Session 4: Test instruments to assess interpretive performance – challenges and opportunities

Overview of Test Set Design and Use

Robert A. Smith, PhD

American Cancer Society

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

Test Sets vs. Audits

Benefits of audits

- Feedback on practice patterns and outcomes
  - Sensitivity, Specificity, PPV, etc.
- Measure performance against a gold standard & peers
- Potential indications for corrective action

Test Sets vs. Audits (2)

Limitations of audits

- Reader volume and variable occurrence of important outcomes are limiting factors
  - Recalls are common; ***cancer is uncommon***
  - Can take several years for poor, or falling performance to be identified
  - May take several years to measure improvement after corrective action
  - General outcome measures tell us very little about areas that need improvement

Test Sets vs. Audits (3)

Limitations of audits

- Not all outcomes are easily measured due to lack of links to cancer registries
- Little empirical data on the effect of reviews of audit data on performance

The Logic for Test Sets

Over the past 25 years, there is been ample data to show wide variation in mammography interpretive skills

- While MQSA requires CME for interpreting physicians, there has been little persuasive evidence that it assured or improved interpretive skills
- The average radiologist has few opportunities to receive feedback on their performance, or to assess their performance

The Logic for Test Sets (2)

Test sets provide an opportunity to **set reference standards**, and measure:

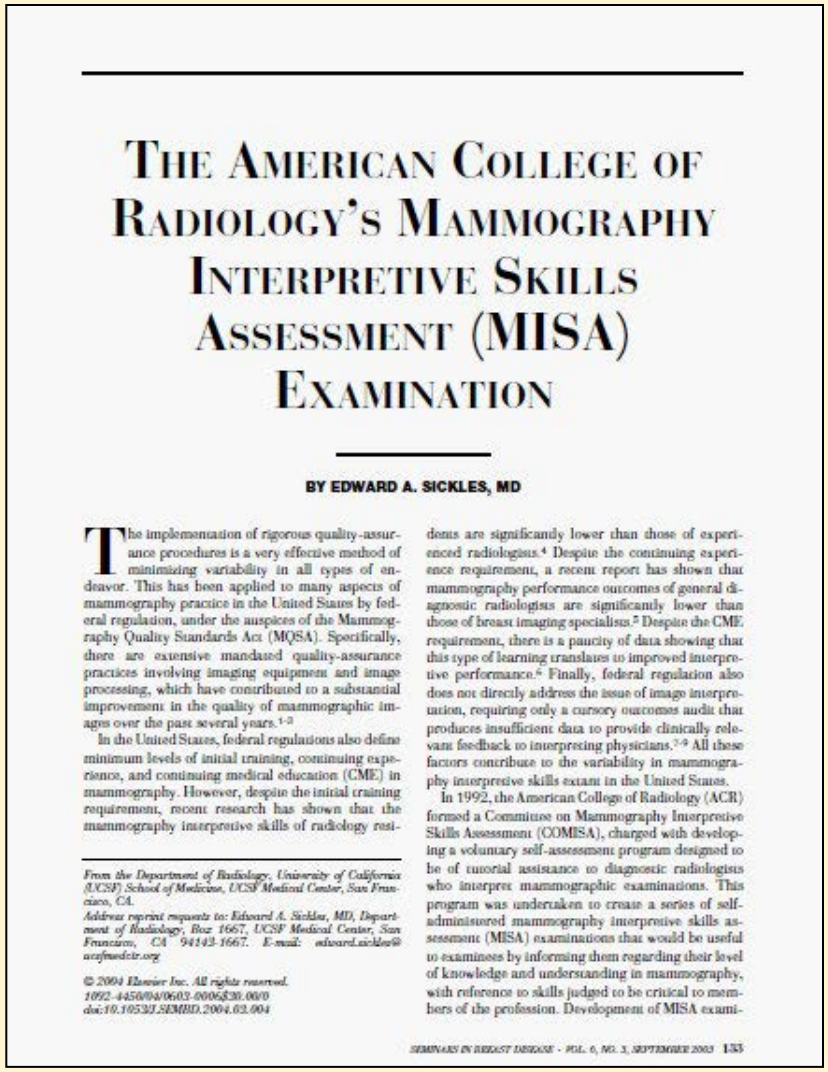
- Qualifying performance
- Pre- and post-intervention performance
  - Overall, or specific to particular needs for improvement
- Performance over time
- Performance on new imaging technology

The Logic for Test Sets (3)

Test sets provide an opportunity to **provide feedback** and learning that may be impossible to provide through audits, i.e., performance based on:

- Appearance of the mammogram (density) and abnormalities (calcifications, lesion characterization, etc.)
- Measures of sensitivity and specificity (truth)
- Judgment about recall, including false positives ***for which recall was appropriate*** (judgment)

Early Test Set Development—ACR—Early 1990s



THE AMERICAN COLLEGE OF RADIOLOGY'S MAMMOGRAPHY INTERPRETIVE SKILLS ASSESSMENT (MISA) EXAMINATION

BY EDWARD A. SICKLES, MD

The implementation of rigorous quality-assurance procedures is a very effective method of minimizing variability in all types of endeavor. This has been applied to many aspects of mammography practice in the United States by federal regulation, under the auspices of the Mammography Quality Standards Act (MQSA). Specifically, there are extensive mandated quality-assurance practices involving imaging equipment and image processing, which have contributed to a substantial improvement in the quality of mammographic images over the past several years.[1-3]

In the United States, federal regulations also define minimum levels of initial training, continuing experience, and continuing medical education (CME) in mammography. However, despite the initial training requirement, recent research has shown that the mammography interpretive skills of radiology residents are significantly lower than those of experienced radiologists.[4] Despite the continuing experience requirement, a recent report has shown that mammography performance outcomes of general diagnostic radiologists are significantly lower than those of breast imaging specialists.[5] Despite the CME requirement, there is a paucity of data showing that this type of learning translates to improved interpretive performance.[6] Finally, federal regulation also does not directly address the issue of image interpretation, requiring only a cursory outcomes audit that produces insufficient data to provide clinically relevant feedback to interpreting physicians.[7-9] All these factors contribute to the variability in mammography interpretive skills extant in the United States.

In 1992, the American College of Radiology (ACR) formed a Committee on Mammography Interpretive Skills Assessment (COMISA), charged with developing a voluntary self-assessment program designed to be of tutorial assistance to diagnostic radiologists who interpret mammographic examinations. This program was undertaken to create a series of self-administered mammography interpretive skills assessment (MISA) examinations that would be useful to examinees by informing them regarding their level of knowledge and understanding in mammography, with reference to skills judged to be critical to members of the profession. Development of MISA exami-

From the Department of Radiology, University of California (UCSF) School of Medicine, UCSF Medical Center, San Francisco, CA.
Address reprint requests to: Edward A. Sickles, MD, Department of Radiology, Box 1667, UCSF Medical Center, San Francisco, CA 94143-1667. E-mail: edward.sickles@ucsfmedctr.org

1092-4450/04/0603-0006$30.00/0
doi:10.1053/J.SEMBD.2004.03.004

In 1992, the American College of Radiology (ACR) formed a Committee on Mammography Interpretive Skills Assessment (COMISA), charged with developing a voluntary self-assessment program (MISA exams) designed to be of tutorial assistance to diagnostic radiologists who interpret mammographic examinations.

Sickles, E. Seminars in Breast Disease 2003;6:(3)

ACR MISA Examinations Were Principally Focused on Identification of Abnormalities and Management

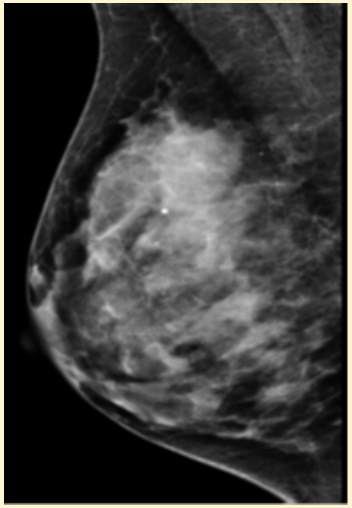
- Early exams were film based, paper & pencil tests
- Cases and questions were extensively field-tested to produce a pool of 400 usable items
- Where alternative management strategies were acceptable, 2 of 5 multiple choice options would be judged to be correct
- Usual exam was 30 cases, and about 125 questions

Explanations of the themes that categorize MISA examination questions.

Detection: Is there an abnormality? Point and click on the finding

Validation: Is it real? Identify quadrant and o'clock position

Analysis: Description of findings. What is the diagnosis?



Management: BI-RADS assessment categories, and management plans

Image Quality: Positioning, contrast, blur, noise, compression, and artifacts

Evolution of MISA from film to digital

Film-based instruction was challenging

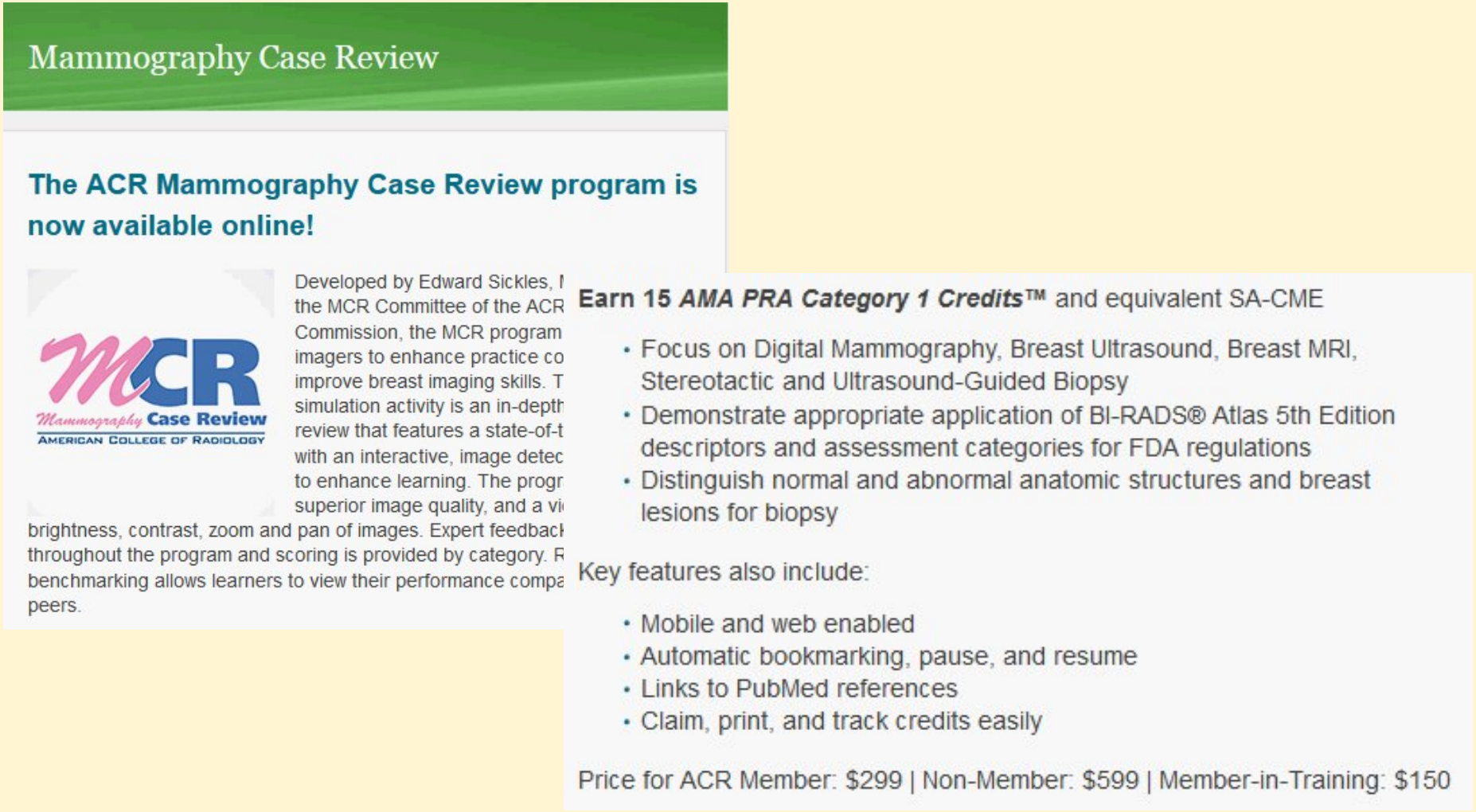
(1) Maintaining a sufficient number of high-quality copy films (films get damaged with use)

(2) Shipping and setting up large numbers of view boxes and renting hotel space for each test administration proved to be very expensive

(3) Staffing each test administration with required supervisory personnel also proved to be expensive

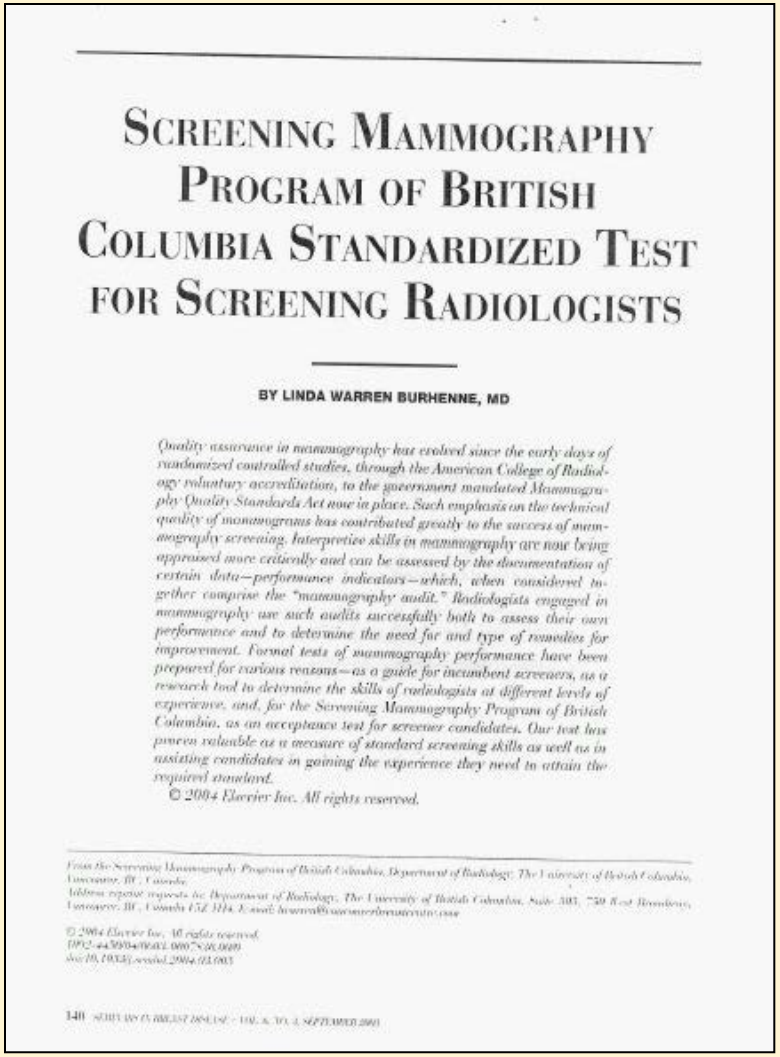
(4) Test results could not be given to users promptly, thus delaying and therefore reducing the value of the feedback

ACR MISA Evolved into a fully digital ACR Mammography Case Review



http://www.acr.org/Education/e-Learning/Mammo-Case-Review

Early Test Set Development—British Columbia Breast Cancer Screening Program



SCREENING MAMMOGRAPHY PROGRAM OF BRITISH COLUMBIA STANDARDIZED TEST FOR SCREENING RADIOLOGISTS

BY LINDA WARREN BURHENNE, MD

Quality assurance in mammography has evolved since the early days of randomized controlled studies, through the American College of Radiology voluntary accreditation, to the government mandated Mammography Quality Standards Act now in place. Such emphasis on the technical quality of mammograms has contributed greatly to the success of mammography screening. Interpretive skills in mammography are now being appraised more critically and can be assessed by the documentation of certain data—performance indicators—which, when considered together comprise the "mammography audit." Radiologists engaged in mammography use such audits successfully both to assess their own performance and to determine the need for and type of remedies for improvement. Formal tests of mammography performance have been prepared for various reasons—as a guide for incumbent screeners, as a research tool to determine the skills of radiologists at different levels of experience, and, for the Screening Mammography Program of British Columbia, as an acceptance test for screener candidates. Our test has proven valuable as a measure of standard screening skills as well as in assisting candidates in gaining the experience they need to attain the required standard.

From the Screening Mammography Program of British Columbia, Department of Radiology, The University of British Columbia, Vancouver, BC, Canada.
Address reprint requests to: Department of Radiology, The University of British Columbia, Suite 500, 750 West Broadway, Vancouver, BC, Canada V5Z 1H4. E-mail: lwarren@mammography.com

1092-4469/04/0303-0007$30.00/0
doi:10.1053/j.sonct.2004.03.003

- The Screening Program of British Columbia was established in 1989
- Organizers established minimum standards for interpreting physicians, including:
  - Recent CME in screening & diagnostic mammography
  - Minimum experience of having read 2500 mammograms
  - Satisfactory performance on a qualifying proficiency test

Screening Program of British Columbia Test Set

Test Set 1: 100 cases (copy films) based on original interpretation

- 39 abnormals (14 cancers)
- 61 normals (based on 2 year follow-up)
- Interpretation: Normal or Abnormal
- Sensitivity: 43% - 100%
- Specificity: 71% -81%
- Agreement on original abnormals: 42% -84%

Test Set 2: 100 cases (original films) based on original interpretation

- 50 malignant lesions
- 50 normal cases, including a mix of normal and abnormal original interpretations
- Interpretation: Normal or Abnormal
- Breast-based outcomes
  - Sensitivity: 73% - 100%
  - Specificity: 71% -94%

Screening Program of British Columbia Test Set

Test Set 3: 120 cases (copies) based on original interpretation

- 40 cancers (50% should be moderate to high suspicion)
  - 25 DCIS; 15 invasive
- 40 non-cancer abnormals
- 40 normals
- One-third all cases involved dense breast tissue

Cancer cases were judged according to detection by 4 expert readers:

- Obvious-- 4/4
- Intermediate-- 2-3/4
- Subtle-- 1/4Test
- **Pass/Fail Rate**
  - ***Sensitivity***-- Average sensitivity of the 4 readers +/- 10%
  - ***Specificity***– Percent of all non-cancer cases called normal by the 4 reviewers

Screening Program of British Columbia—Interplay between test sets, audits, and monthly reviews

1. Qualify on test set
   - Training available for those who do not qualify
2. Bi-monthly review of all screen-detected and interval cancers
3. Annual review of individual and program performance

Burhenne, LW, Seminars in Breast Disease 2004;6(3)

Early Test Set Development—Proficiency Testing in Italy, Ciatto, et al. , 1999

J Med Screen 1999;6:149–151 149

Proficiency test for screening mammography: results for 117 volunteer Italian radiologists

S Ciatto, D Ambrogetti, S Catarzi, D Morrone, M Rosselli Del Turco

Abstract

Objective—To analyse the performance of a large sample of Italian radiologists undergoing a proficiency test for screening mammography.

Design—Evaluation of performance indicators according to reference standards determined by a panel of experts (sensitivity (reference standard ≥80%), recall rate (reference standard ≤15%)).

Setting—117 Italian radiologists of varying experience (years of practice 0.5–18, average 5.9; mammograms read 500–51 000, average 13 000), all currently reporting clinical mammography and planning to take part in screening in the near future.

Results—Eighty four of 117 (72%) radiologists reached the standard for sensitivity, 88 (75%) reached the standard for recall rate, and only 59 (50%) reached both standards and passed the proficiency test. The probability of passing the test was significantly correlated with mammographic practice (p=0.015), mammograms read (p=0.024), and mammograms read/year (p=0.043).

Discussion—The performance of a large sample of Italian radiologists currently reporting clinical mammography was disappointing, indicating the need for proper training of at least 50% of the tested subjects. When implementing organised screening the health authority should set up a proper process for training and accrediting radiologists, and a proficiency test should be part of such a process.

(J Med Screen 1999;6:149–151)

Keywords: breast cancer; mammography; quality control; training; proficiency test

Centro per lo Studio e la Prevenzione Oncologica, viale A. Volta 171, I-50131 Florence, Italy
Department of Diagnostic Imaging
S Ciatto, head of unit

Breast Unit
D Ambrogetti, radiologist
S Catarzi, radiologist
D Morrone, radiologist
M Rosselli Del Turco, head of unit

Correspondence to:
Dr Ciatto

Accepted for publication
18 August 1999

The variability of radiologists in interpreting mammograms is well known and has been documented in several studies.[1–4] As accuracy in mammography is expected to be related to experience,[5] training is currently recommended for radiologists before they read screening mammograms.[6]

Population based mammographic screening has been taking place in Florence since 1970[7] at the Centro per lo Studio e la Prevenzione Oncologica (CSPO), and for many years this has been the only screening experience in Italy. CSPO is a reference centre for breast cancer screening in Italy, and many radiologists, generally interested in the screening issue and planning to take part in a screening programme in their region, do currently attend the CSPO for training. As part of the training programme available at the CSPO, a proficiency test was developed in 1997.

This report evaluates the results of the proficiency test on a large series of Italian radiologists who volunteered for training. The implications of test results on the feasibility of screening in the Italian scenario are discussed.

Material and methods

The test file was prepared by one of us (SC). Seventeen mammograms from patients with cancer were interspersed among 133 non-cancer original screening mammograms (two views: oblique + craniocaudal), giving a total of 150 cases in the set. Non-cancer cases were randomly selected from screening archives, having been obtained during the 1993 screening round of the Florence district programme. In non-cancer cases, cancer was excluded as (a) the mammogram had been reported as negative in 1993 and at the subsequent screening test in 1995–96, or (b) a benign lesion had been confirmed at histological examination in a minority of cases undergoing open biopsy. Cancer cases were selected randomly from among screen detected breast cancers diagnosed during the 1993 screening round. Cases were randomly mixed, numbered in sequence, and mounted on a multiple viewer in progressive numerical order. The cost of preparing the test was around 1500 euros.

When running the test, operators had to report the following on a predefined proforma: (a) the progressive number, and (b) the side (right or left) of cases in which they identified abnormalities prompting referral for further diagnostic assessment. Indicators of test performance were sensitivity (SENS = proportion of cancer cases referred) and referral rate (RR = proportion of non-cancer cases referred). To establish performance standards a panel of four experienced (>20 000 screening mammograms read) CSPO radiologists (MRDT, DA, DM, SCa) performed the test. Average SENS was 95% (range 88–100) and average RR 8% (range 3–14). Standards for SENS and RR were set at >80% (about 15% less than the average panel result, 8% lower than the worst panel result) and ≤15% (double the average panel result, higher than the worst panel result) respectively. Both SENS and RR standards had to be reached to pass the test.

The test has been available at the CSPO in Florence since September 1997. Information on the test was circulated at meetings, congresses, and at the Italian Group for Mammographic Screening (GISMA), a spontaneous working group operating since 1995 and coordinating the results of all Italian centres with a current

- In Florence Italy, the Centro per lo Studio e la Prevenzione Oncologica (CSPO) was responsible for training radiologists in mammography
- In 1997 CSPO developed a proficiency test consisting of 150 cases, incluidng
  - 17 cases with breast cancer
  - 133 normal cases (previously read as normal, or recalled, but determined to be benign
- Reference Standard:
  - Sensitivity ≥ 80%
  - Recall rate ≤ 15%

CSPO Test Set Results, 1999

Table 1 Distribution of subjects according to test results

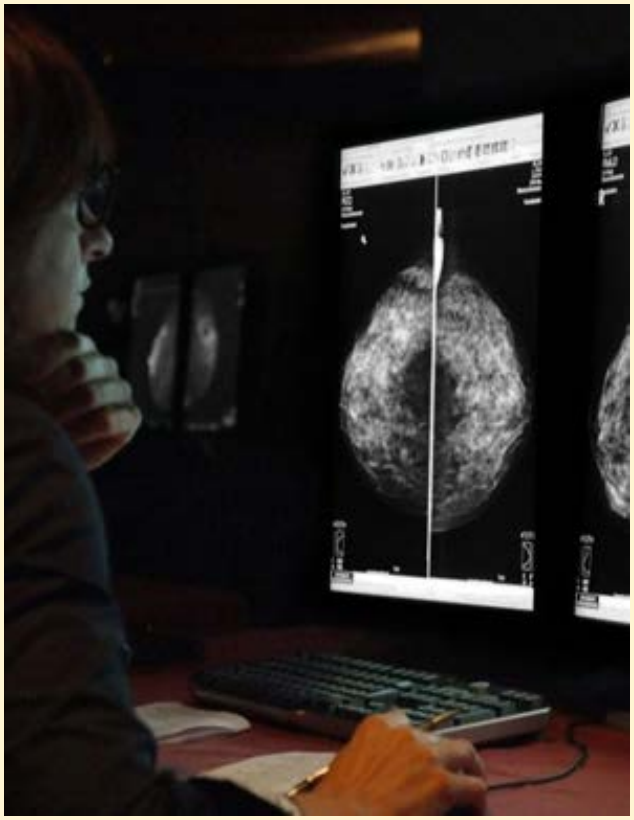
Sensitivity (%)	*Recall rate (%)*	*Number (%) of subjects*
>80	⩽15	59 (50)
>80	>15	25 (21)
⩽80	⩽15	29 (25)
⩽80	>15	4 (3)

- Average sensitivity was 82%, and 72% (84/117) met the reference standard
- Average recall rate was 12.6%, and 75% (88/117) met the standard
- *50% radiologists did not meet pre-set reference criteria*

CSPO Test Set--Conclusion

- “Previous experience in reading mammograms is associated with good test results, but experience indicators are not sufficient in themselves to accredit radiologists to read screening mammograms.”

- “Clearly, as for other countries, the health authority responsible for the implementation of a screening programme should provide proper training and accreditation of radiologists. ***A proficiency test, such as the one evaluated here, should be part of such a process.”***



Early Test Set Development—Investigations by Beam, et al (1996)

ORIGINAL INVESTIGATION

Variability in the Interpretation of Screening Mammograms by US Radiologists

Findings From a National Sample

Craig A. Beam, PhD; Peter M. Layde, MD, MSc; Daniel C. Sullivan, MD

Objective: To evaluate the effectiveness of screening mammography by estimating the variability in radiologists' ability to detect breast cancer within the US population of radiologists at mammography centers accredited by the American College of Radiology.

Methods: A two-way sample survey design was used as follows. Fifty mammography centers having an American College of Radiology–accredited unit were randomly sampled from across the United States. One hundred eight radiologists from these centers gave blinded interpretation to the same set of 79 randomly selected screening mammograms. The mammograms were from women who had been screened at a large screening center. Before their sampling, these women had been stratified by their breast disease status, established either by biopsy or by 2-year follow-up. Rates of biopsy recommendations were summarized by the mean, median, minimum, maximum, and range of sensitivity and specificity. Overall cancer detection ability was summarized by similar statistics for receiver operating characteristic curve areas. Ninety-five percent lower confidence bounds on the ranges in accuracy measures were established by bootstrapping.

Results: There is a range of at least 40% among US radiologists in their screening sensitivity. There is a range of at least 45% in the rates at which women without breast cancer are recommended for biopsy. As indicated by receiver operating characteristic curve areas, the ability of radiologists to detect cancer mammograms varies by as much as 11%.

Conclusions: Our findings indicate that there is wide variability in the accuracy of mammogram interpretation in the population of US radiologists. Current accreditation programs that certify the technical quality of radiographic equipment and images but not the accuracy of the interpretation given to mammograms may not be sufficient to help mammography fully realize its potential to reduce breast cancer mortality.

(Arch Intern Med. 1996;156:209-213)

From the Department of Family and Community Medicine, Medical College of Wisconsin, Milwaukee (Drs Beam and Layde), and the Department of Radiology, Duke University Medical Center, Durham, NC (Dr Sullivan). Dr Sullivan is now with the Department of Radiology, University of Pennsylvania Medical Center, Philadelphia.

IT IS WIDELY recognized that differences in mammogram interpretation among radiologists can influence the number and stage of progression of detected cancers and thereby influence any effect of screening with mammography on breast cancer mortality.[1] Although variability in the accuracy of mammogram interpretation has been investigated, to our knowledge none of the studies published to date provide findings that are directly generalizable to populations of practicing radiologists. This limitation comes about because each of the previous studies used only a small number of nonrandomly selected readers and consequently failed to reflect the actual extent of variability in practice. The most recent example is a study based on 10 radiologists nonrandomly selected from academic and private practices.[2] Another important limitation to the published studies is that none of them have accounted for the influence of sampling error on their findings and conclusions.

We present findings from the first study (to our knowledge) of variability in the diagnostic accuracy and interpretation of mammographic screening based on a large random sample of US radiologists from centers with mammography screening units accredited by the American College of Radiology (ACR).

RESULTS

Figure 1 presents box plots displaying the distribution by disease status of women

See Subjects and Methods on next page

- Fifty U.S. mammography centers having an ACR accredited unit were randomly sampled
- 108 radiologists gave blinded interpretation to the same set of 79 randomly selected screening mammograms.
- Wide range of sensitivity and specificity identified among participating radiologists

Beam C, et al. Arch Intern Med. 1996;156:209-213

Composition of the Beam, et al. Test Set—
(A) Film-based (Copies) Mailed to Study Subjects;
(B) Cancers Comprised About 57% of Cases
(C) Cancer Prevalence Varied by Age

Table 1. Age and Disease Composition of Sampled Women

Disease Group	By Age Group, No. (% of Disease Group) 40-49	50-59	60-69	All Age Groups, No. (% of All Patients)
Normal*	5 (26.4)	7 (36.8)	7 (36.8)	19 (24.1)
Benign	5 (33.3)	4 (26.7)	6 (40.0)	15 (19.0)
Cancer	9 (20.0)	15 (33.3)	21 (45.0)	45 (56.9)
Total (% of all women)	**19** (24.1)	**26** (32.9)	**34** (43.0)	**79**

**Women with normal mammograms and women with benign mammographic findings over 2 years of follow-up.*

Challenges in Assessing Reader Performance

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Review

Assessing reader performance in radiology, an imperfect science: Lessons from breast screening

B.P. Soh*, W. Lee, P.L. Kench, W.M. Reed, M.F. McEntee, A. Poulos, P.C. Brennan

Medical Image Optimisation and Perception Group (MIOPeG), Faculty of Health Sciences, University of Sydney, Lidcombe, NSW, Australia

ARTICLE INFORMATION

Article history:
Received 24 November 2011
Received in revised form
7 February 2012
Accepted 13 February 2012

The purpose of this article is to review the limitations associated with current methods of assessing reader accuracy in mammography screening programmes. Clinical audit is commonly used as a quality-assurance tool to monitor the performance of screen readers; however, a number of the metrics employed, such as recall rate as a surrogate for specificity, do not always accurately measure the intended clinical feature. Alternatively, standardized screening test sets, which benefit from ease of application, immediacy of results, and quicker assessment of quality improvement plans, suffer from experimental confounders, thus questioning the relevance of these laboratory-type screening test sets to clinical performance. Four key factors that impact on the external validity of screening test sets were identified: the nature and extent of scrutiny of one's action, the artificiality of the environment, the oversimplification of responses, and prevalence of abnormality. The impact of these factors on radiological and other contexts is discussed, and although it is important to acknowledge the benefit of standardized screening test sets, issues relating to the relevance of test sets to

Factors Affecting External Validity of Mammography Test Sets

- The nature and extent of scrutiny of one's action
- The artificiality of the environment
- The oversimplification of responses
- The prevalence of abnormalities

Source: Soh BP, et al. Clin Radiol 2012;67:623-8.

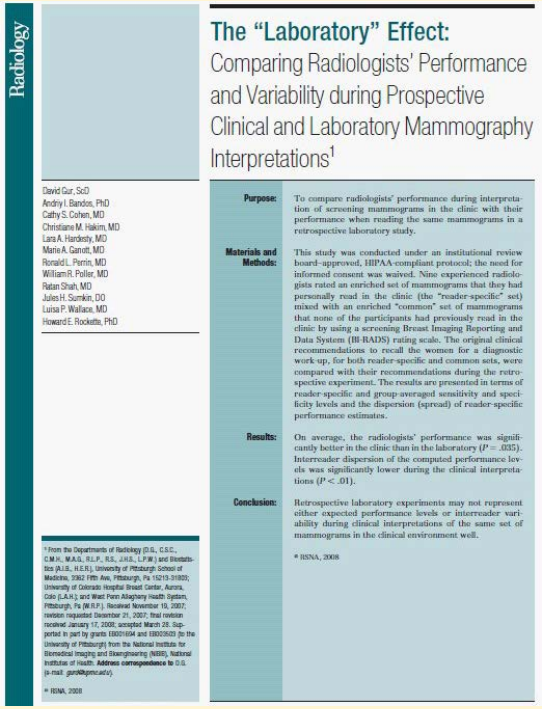
Factors Affecting External Validity of Mammography Test Sets

The artificiality of the environment, & the nature and extent of scrutiny of one's action, i.e.

- *People behave differently when they are being observed*
- *The reading environment is different*
- *The implications of correct and incorrect judgment is different*

Source: Soh BP, et al. Clin Radiol 2012;67:623-8.

The “Laboratory Effect”—Performance in the Clinic vs. the Laboratory



Radiology

The “Laboratory” Effect:
Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations[1]

David Gur, ScD
Andriy I. Bandos, PhD
Cathy S. Cohen, MD
Christiane M. Hakim, MD
Lara A. Hardesty, MD
Marie A. Ganott, MD
Ronald L. Perrin, MD
William R. Poller, MD
Ratan Shah, MD
Jules H. Sumkin, DO
Luisa P. Wallace, MD
Howard E. Rockette, PhD

Purpose: To compare radiologists' performance during interpretation of screening mammograms in the clinic with their performance when reading the same mammograms in a retrospective laboratory study.

Materials and Methods: This study was conducted under an institutional review board–approved, HIPAA-compliant protocol; the need for informed consent was waived. Nine experienced radiologists rated an enriched set of mammograms that they had personally read in the clinic (the "reader-specific" set) mixed with an enriched "common" set of mammograms that none of the participants had previously read in the clinic by using a screening Breast Imaging Reporting and Data System (BI-RADS) rating scale. The original clinical recommendations to recall the women for a diagnostic work-up, for both reader-specific and common sets, were compared with their recommendations during the retrospective experiment. The results are presented in terms of reader-specific and group-averaged sensitivity and specificity levels and the dispersion (spread) of reader-specific performance estimates.

Results: On average, the radiologists' performance was significantly better in the clinic than in the laboratory ($P = .035$). Interreader dispersion of the computed performance levels was significantly lower during the clinical interpretations ($P < .01$).

Conclusion: Retrospective laboratory experiments may not represent either expected performance levels or interreader variability during clinical interpretations of the same set of mammograms in the clinical environment well.

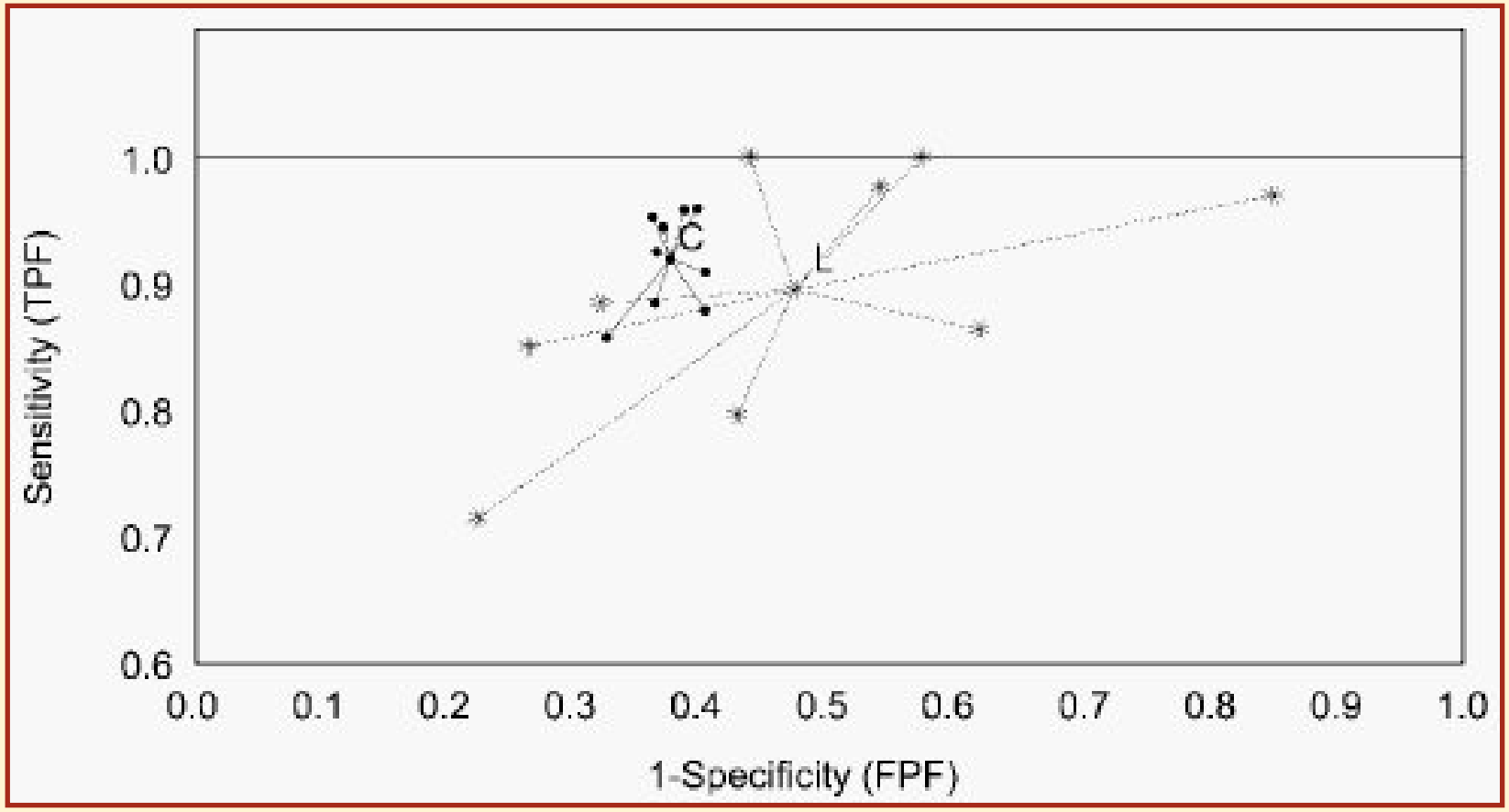
* RSNA, 2008

* From the Departments of Radiology (D.G., C.S.C., C.M.H., M.A.G., R.L.P., R.S., J.H.S., L.P.W.) and Biostatistics (A.I.B., H.E.R.), University of Pittsburgh School of Medicine, 3362 Fifth Ave, Pittsburgh, Pa 15213-3180; University of Colorado Hospital Breast Center, Aurora, Colo (L.A.H.); and West Penn Allegheny Health System, Pittsburgh, Pa (W.R.P.). Received November 19, 2007; revision requested December 21, 2007; final revision received January 17, 2008; accepted March 28. Supported in part by grants EB001694 and EB003503 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. **Address correspondence to** D.G. (e-mail: gurd@upmc.edu).

Radiology: Volume 249: Number 1—October 2008

- Nine experienced radiologists rated an enriched set of mammograms that they had personally read in the clinic (the “reader-specific” set) mixed with an enriched “common” set of mammograms that none of the participants had previously read in the clinic by using a screening BI-RADS rating scale.
- On average, the radiologists’ performance was significantly better in the clinic than in the laboratory. Inter reader dispersion of the computed performance levels was significantly lower during the clinical interpretations.

Performance levels (sensitivity & specificity) and overall average performance (center points and spread) for reader-specific sets of mammograms in the clinic (C) and laboratory (L).



Radiology: Volume 249: Number 1—October 2008

Journal of Medical Imaging and Radiation Oncology •• (2015) ••–••

RADIOLOGY—ORIGINAL ARTICLE

Certain performance values arising from mammographic test set readings correlate well with clinical audit

BaoLin Pauline Soh,[1] Warwick Bruce Lee,[2] Claudia Mello-Thoms,[3] Kriscia Tapia,[3] John Ryan,[4] Wai Tak Hung,[2] Graham Thompson,[2] Rob Heard[3] and Patrick Brennan[3]

1 Department of Diagnostic Radiology, Singapore General Hospital, Singapore
2 Cancer Institute NSW, Sydney, New South Wales, Australia
3 Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences (C42), University of Sydney, Sydney, New South Wales, Australia
4 Ziltron, Dublin, Ireland

- Test set consisted of 20 cancers, and 40 normal exams
- Clinical audit data was generated for 20 radiologists over 2 years
- Significant correlations were observed for:
  - Recall rate at 1st exam
  - Rate of small invasive cancers per 10,000 reads
  - Sensitivity
  - Missed cancers

Soh BPJ Med Imaging Radiat Oncol 2015.

Journal of Medical Imaging and Radiation Oncology •• (2015) ••–••

RADIOLOGY—ORIGINAL ARTICLE

Certain performance values arising from mammographic test set readings correlate well with clinical audit

BaoLin Pauline Soh,[1] Warwick Bruce Lee,[2] Claudia Mello-Thoms,[3] Kriscia Tapia,[3] John Ryan,[4] Wai Tak Hung,[2] Graham Thompson,[2] Rob Heard[3] and Patrick Brennan[3]

1 Department of Diagnostic Radiology, Singapore General Hospital, Singapore
2 Cancer Institute NSW, Sydney, New South Wales, Australia
3 Medical Image Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences (C42), University of Sydney, Sydney, New South Wales, Australia
4 Ziltron, Dublin, Ireland

- The test set did not correlate as well with Specificity, likely due to:
  - Restrictions in the clinical setting that did not carry over to the laboratory setting, i.e.,
    - ≤ 10% recall on initial screening exams
    - ≤ 5% recalls on subsequent exams
  - Radiologists were informed that the test set was enriched with cancers, likely leading to a greater tendency to recall

Soh BPJ Med Imaging Radiat Oncol 2015.

Factors Affecting External Validity of Mammography Test Sets

The oversimplification of responses, *i.e., often a simple answer is the only choice to a complex situation*

- ***Prior studies may not be available***
- ***Other features of the image may prompt questions that can't be answered***

Source: Soh BP, et al. Clin Radiol 2012;67:623-8.

Factors Affecting External Validity of Mammography Test Sets

The prevalence of abnormalities, *i.e., it always will exceed the normal prevalence of disease*

- ***A higher prevalence of cancers, leads to a heightened level of suspicion***
- ***Overall and individual effects are unclear***

Source: Soh BP, et al. Clin Radiol 2012;67:623-8.

Simulation usually aims to increase the prevalence of rare events—consider the following evaluation of a crew's ability to handle in flight emergencies

1. Tire explodes in the wheel well after takeoff, damages control surfaces on right wing
2. Right fuel gauge indicates fuel loss, likely due to damaged fuel lines from explosion
3. During final approach, right engine catches on fire, requiring engine shut down
4. ***Not a normal day in the cockpit !***

Challenges in the Development of Test Sets — *these still are unresolved*

- Number of exams?
- Mix of difficulty?
- Ratio of cancers to normal exams?
- Measures of truth—biopsy-confirmed, or expert consensus?
  - Cases that should be recalled vs….
  - those that are indeterminate vs….
  - those that should not be recalled

Challenges in the Development of Test Sets (2)

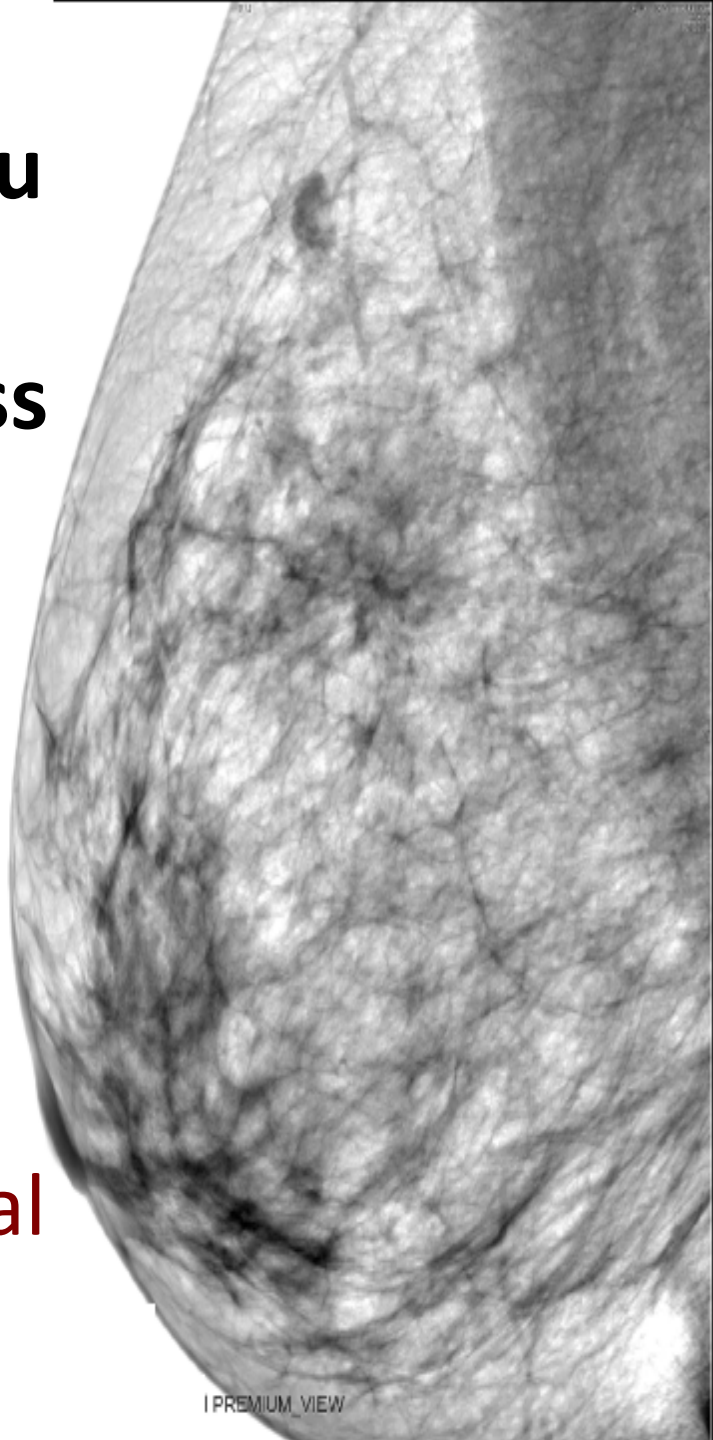
- Who should be tested, and how often?
- Should test sets be manufacturer-specific?
- How often should images be refreshed?
- How should performance be evaluated to improve test set composition?
- Are differences between clinic performance and laboratory performance a function of single occasion testing?

New Directions in Evaluating Performance

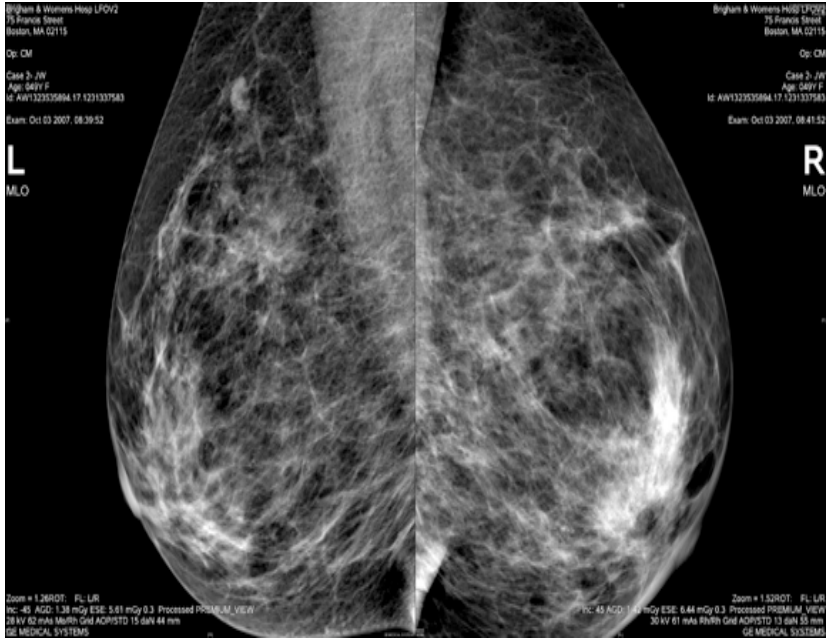
The influence of the low prevalence of cancers in the screening cohort has been proposed as a contributing factor in missed cancer error rates

If You Don't Find It Often, You Often Don't Find It: Disease Prevalence Is a Source of Miss Errors in Screening Mammography

Jeremy M Wolfe, PhD
Robyn L Birdwell, MD
Karla K Evans, PhD
Brigham and Women's Hospital
& Harvard Medical School



Basic 2-Arm Design



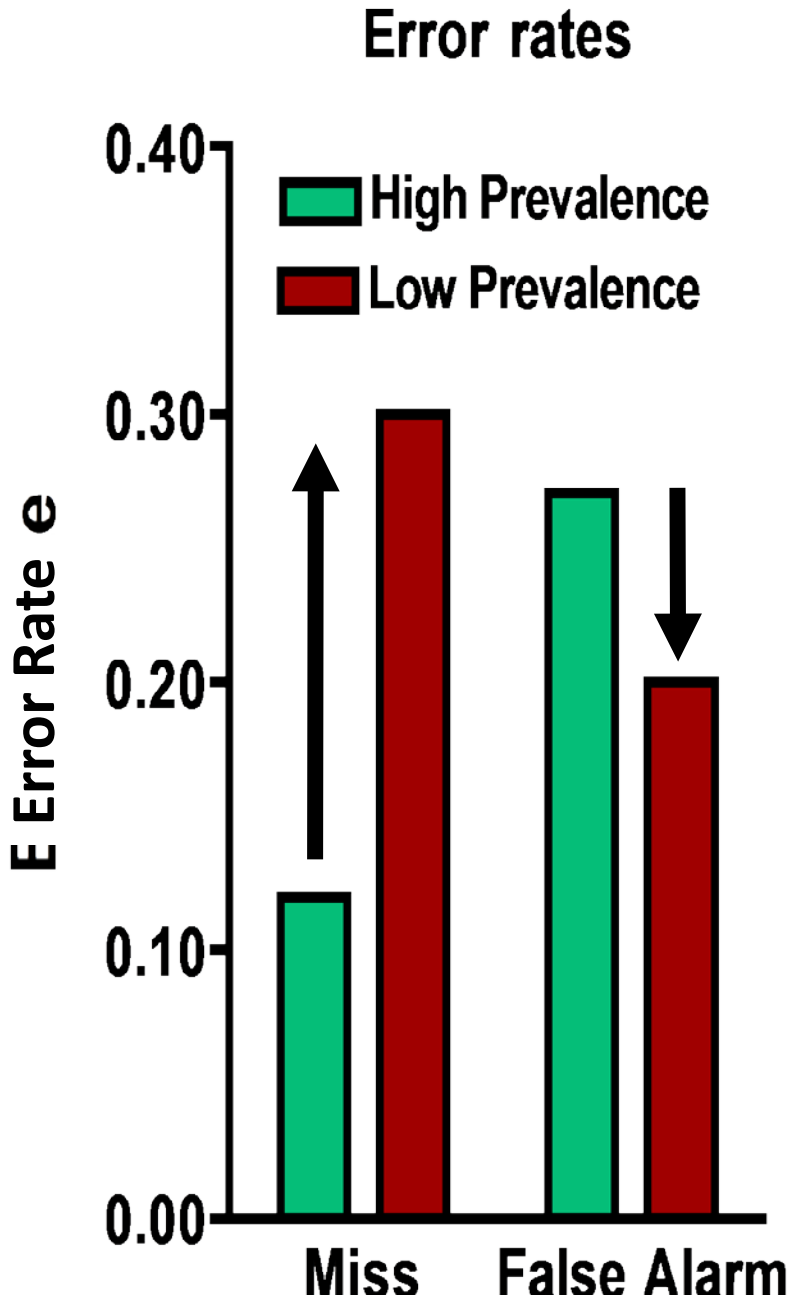
Low Prevalence Arm

- 100 cases (50 positive, 50 negative)
- These are inserted into the normal workflow of the Women's Imaging practice at Brigham and Women's Hospital
- The 100 cases were viewed over the course of 9 months during which another 9826 other cases were screened.
- Estimated prevalence = 0.8%. Data are the call back decisions.
- 14 radiologists, reading unequal numbers of these cases

Basic 2-Arm Design

High Prevalence Arm

- 100 cases (50 positive, 50 negative): Prevalence = 50%
- All 100 were read by each of 6 of the 14 radiologists from the low prevalence arm.
- Reading the 100 cases took about 3 hours.
- Data are the call back decisions and a 0-10 rating from negative to clearly abnormal.



The Key Result

Miss error rates are substantially higher at low prevalence

False alarm rates are somewhat lower at low prevalence

Modern Proficiency Testing—New Directions?

OPEN ACCESS Freely available online

PLOS ONE

If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening

Karla K. Evans[1]*, Robyn L. Birdwell[2], Jeremy M. Wolfe[1]

1 Visual Attention Lab, Brigham and Women's Hospital, Harvard Medical School, Cambridge, Massachusetts, United States of America, 2 Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Mammography is an important tool in the early detection of breast cancer. However, the perceptual task is difficult and a significant proportion of cancers are missed. Visual search experiments show that miss (false negative) errors are elevated when targets are rare (low prevalence) but it is unknown if low prevalence is a significant factor under real world, clinical conditions. Here we show that expert mammographers in a real, low-prevalence, clinical setting, miss a much higher percentage of cancers than are missed when the mammographers search for the same cancers under high prevalence conditions. We inserted 50 positive and 50 negative cases into the normal workflow of the breast cancer screening service of an urban hospital over the course of nine months. This rate was slow enough not to markedly raise disease prevalence in the radiologists' daily practice. Six radiologists subsequently reviewed all 100 cases in a session where the prevalence of disease was 50%. In the clinical setting, participants missed 30% of the cancers. In the high prevalence setting, participants missed just 12% of the same cancers. Under most circumstances, this low prevalence effect is probably adaptive. It is usually wise to be conservative about reporting events with very low base rates (Was that a flying saucer? Probably not.). However, while this response to low prevalence appears to be strongly engrained in human visual search mechanisms, it may not be as adaptive in socially important, low prevalence tasks like medical screening. While the results of any one study must be interpreted cautiously, these data are consistent with the conclusion that this behavioral response to low prevalence could be a substantial contributor to miss errors in breast cancer screening.

Citation: Evans KK, Birdwell RL, Wolfe JM (2013) If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. PLoS ONE 8(5): e64366. doi:10.1371/journal.pone.0064366

Editor: Michael J. Proulx, University of Bath, United Kingdom

Received March 20, 2013; Accepted April 12, 2013; Published May 30, 2013

Funding: This work was supported by NIH EY017001 and ONR MURI N000141010278 to J.M.W. K.K.E. was supported by NIH/NEI 1F32EY019814-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

*E-mail: kevans@search.bwh.harvard.edu

Introduction

Mammographic screening is an important tool in the early detection of breast cancer [1] but it is a difficult perceptual task and error-prone [2] with reported false negative rates of 20–30% [3,4]. The signs of breast cancer are often ambiguous and/or hard to see, with some proportion of errors attributable to the perceptual difficulty of the task. However, a significant proportion of miss errors cannot be attributed to a lack of a clear signal. In many cases, if disease is detected in the current exam, it can also be seen in retrospect on the previous exam. These "retrospectively visible" or "actionable" cancers could have been found but were missed on that previous exam [5–7]. They are either errors in perception (failures of search) [8], or alternatively errors in interpretation. Here we consider one contributor to those failures, namely the low prevalence of disease in screening mammograms.

Breast cancer screening by mammography is a difficult visual search task, characterized by a low prevalence of positive findings. Experiments with non-experts in a laboratory setting show that more targets are missed during vigilance tasks when observers monitor displays for targets that appear infrequently [9,10]. Attention fluctuates and targets can come and go without being noticed. More recently, it has been shown that these prevalence effects occur in visual search tasks even though observers can view displays for as long as they want. Even when observers must actively reject a display before it will be removed, more targets are missed at low prevalence than at higher prevalence [11].

The opposite type of error, false positives, tend to decline at low prevalence [12] because the primary effect of prevalence is a criterion shift, with observers in low prevalence situations less likely to call an ambiguous stimulus a target and more likely to terminate search [13]. In clinical settings, neither false positives nor false negative errors are desirable but it seems reasonable to assert that false negative errors are less desirable. Thus, if low prevalence produces more false negative errors in a clinical setting, even if the false positive errors decline, that would be important information.

It is important to note that prevalence effect could have two different types of effect on performance in mammography [14–16]. Gur and his colleagues have shown that, in a laboratory setting, prevalence did not change the area under the receiver operating characteristic curve (AUC) [14]. However, even if AUC is unchanged, it would be of interest to find the change in the pattern of errors that would follow a change in criterion, the bias to call a case actionable or non-actionable. In reanalyzing the 2003 data, Gur et al. (2008) reported a change in confidence ratings with prevalence that would be consistent with a criterion shift and, as noted above, criterion shifts have been a hallmark of prevalence effects outside radiology [15]. Our particular interest was in looking for evidence for a prevalence effect in the clinic with professionals carrying out a critical task in their area of expertise.

What can be done? General methods to improve performance include:

- Training
- Experience
- Continuing education,
- Prospective double reading,
- Retrospective evaluation of missed cases

"Given the present results, we should now determine if there are practical changes in clinical settings that can reduce errors due to target prevalence."

Thank you