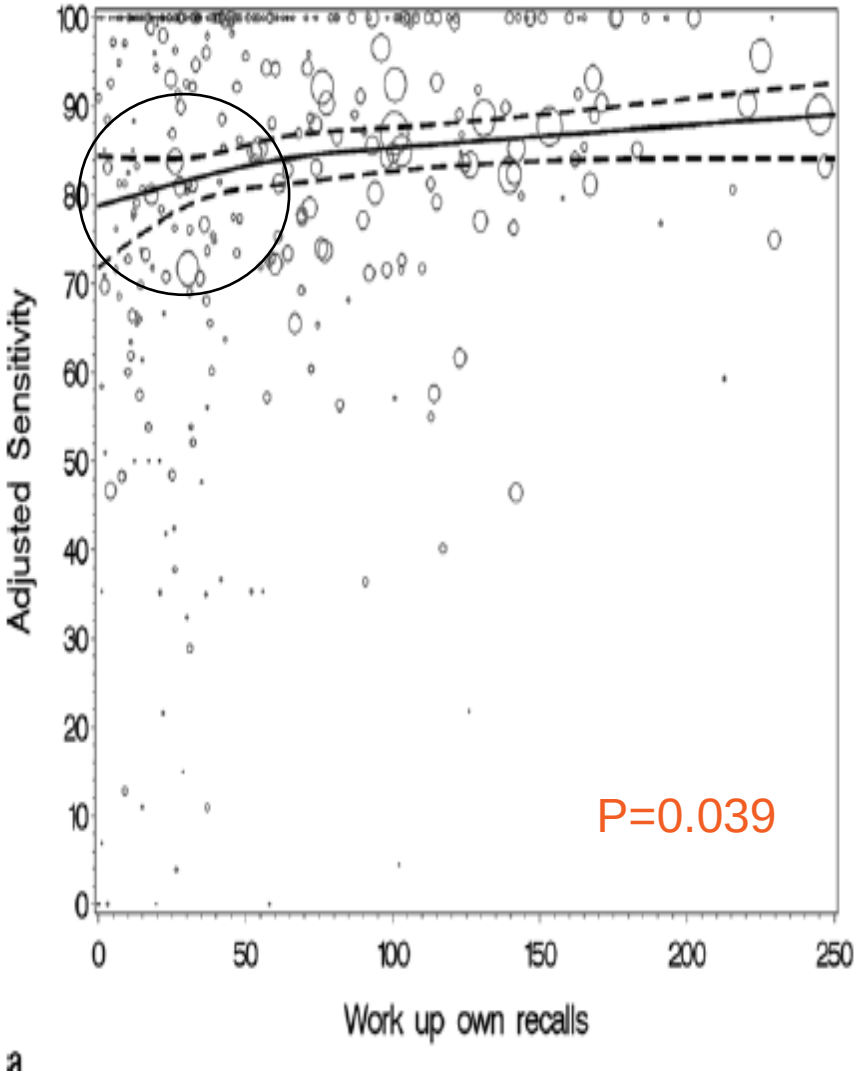
of any recalled screening exam regardless of who recalled

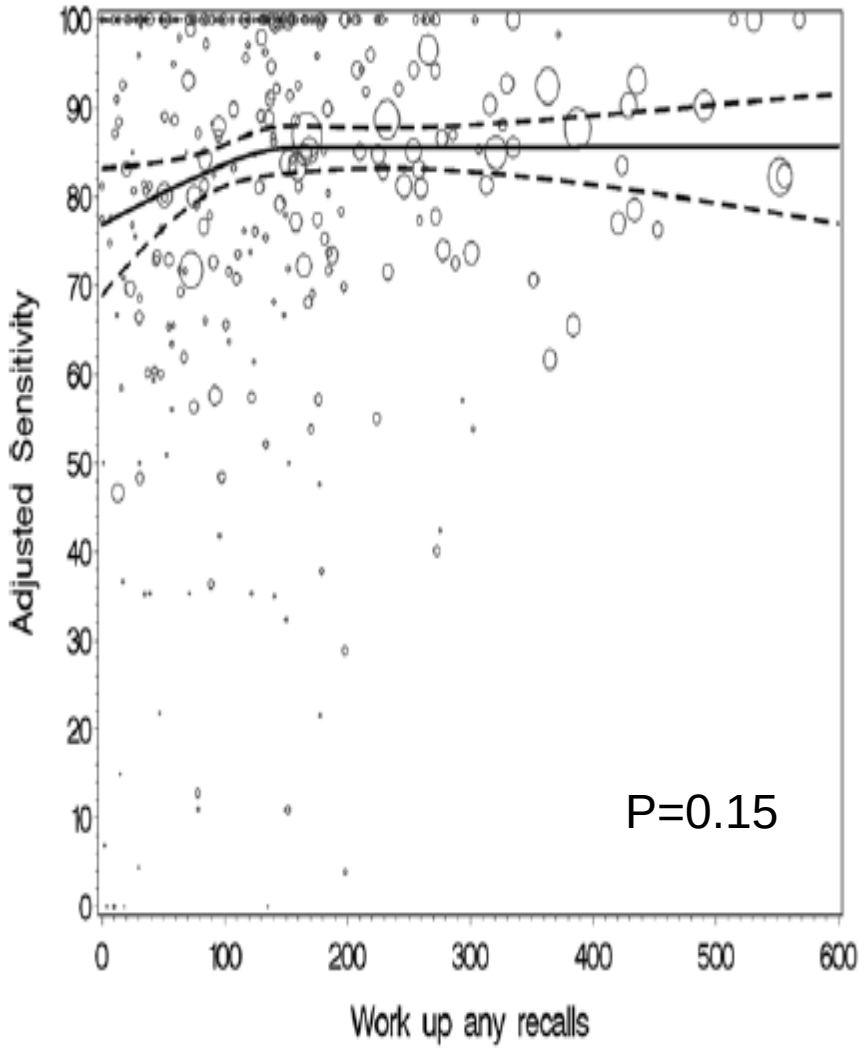
	annual ave	% distribution
Low	<50	24%
Med	51-125	32%
High	>125	44%

These are different measures
Pearson's correlation= 0.49

Buist *Radiology* 2014

Sensitivity by number of work-ups



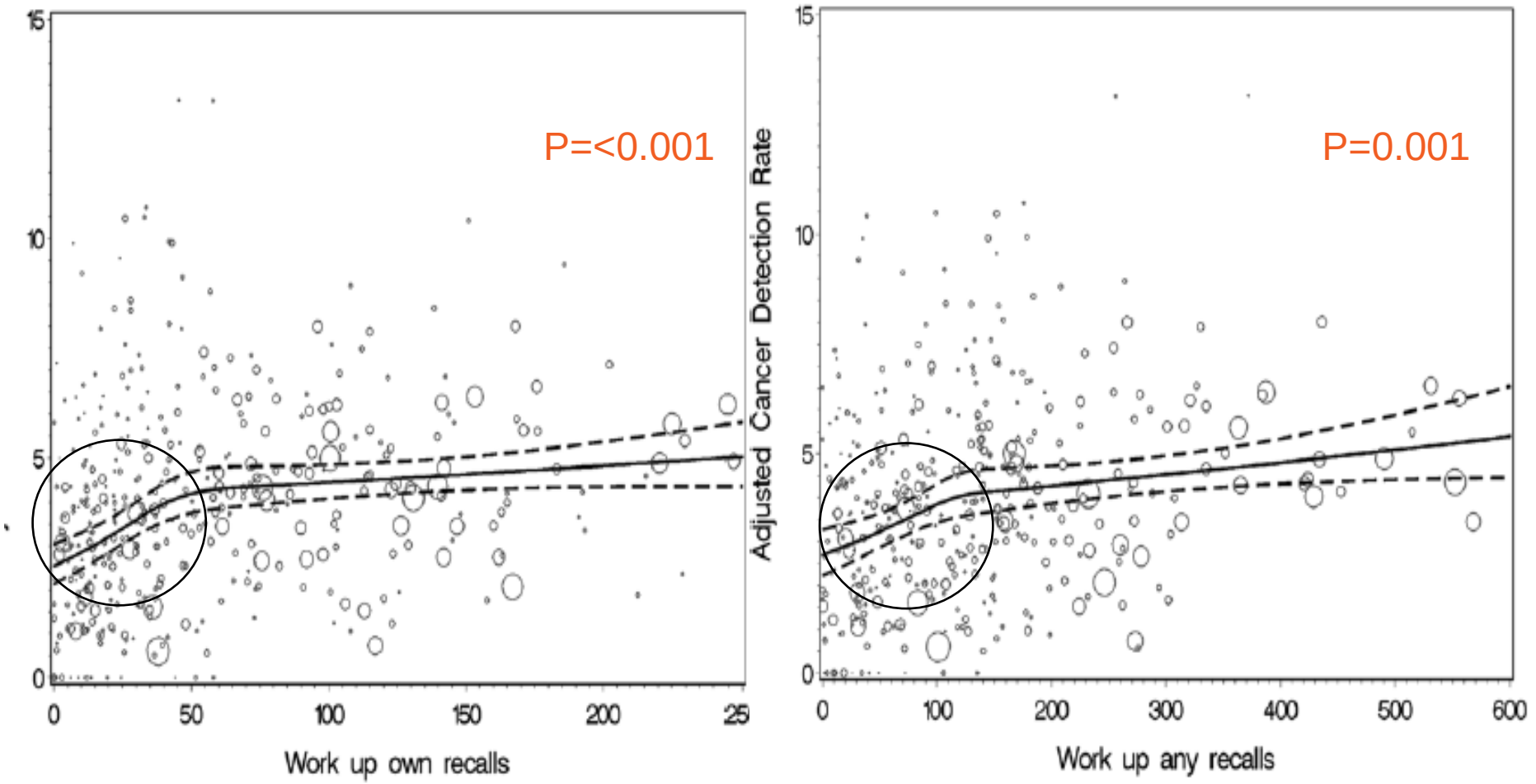


OWN

ANY

Buist *Radiology* 2014

Cancer detection rate by number of work-ups

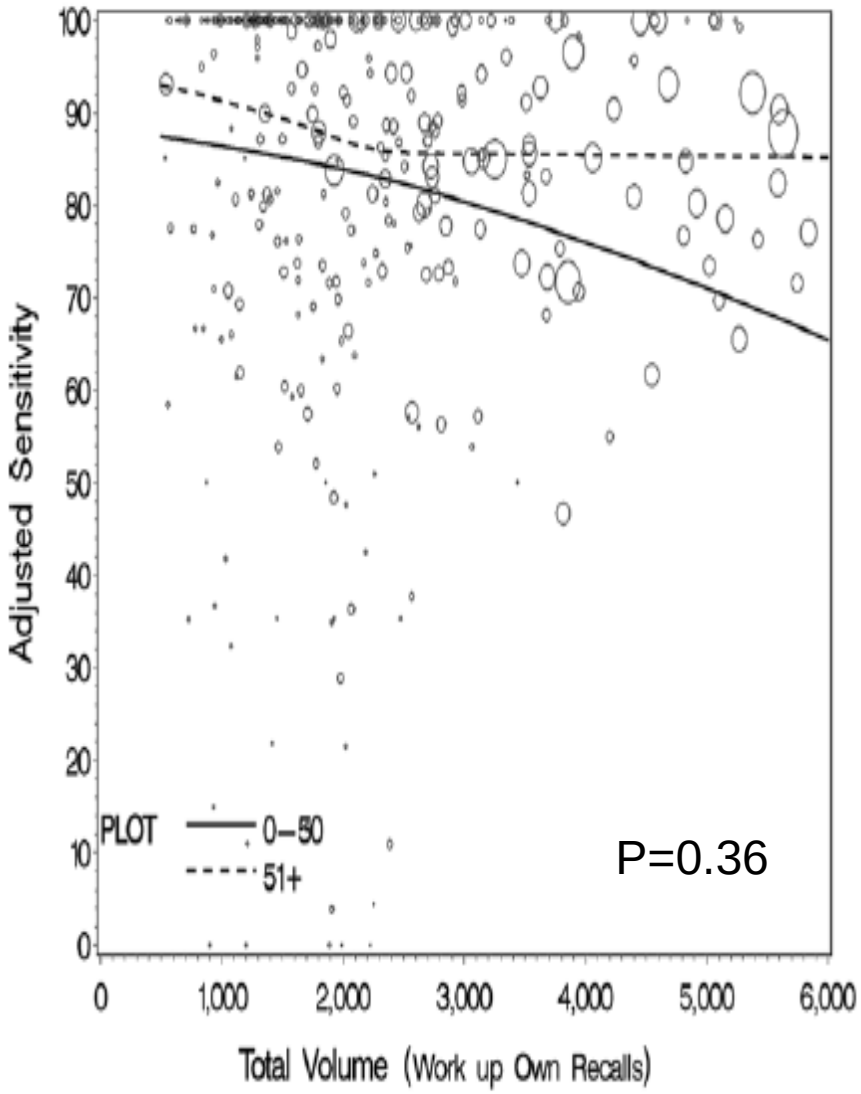


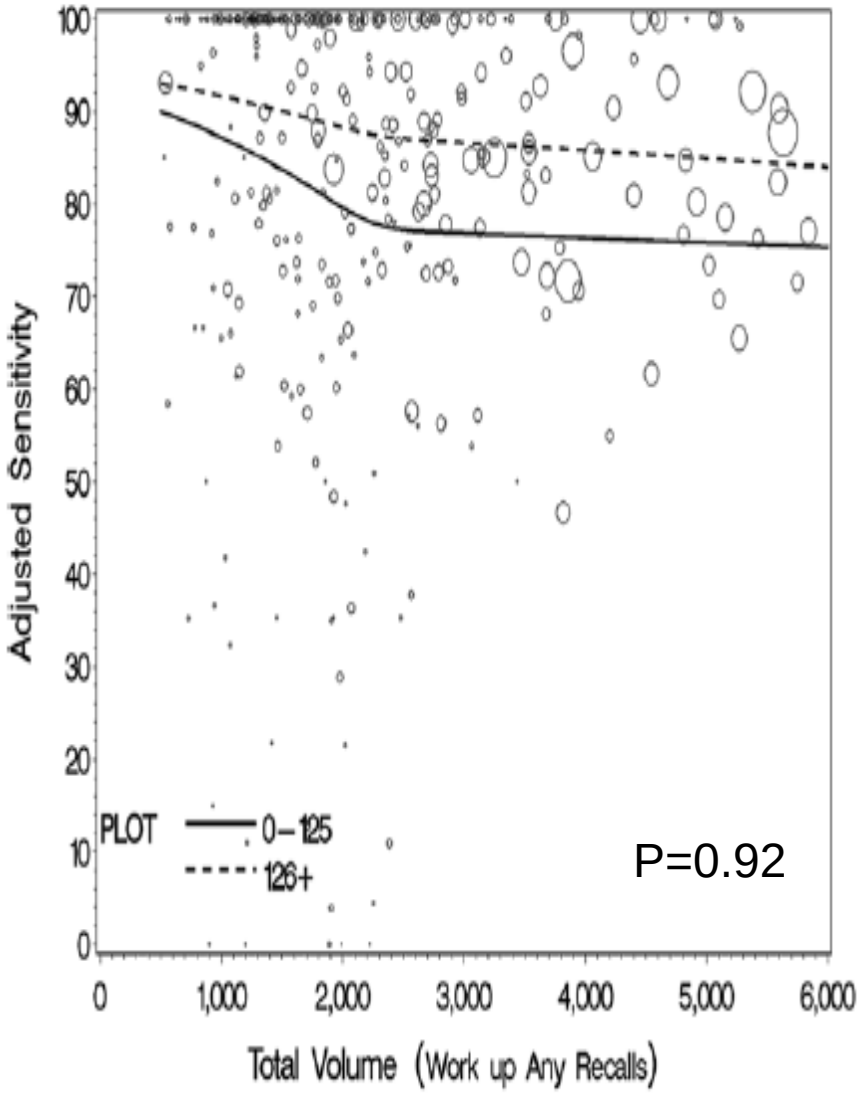
OWN

ANY

Buist *Radiology* 2014

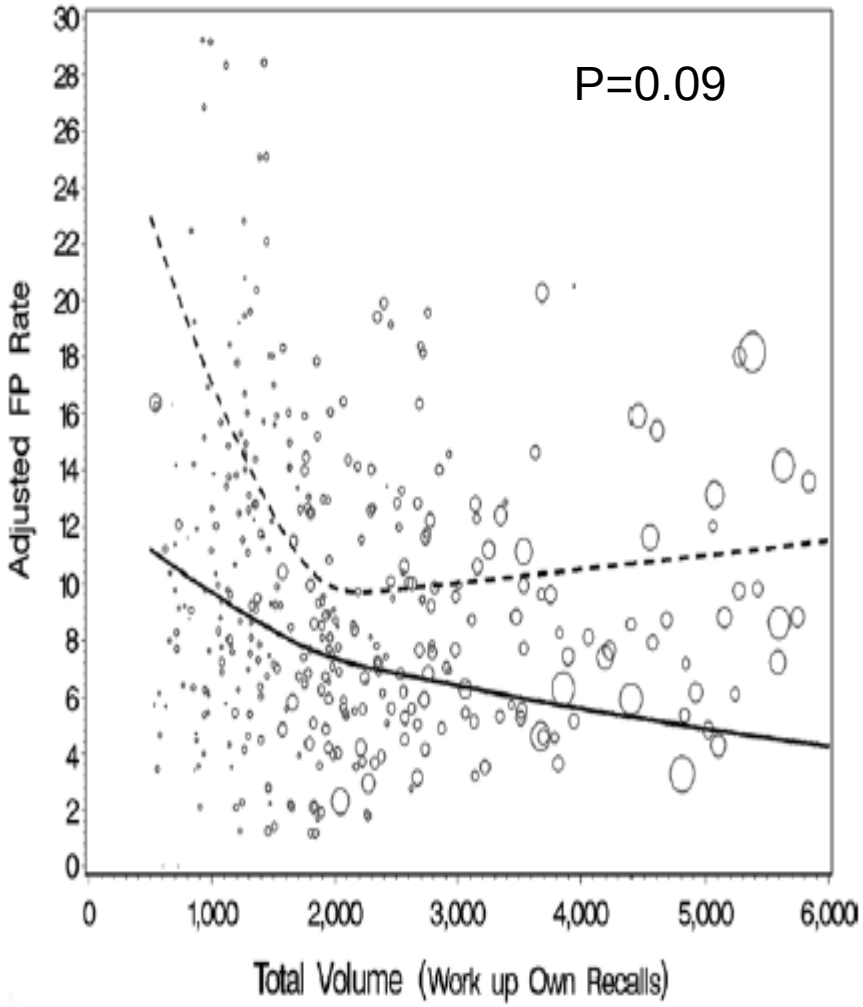
Sensitivity by total volume stratified by number of work-ups – low (solid), high (dashed)

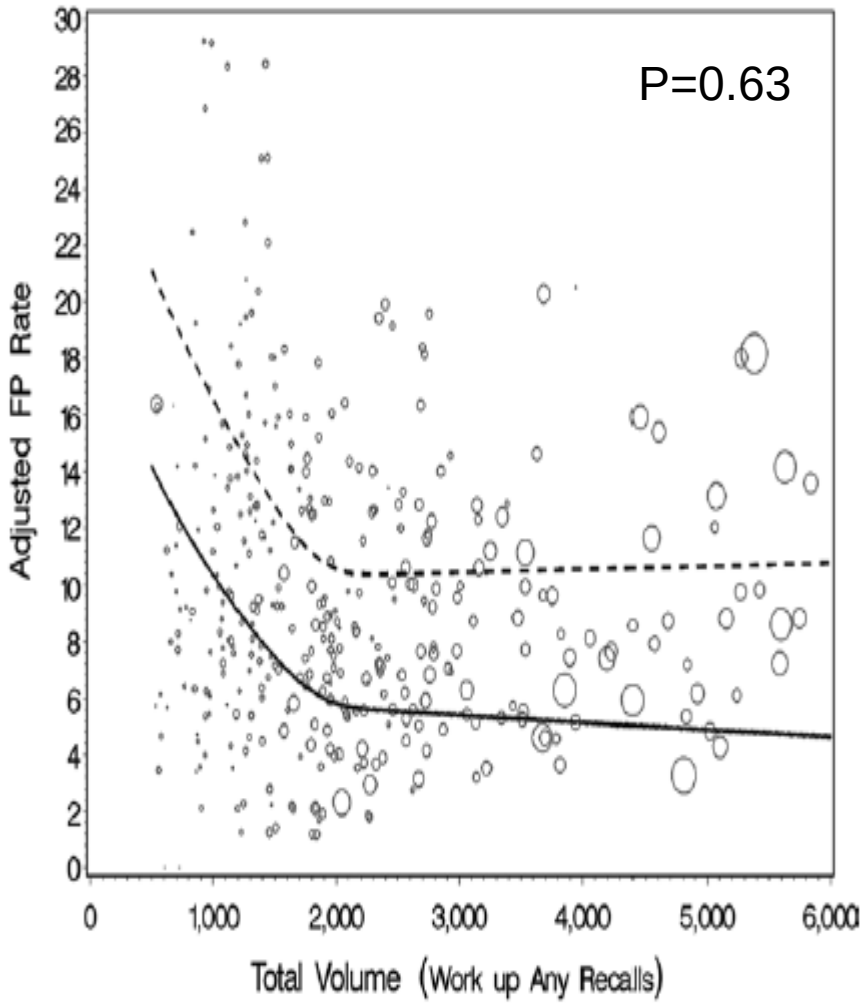




Buist *Radiology* 2014

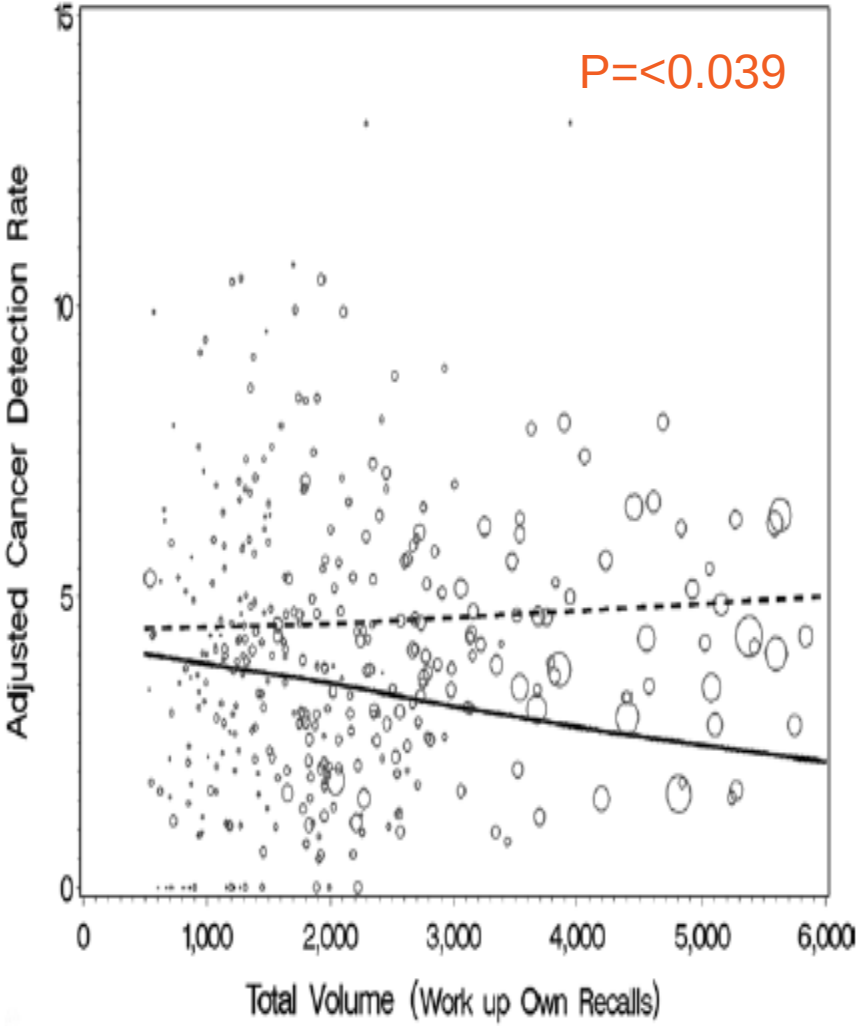
False-positive rate by total volume stratified by number of work-ups – low (solid), high (dashed)

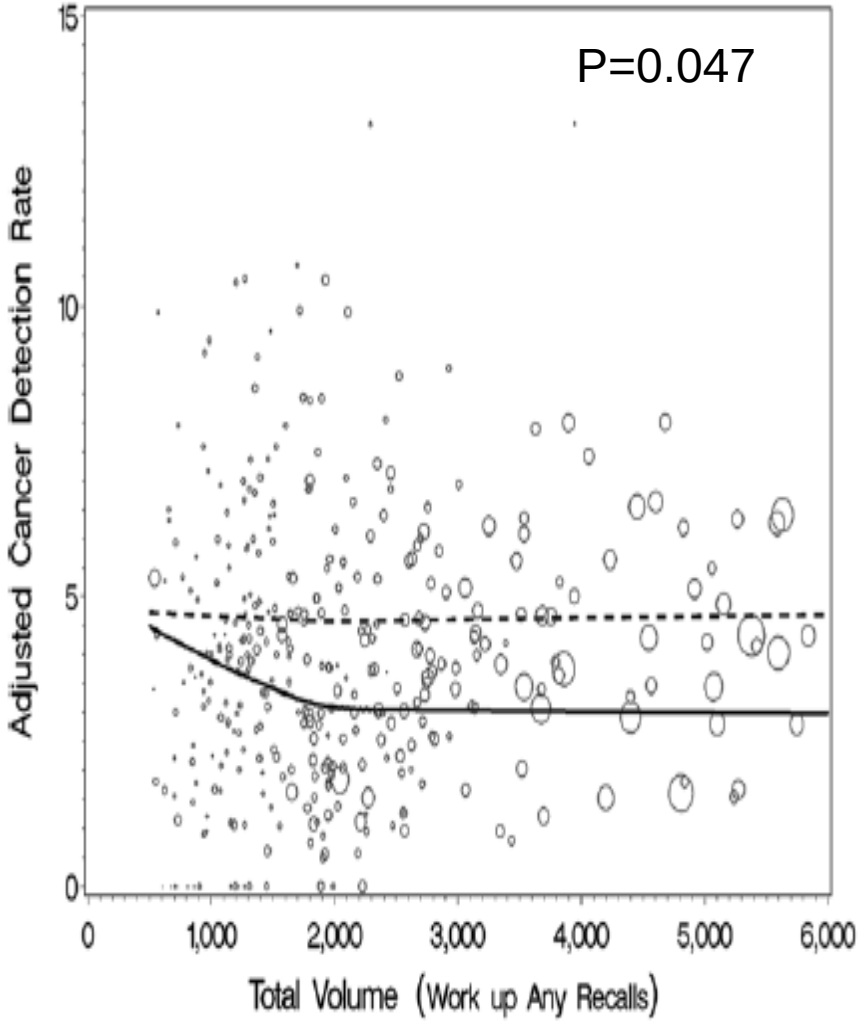




Buist *Radiology* 2014

Cancer detection rate by total volume stratified by number of work-ups – low (solid), high (dashed)



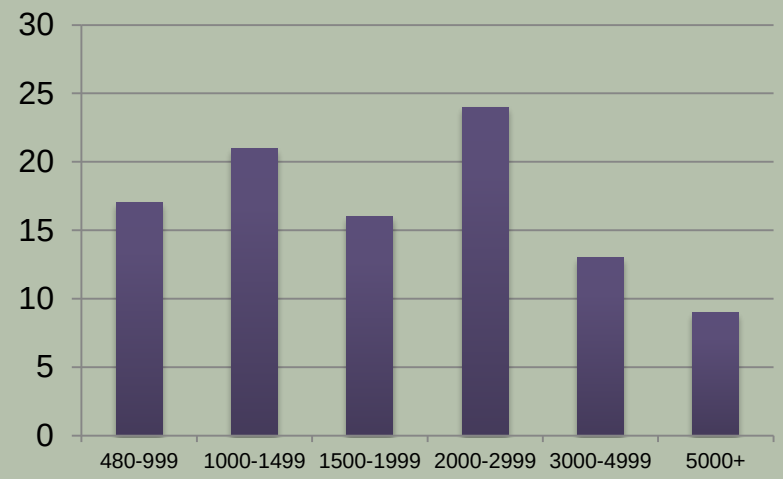


Buist *Radiology* 2014

Bringing it all together – Volume & Performance in the US

- Suggest consideration for increasing minimum interpretation volume requirements in the US
- Include a minimal diagnostic interpretation requirement – better to be some proportion of total volume
- Consider requiring a minimum number of diagnostic work-ups that resulted from radiologists' own recall

Impact



Own recalled exams

	annual ave	% distribution
Low	<25	38%
Med	26-50	24%
High	>50	39%

National Impact:

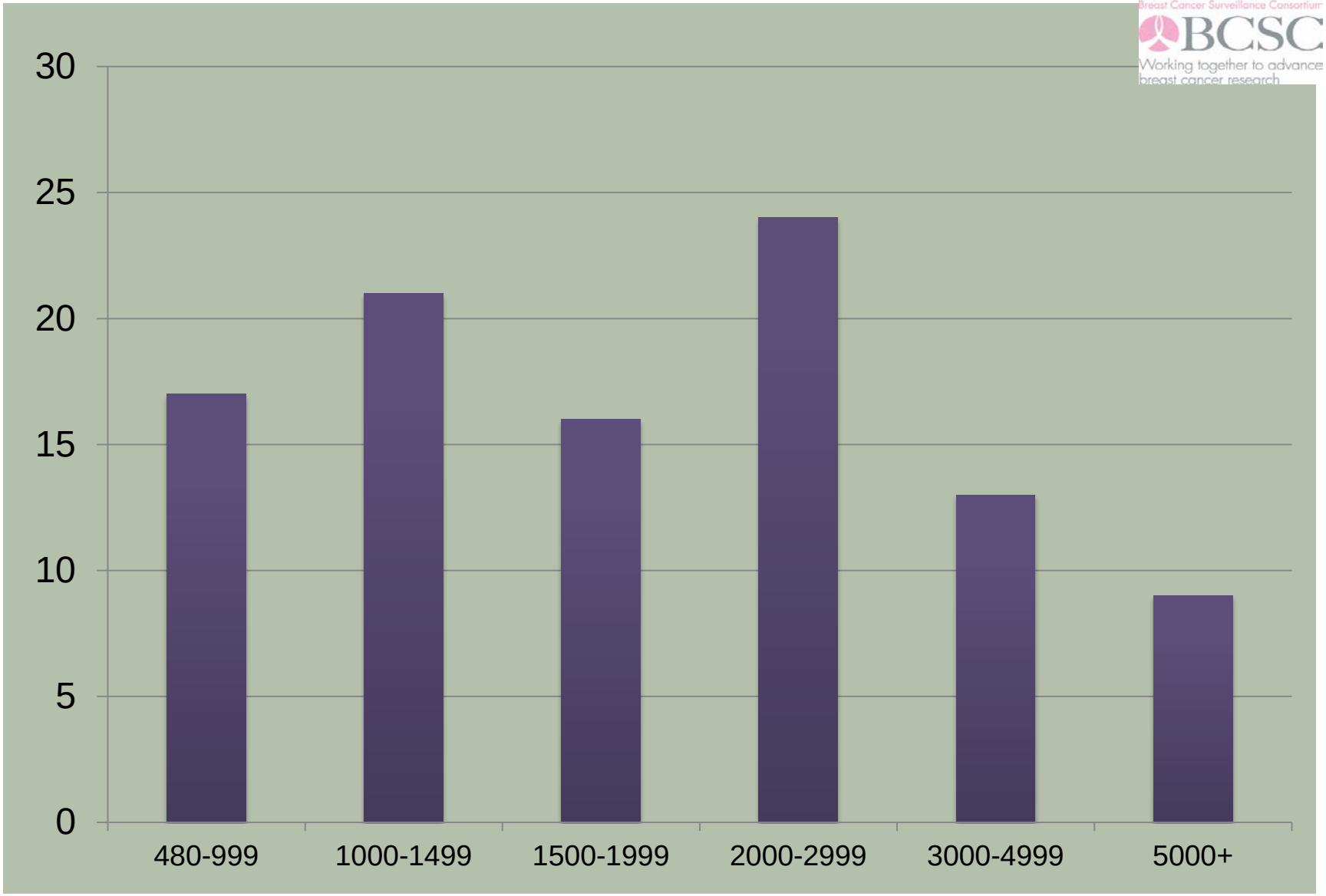
- Breast cancer screening costs >$3.6 billion annually in the United States
- ~$1.6 billion/year for costs of false positives (~$1.6 billion per year, or avoid time, trouble, and anxiety for women)
- Increasing annual screening increase would lower FP work-ups by
  - >1000: $35.6 million/year
  - >1500: $58.6 million/year

Have not updated simulation based on any joint criteria of total volume and composition of volume

*based on $107 US/screening mammogram

Thank you!

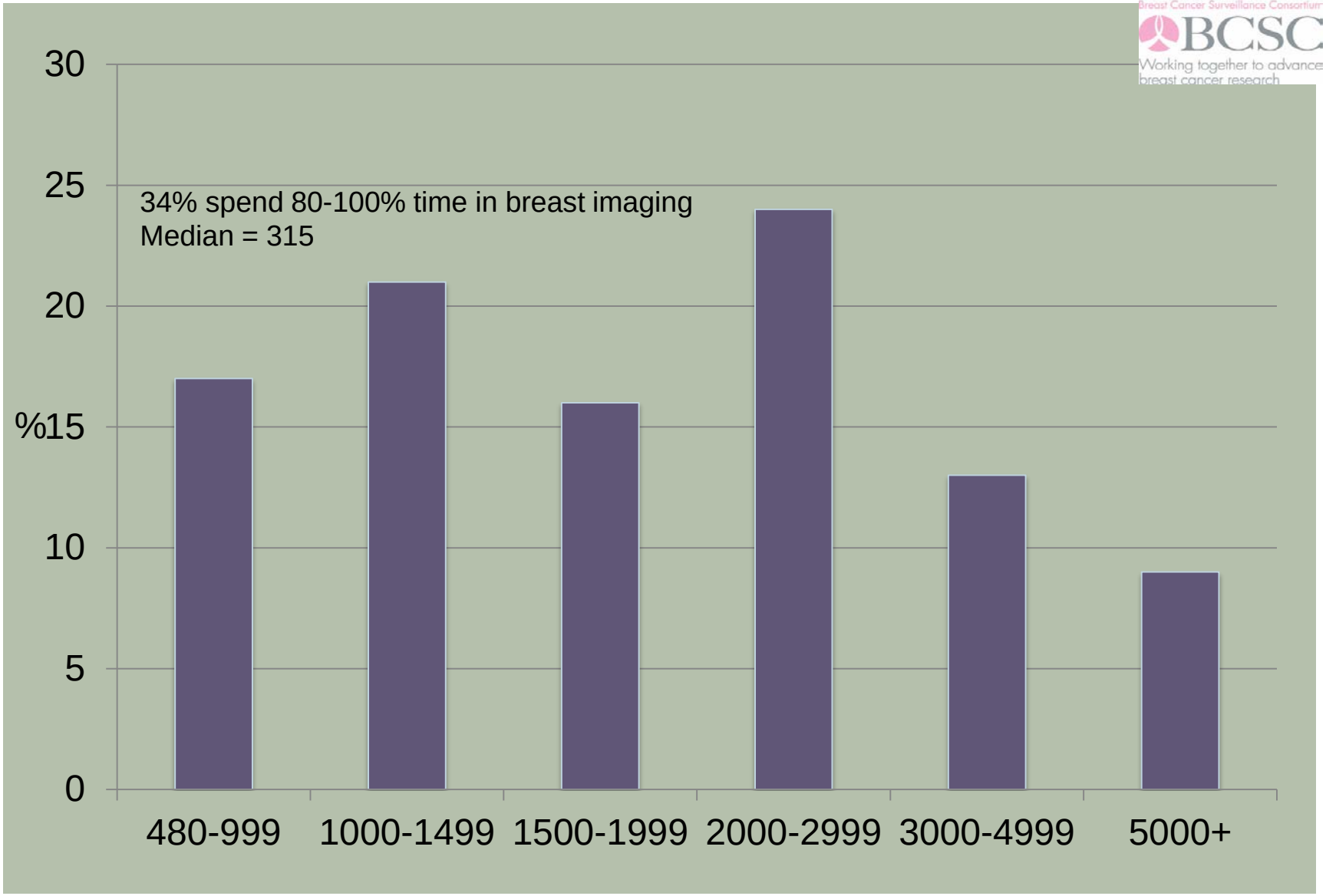
Distribution of annual screening volume interpretation in BCSC radiologists, US



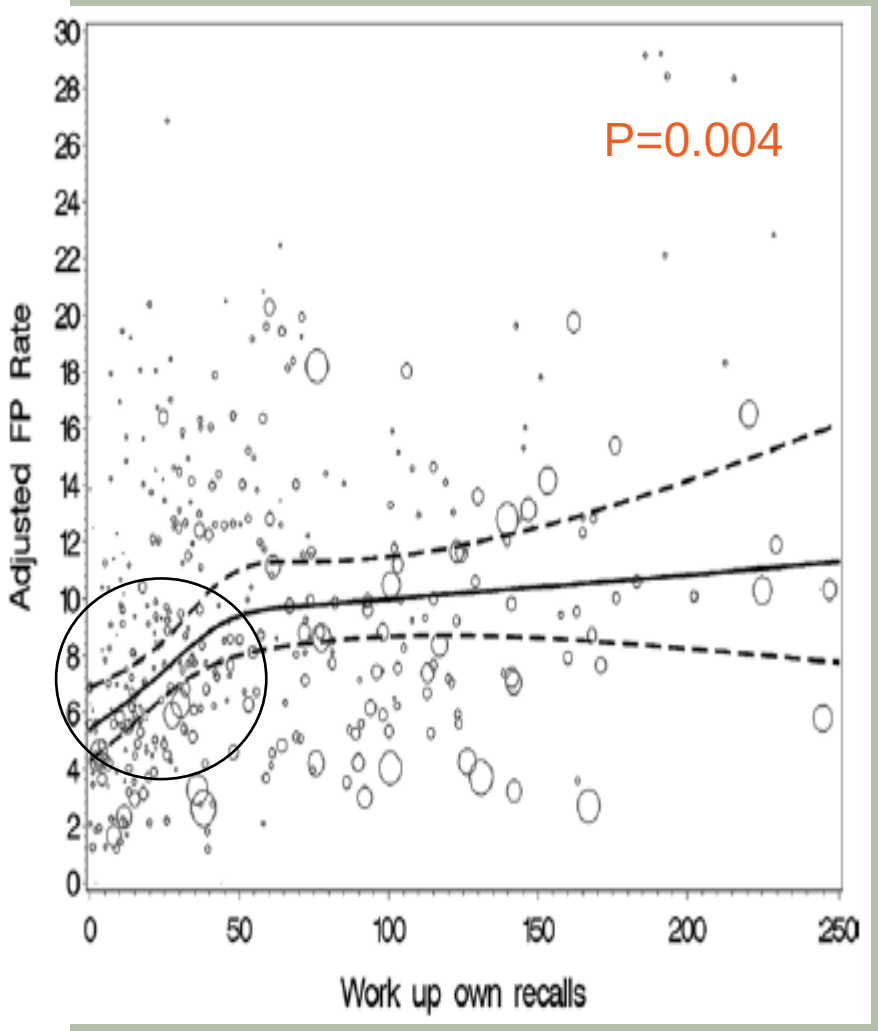
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
Advising the nation • Improving health

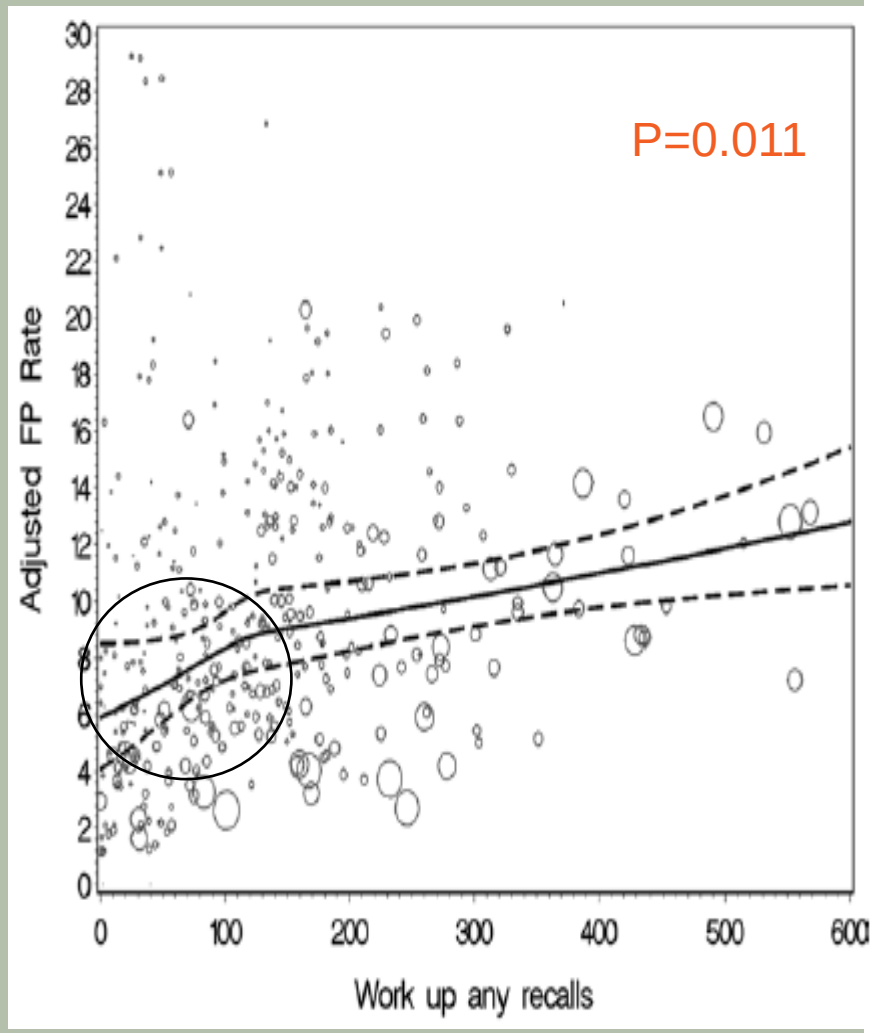
Buist *Radiology* 2011

Distribution of annual diagnostic volume interpretation in BCSC radiologists, US



False-positive rate by number of work-ups





INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
Advising the nation • Improving health

Buist *Radiology* 2014