LRCB dutch reference centre for screening

Radboudumc

Overview of international test sets

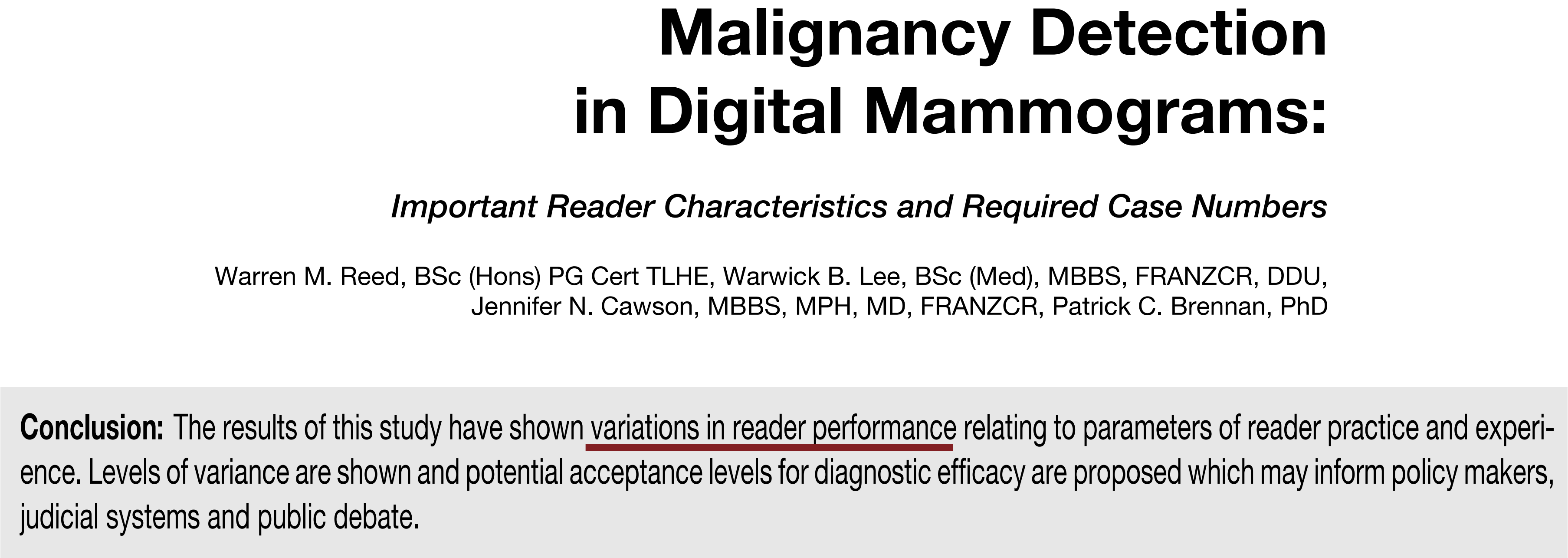
Mireille Broeders, PhD

Dutch Reference Centre for Screening and Dept for Health Evidence, Radboudumc, Nijmegen, NL

IOM Workshop, Washington, May 12, 2015

A self test

… to measure performance



Malignancy Detection in Digital Mammograms:

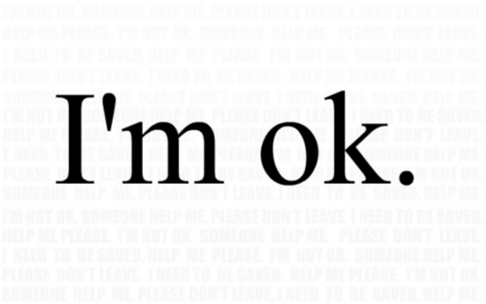
Important Reader Characteristics and Required Case Numbers

Warren M. Reed, BSc (Hons) PG Cert TLHE, Warwick B. Lee, BSc (Med), MBBS, FRANZCR, DDU,
Jennifer N. Cawson, MBBS, MPH, MD, FRANZCR, Patrick C. Brennan, PhD

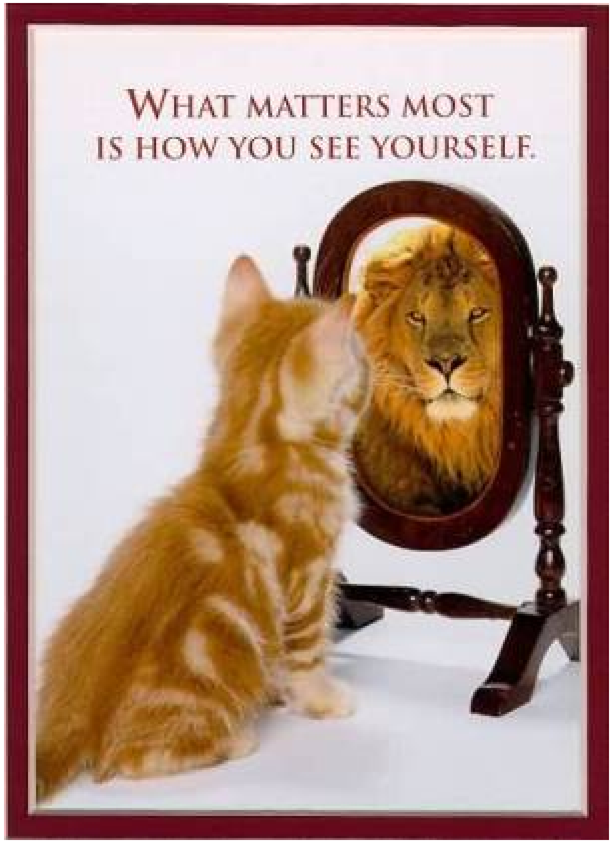
Conclusion: The results of this study have shown variations in reader performance relating to parameters of reader practice and experience. Levels of variance are shown and potential acceptance levels for diagnostic efficacy are proposed which may inform policy makers, judicial systems and public debate.

Acad Radiol 2010; 17:1409-1413

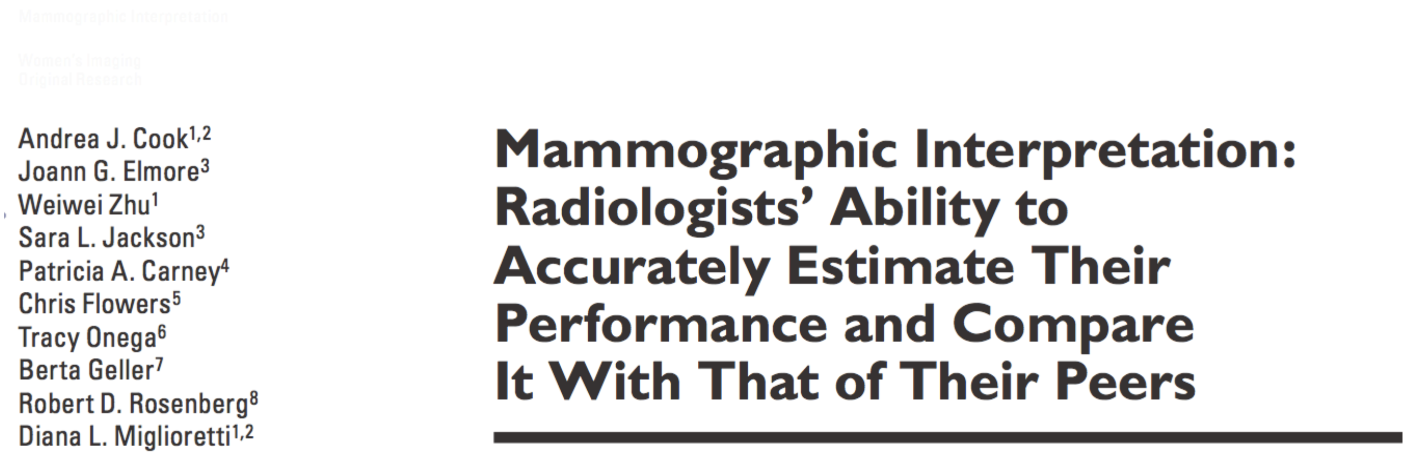
… to provide feedback


I'm ok.
I'M NOT OK. SOMEONE HELP ME. PLEASE DON'T LEAVE. I NEED TO BE SAVED.

I'm the BEST !



… to provide feedback

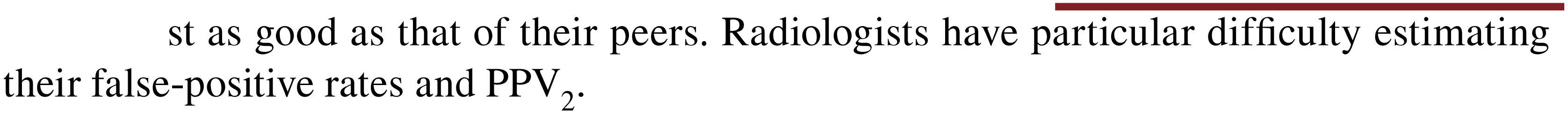


Mammographic Interpretation

Women's Imaging
Original Research

Andrea J. Cook[1,2]
Joann G. Elmore[3]
Weiwei Zhu[1]
Sara L. Jackson[3]
Patricia A. Carney[4]
Chris Flowers[5]
Tracy Onega[6]
Berta Geller[7]
Robert D. Rosenberg[8]
Diana L. Miglioretti[1,2]

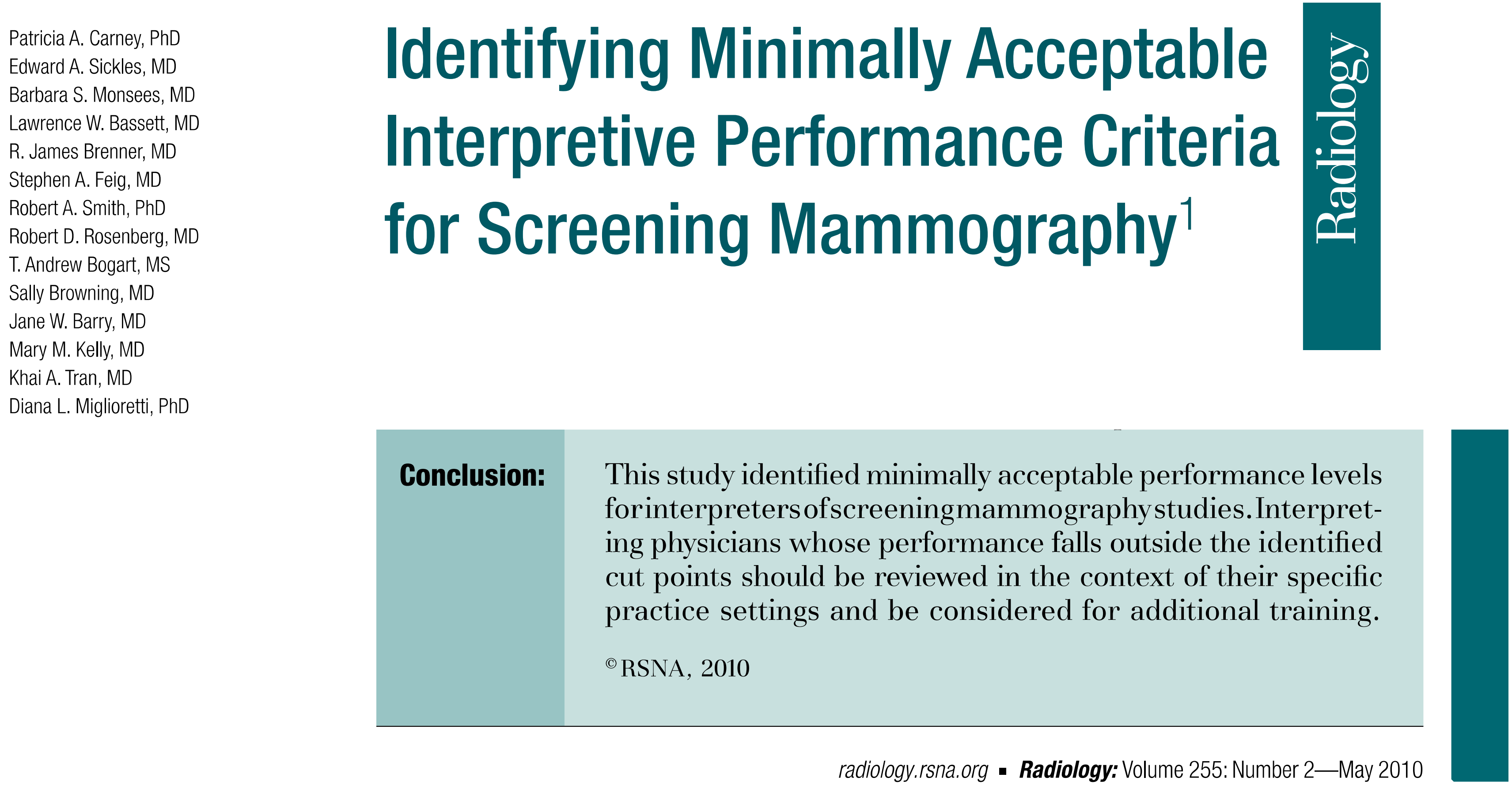
Mammographic Interpretation: Radiologists' Ability to Accurately Estimate Their Performance and Compare It With That of Their Peers



st as good as that of their peers. Radiologists have particular difficulty estimating their false-positive rates and PPV_2.

AJR 2012;199:695-702

… to set reference standards



Patricia A. Carney, PhD
Edward A. Sickles, MD
Barbara S. Monsees, MD
Lawrence W. Bassett, MD
R. James Brenner, MD
Stephen A. Feig, MD
Robert A. Smith, PhD
Robert D. Rosenberg, MD
T. Andrew Bogart, MS
Sally Browning, MD
Jane W. Barry, MD
Mary M. Kelly, MD
Khai A. Tran, MD
Diana L. Miglioretti, PhD

Identifying Minimally Acceptable Interpretive Performance Criteria for Screening Mammography[1]

Radiology

Conclusion: This study identified minimally acceptable performance levels for interpreters of screening mammography studies. Interpreting physicians whose performance falls outside the identified cut points should be reviewed in the context of their specific practice settings and be considered for additional training.

radiology.rsna.org ▪ ***Radiology:*** Volume 255: Number 2—May 2010

A self test

- ✓ Simple
- ✓ Instant feedback
- ✓ Identify individual training needs
- ✓ Benchmark

- ? Representative
- ? How many cancers to put in
- ? Laboratory environment
- ? Relation real life and testing

Current test sets in use

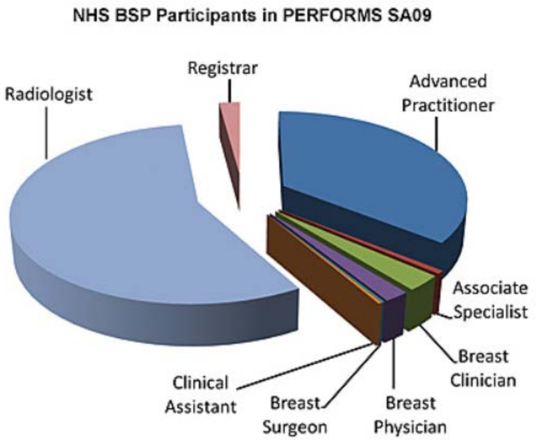
As part of quality assurance in national screening programs:

- UK: PERFORMS
- Australia/New Zealand: BREAST

Other than these two, standardized test sets for self-assessment do not (yet) seem to be integrated in quality assurance

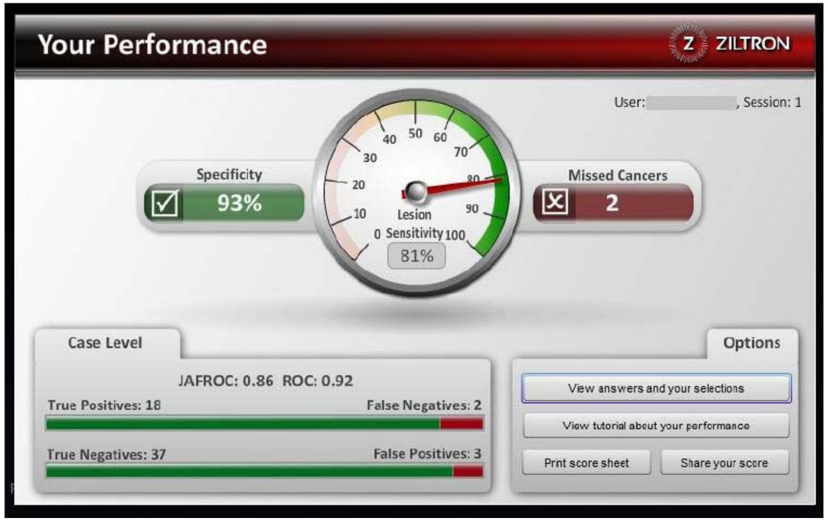
PERFORMS

- PERFORMS = Personal Performance in Mammography Screening
- Implemented in the UK National Programme in 1991
- Educational self assessment and training scheme for breast screening professionals
- Challenging set of 60 cases, no prior images, submitted by screening centres
- Feedback directly after taking the test



BREAST

- BREAST = Breastscreen REader ASsessment Strategy
- Developed in 2011, now rolled out to all to BreastScreen programs in all states in Australia and to BreastScreen Aotearoa (i.e. New Zealand)
- 60 cases, no prior images, uses web-based software
- Immediate feedback on performance and annotated images


Your Performance

User: [redacted], Session: 1

Specificity ☑ 93%

Lesion Sensitivity 81%

Missed Cancers ☒ 2

Case Level

JAFROC: 0.86 ROC: 0.92

True Positives: 18 False Negatives: 2

True Negatives: 37 False Positives: 3

Options

Z ZILTRON
0
10
20
30
40
50
60
70
80
90
100
View answers and your selections
View tutorial about your performance
Prnt score sheet
Share your score

Dutch self-test

Experiences with a self-test for Dutch breast screening radiologists: lessons learnt

J. M. H. Timmers · A. L. M. Verbeek · R. M. Pijnappel ·
M. J. M. Broeders · G. J. den Heeten

Eur Radiol (2014) 24:294–304

- Feedback for Dutch screening radiologists is organized locally and every 3 years through audits
- Self-test was developed to offer individual evaluation to screening radiologists and identify areas for training
- The purpose of this study was to evaluate the use of the test set and improve the method of testing

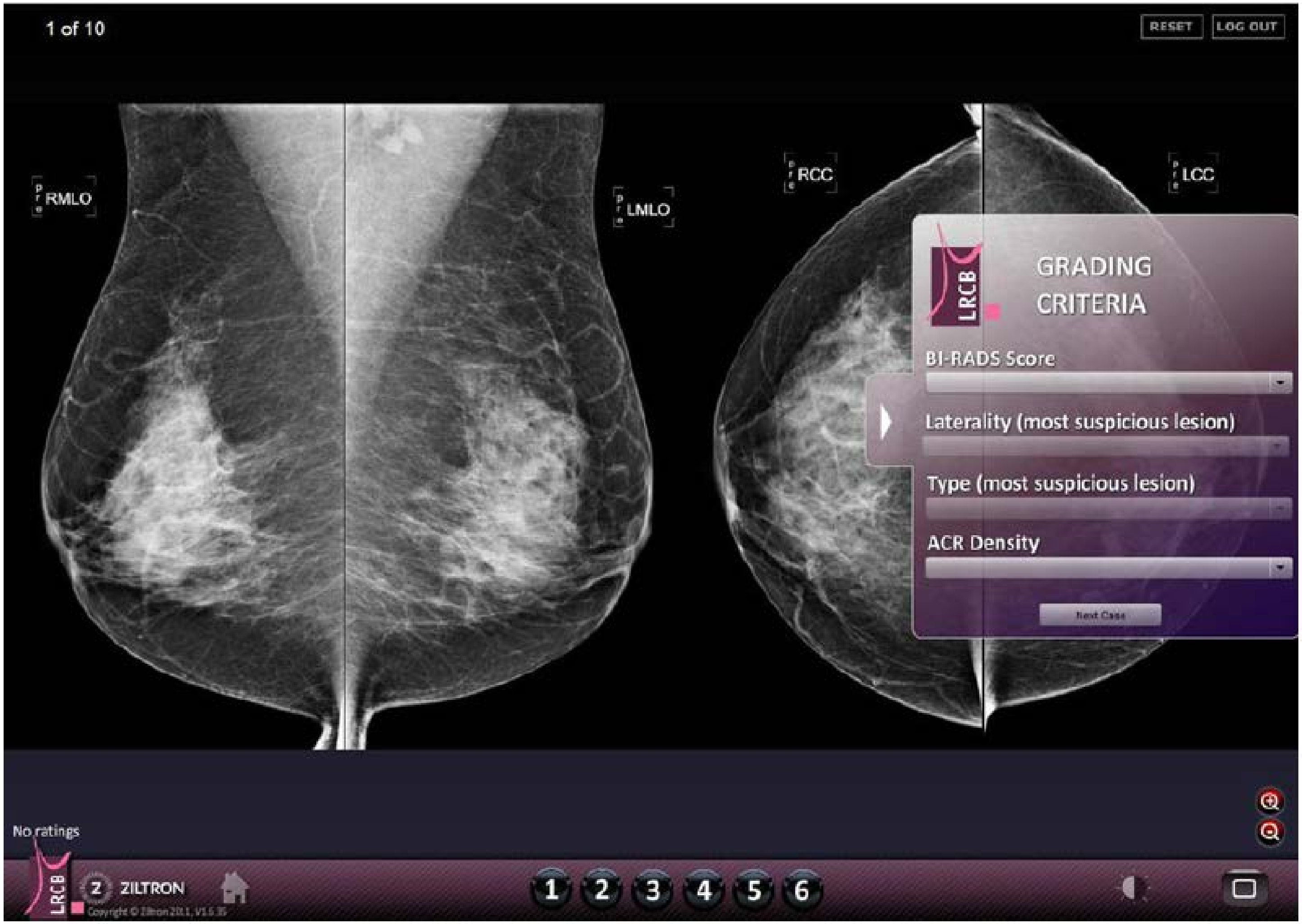
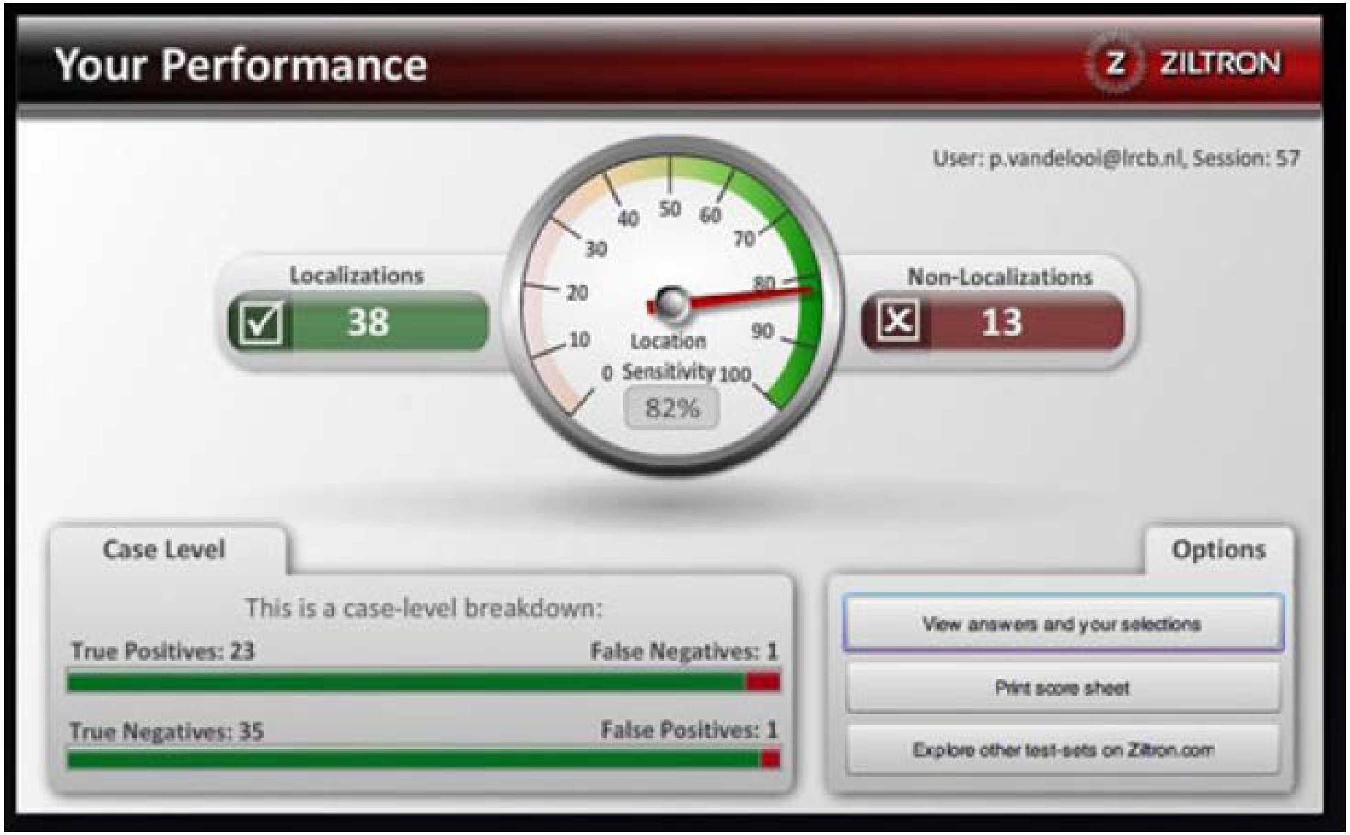


Fig. 1 Screen shot of self-test with grading criteria and confidence scale of suspicion for malignancy



Your Performance

User: p.vandelooi@lrcb.nl, Session: 57

Localizations ☑ 38

Non-Localizations ☒ 13

Location Sensitivity 82%

Case Level

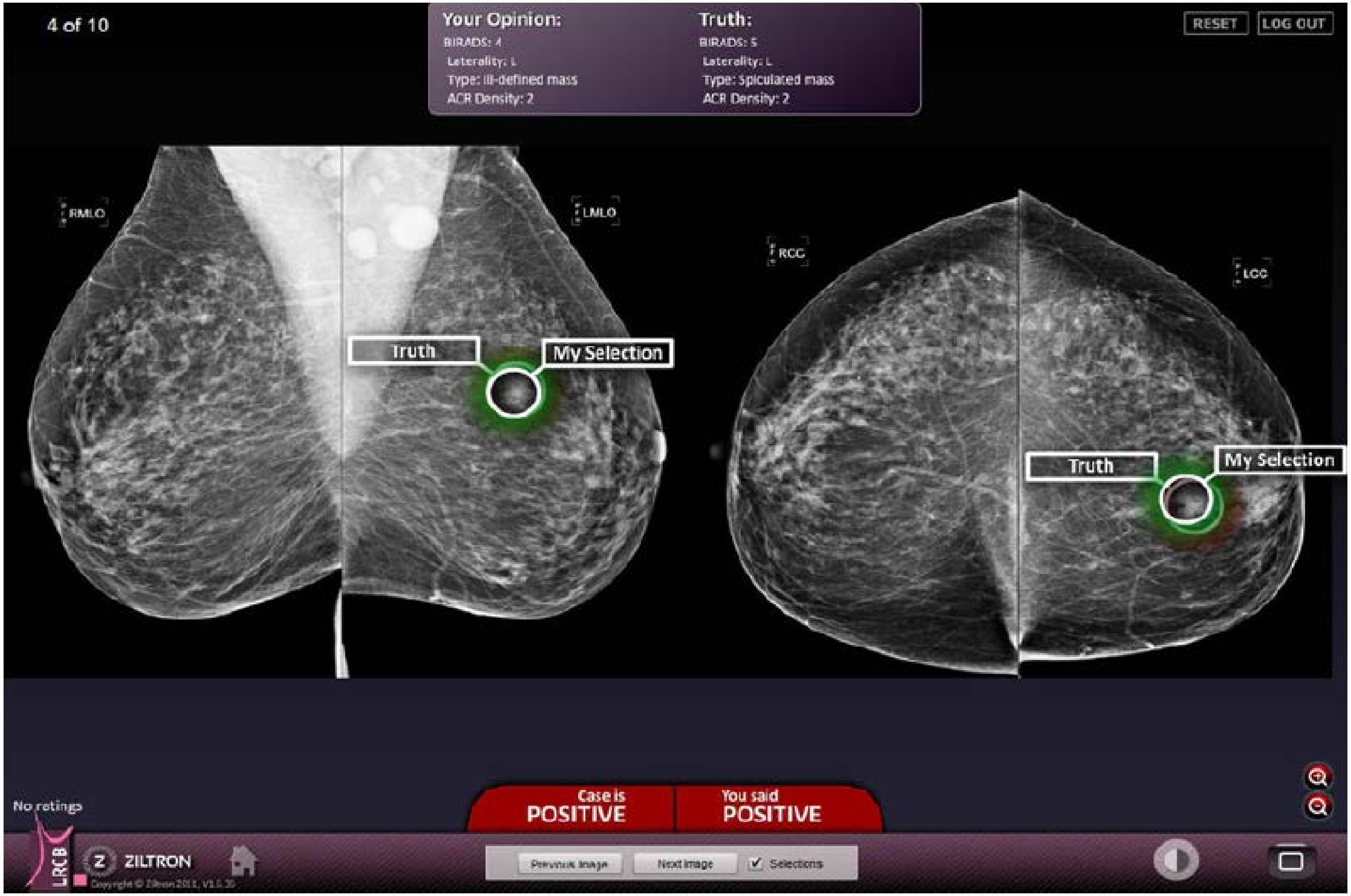
This is a case-level breakdown:

True Positives: 23 False Negatives: 1

True Negatives: 35 False Positives: 1

Z ZILTRON
0
10
20
30
40
50
60
70
80
90
100
Options
View answers and your selections
Print score sheet
Explore other test-sets on Ziltron.com

Fig. 2 Feedback module on overall individual performance showing performance parameters such as correctly assigned lesions or number of true positive lesions



Your Opinion:
BIRADS: 4
Laterality: L
Type: Ill-defined mass
ACR Density: 2

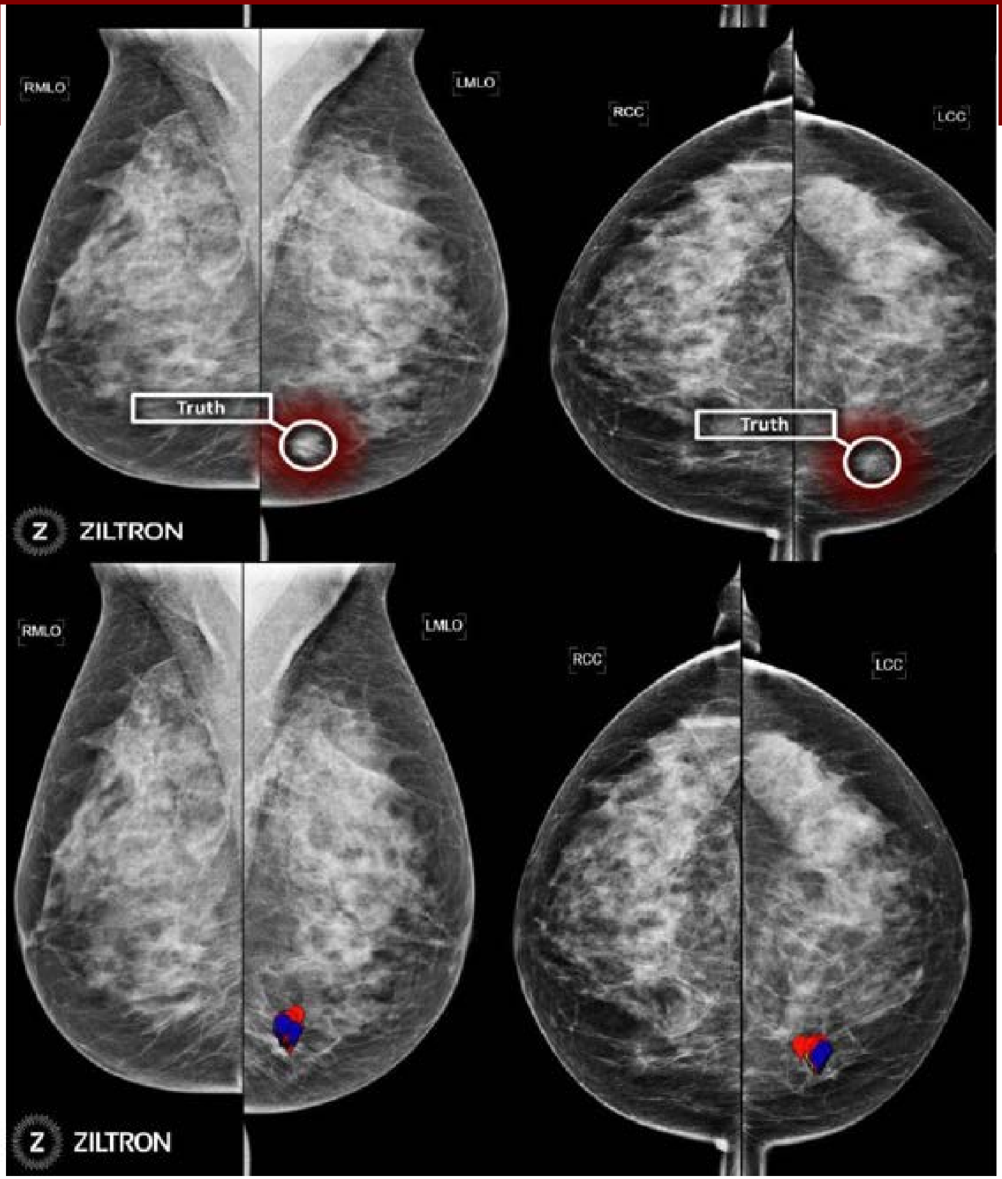
Truth:
BIRADS: 5
Laterality: L
Type: Spiculated mass
ACR Density: 2

Fig. 3 Feedback at case level showing the selection of the participant and the selection of the expert panel

Main results

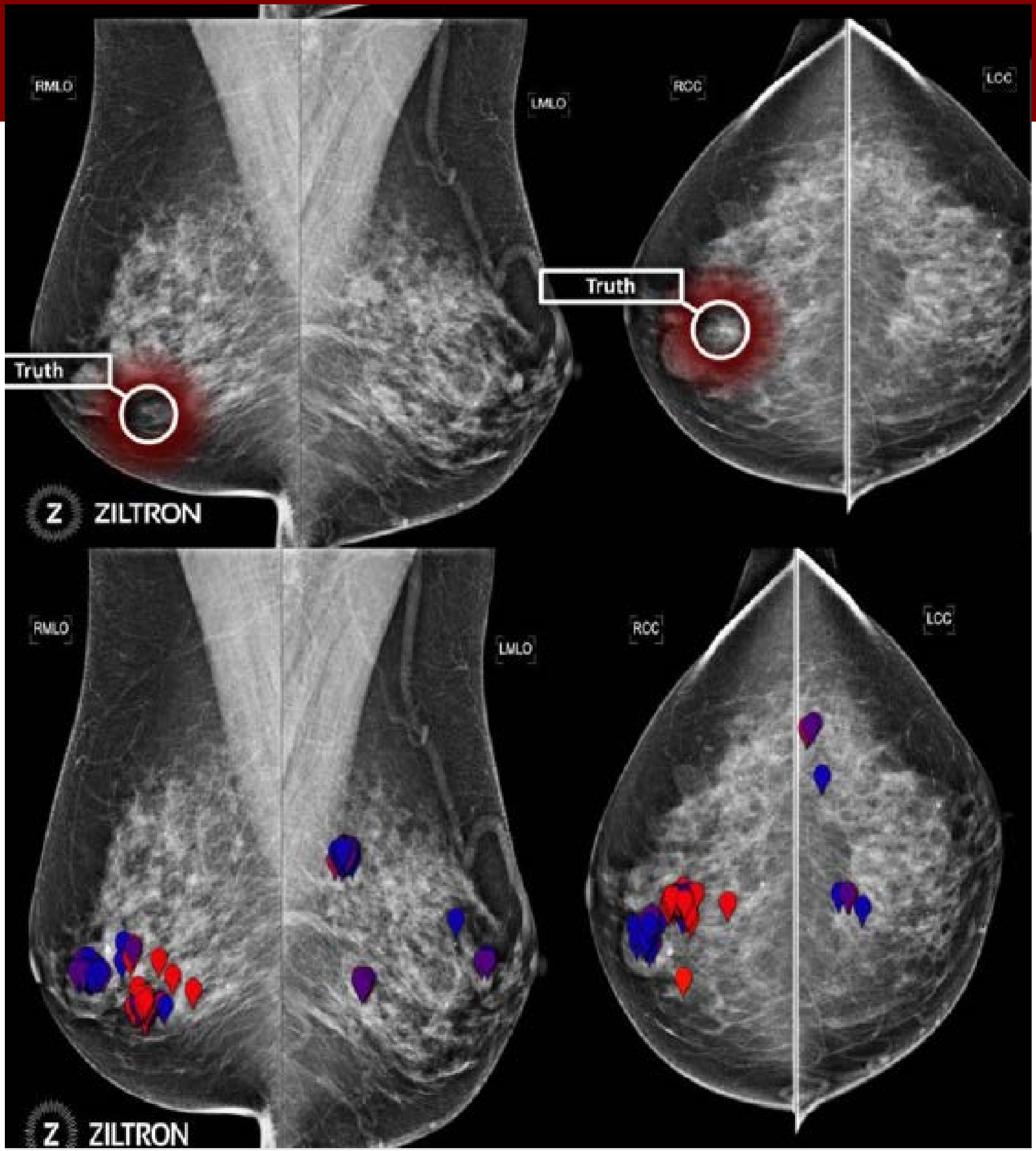
- ROC curves, sensitivity, recall agreement satisfactory

- Agreement in BI-RADS interpretation and lesion type could be improved



Main results

- ROC curves, sensitivity, recall agreement satisfactory
- Agreement in BI-RADS interpretation and lesion type could be improved



Lessons learnt

- Well accepted by participants (78%): representative of invited radiologists?

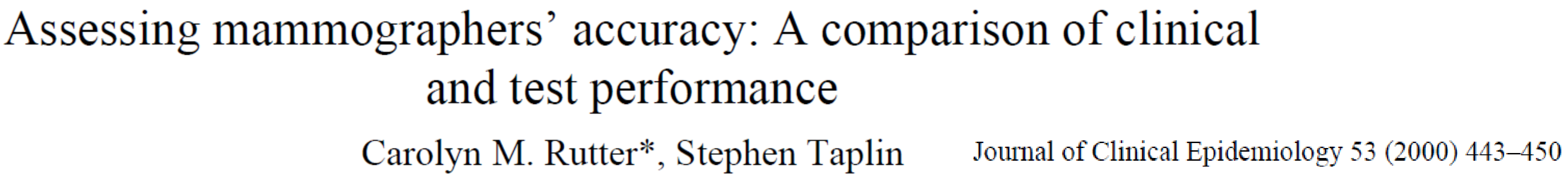
- Consider using fixed test dates and locations:
  - to prevent user problems
  - to prevent possible sharing of results

- Confidence scale is important for ROC analysis: we found that the use of the confidence scale was not clear to at least 19 readers (14 %)

LRCB dutch reference centre for screening

Radboudumc

Test set vs. real life

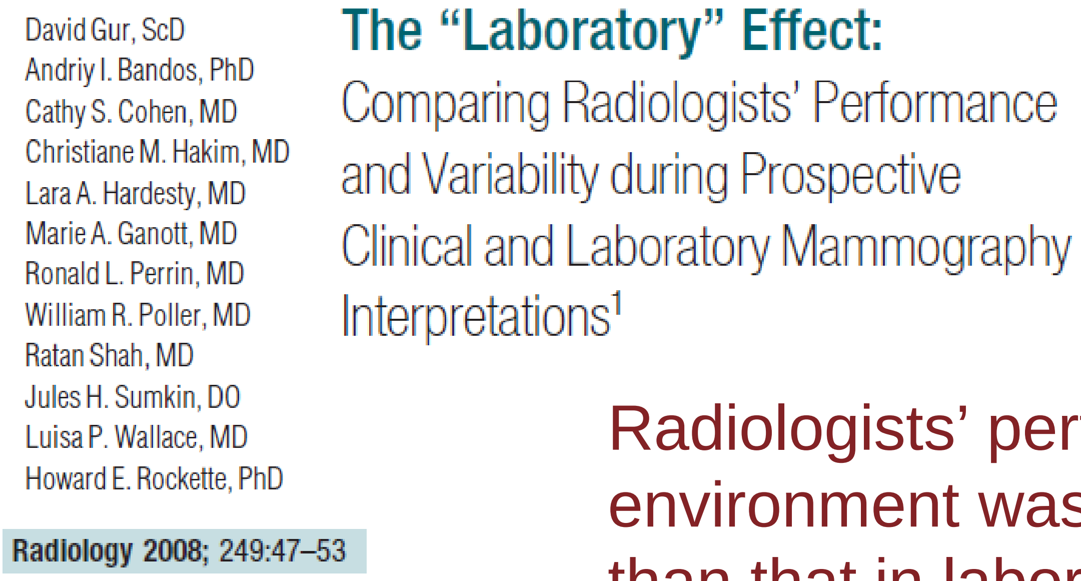


Assessing mammographers' accuracy: A comparison of clinical and test performance

Carolyn M. Rutter*, Stephen Taplin

Journal of Clinical Epidemiology 53 (2000) 443–450

No evidence of correlation between clinical and test accuracy



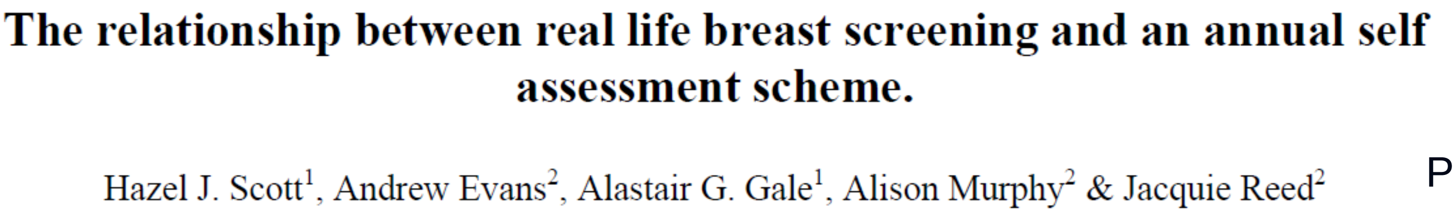
David Gur, ScD
Andriy I. Bandos, PhD
Cathy S. Cohen, MD
Christiane M. Hakim, MD
Lara A. Hardesty, MD
Marie A. Ganott, MD
Ronald L. Perrin, MD
William R. Poller, MD
Ratan Shah, MD
Jules H. Sumkin, DO
Luisa P. Wallace, MD
Howard E. Rockette, PhD

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

Radiology 2008; 249:47–53

Radiologists' performance in the clinical environment was significantly different than that in laboratory retrospective studies

Test set vs. real life



The relationship between real life breast screening and an annual self assessment scheme.

Hazel J. Scott[1], Andrew Evans[2], Alastair G. Gale[1], Alison Murphy[2] & Jacquie Reed[2]

Proc SPIE 2009:7263

Performance at test set broadly reflected clinical performance

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]

J Med Imaging Radiat Oncol 2015

…. although caution needs to be exercised when generalising test set specificity to the clinical situation

Summary so far

- Test sets are provided by some screening programs to measure readers’ performance:
  - usually as part of quality assurance procedures
  - usually as a complement to clinical audits
  - usually not obligatory
- There is still uncertainty on how performance in test sets is correlated with performance in the real life screening setting

Given that ...

- in many settings, screening radiologists have little opportunity for feedback
- there is still a void of information on the value of test sets
- factors that influence screening interpretation outcomes need to be better understood

⟶ ***international test set project***

LRCB
dutch reference centre
for screening

Radboudumc

ICSN
International Cancer
Screening Network

American Cancer Society®

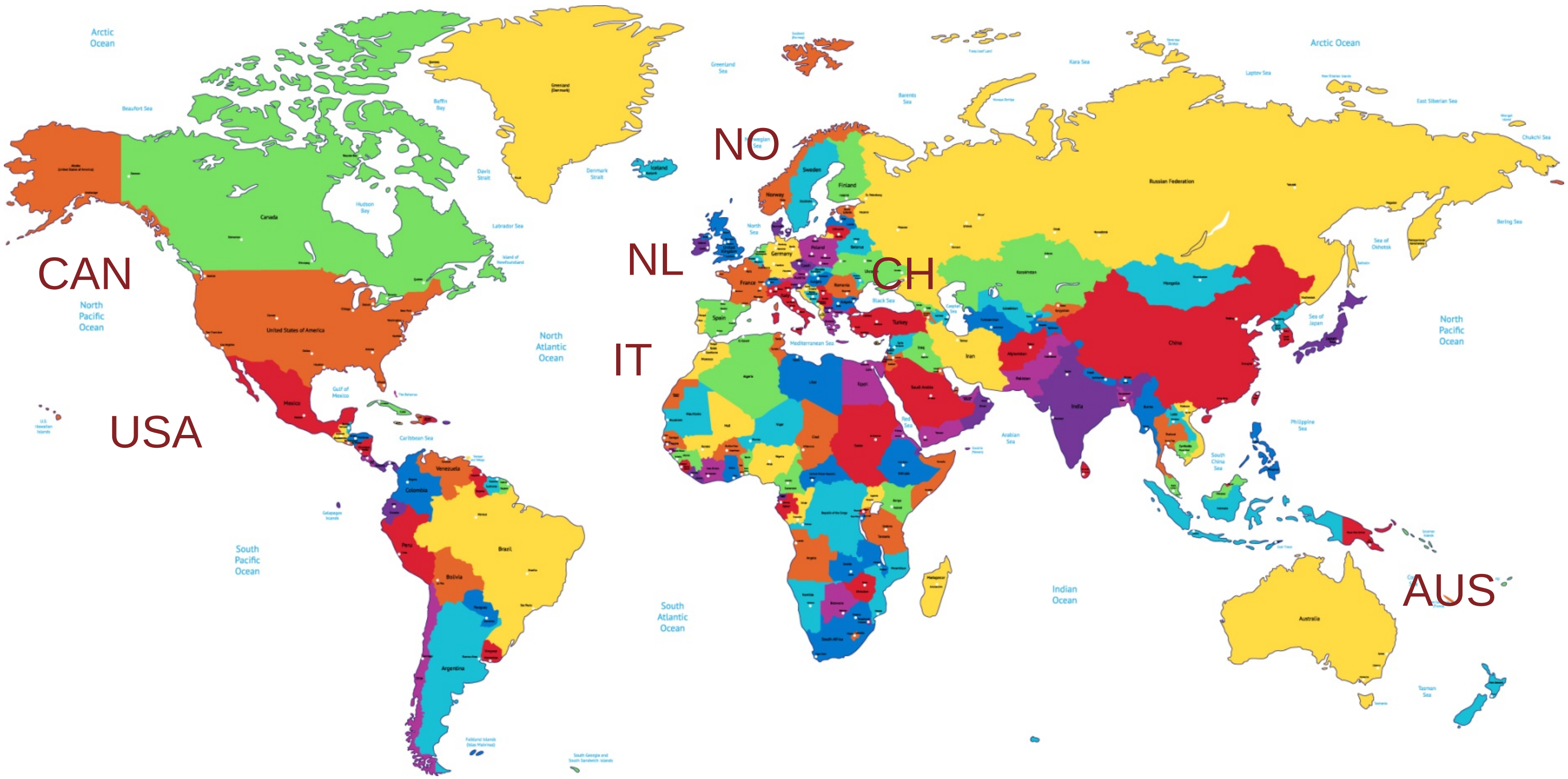
Assessment of International Mammography Screening Skills

AIMSS

Objectives AIMSS

- Develop a test set that will measure proficiency of radiologists
- Develop a system that gives participating radiologists immediate feedback
- Allow for international comparison of radiologists active in screening mammography in different settings

Working group AIMSS



- B. Yankaskas
- M. Broeders
- R. Smith
- D. Miglioretti

- J.L. Bulliard
- A. Frigerio
- R. Pijnappel
- B. Monsees

- S. Hofvind
- A. Chiarelli

Progress AIMSS

- **ICSN 2010**: presentations from five countries on existing test sets

- **ICSN 2012**: presentation on development of the test set based on specific criteria

- **ICSN 2015**: presentation on ongoing pilot with ACS funding – purpose of pilot is to fine tune the procedures for introduction and recruitment

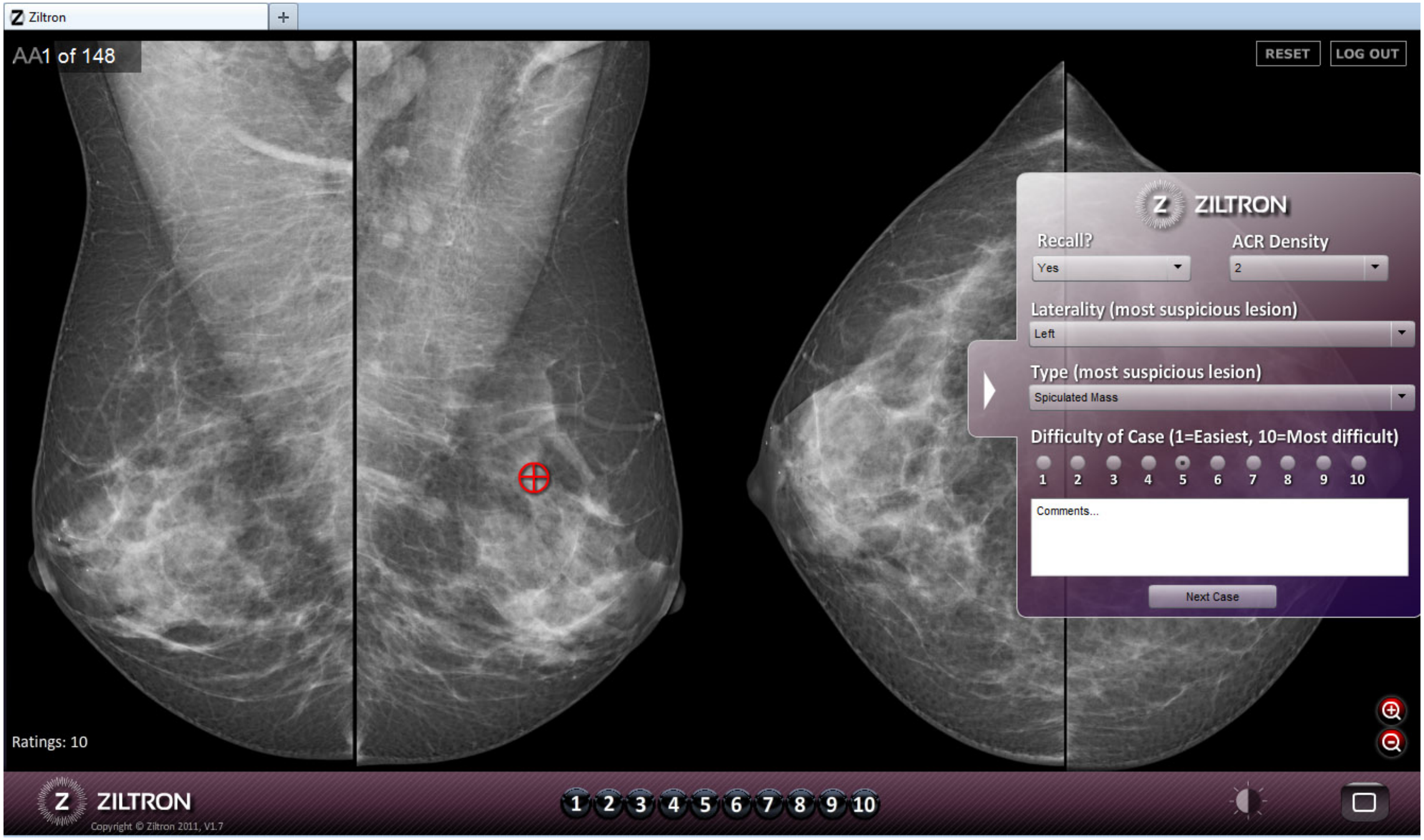
Test set criteria

- Cases of women 40-79 years of age
- Subsequent screening mammograms will have priors, preferably 2 years before
- Only a few obvious cancers, and a few extremely difficult cases will be included
- Exclude interval cancers
- Normal cases should include difficult normal, and have no cancer diagnosed within 2 years

Developing the test set

- 148 cases provided for review (IT, NL, AUS, USA) to select 80 cases for test set
- 10 radiology experts downloaded cases to their own workstations
- Experts used Ziltron software to view the cases on the web and to collect information
- Ziltron provided results of the experts, aggregated and by case
- Results reviewed at consensus meeting with 5 radiology experts

Ziltron interface



Recall? Yes

ACR Density 2

Laterality (most suspicious lesion) Left

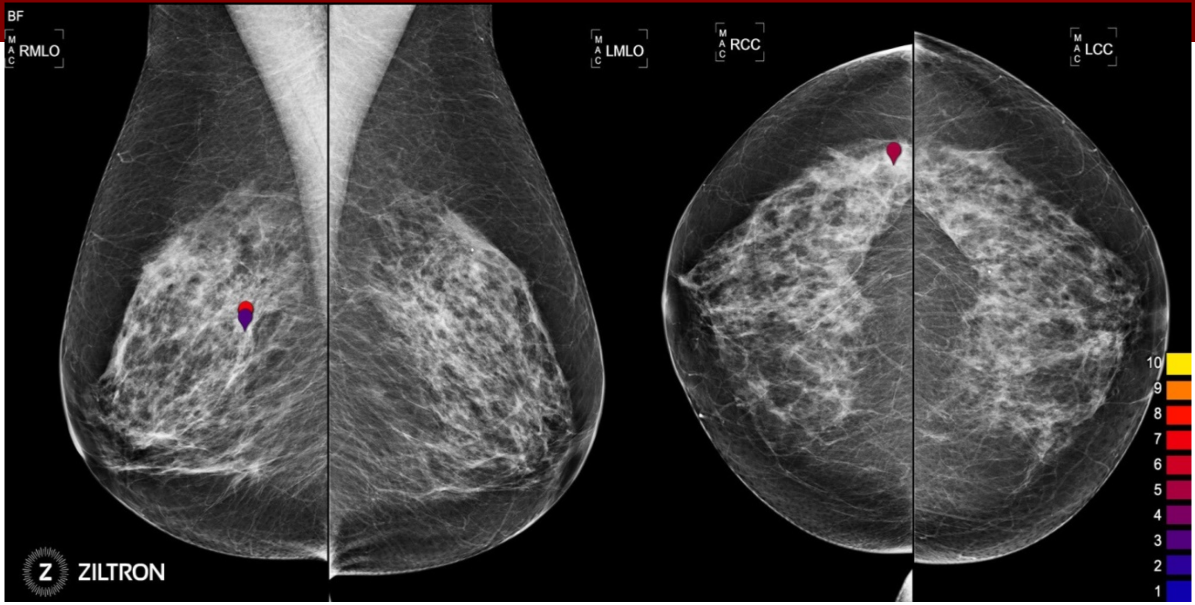
Type (most suspicious lesion) Spiculated Mass

Difficulty of Case (1=Easiest, 10=Most difficult)

○ 1 ○ 2 ○ 3 ○ 4 ● 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Comments...

Consensus meeting



User_ID	Cancer	Recall	TP	FP	TN	FN	Laterality	Type	Density	Difficulty
1	Yes	1	1	0	0	0	1	6	2	4
2	Yes	2	0	0	0	1	0	0	3	1
3	Yes	2	0	0	0	1	0	0	3	7
4	Yes	1	1	0	0	0	1	5	3	5
5	Yes	1	1	0	0	0	1	6	3	7

Recall
1=yes
2=no

Current status AIMSS

- Teaching points for test set under development
- Customize Ziltron simulator to AIMSS
- First invitation for participation in pilot

Practical challenges

- Development of the test set – number of cases, mixture of cases, difficulty of the cases
- International context – appropriate or inappropriate recall – teaching point
- Use of confidence scale will help evaluate recall decisions, but has to be well-explained
- Consider how vendors and display options may influence interpretive performance

Points to consider

- Further develop the methodology for designing test-sets and determine their optimal use
  - Include prior mammograms to resemble screening?
  - Different test sets for detecting cancers and reducing recalls?
  - Different test sets to improve e.g. agreement on lesion type?
  - Different test sets for radiologists and technologists?

- Continue efforts to understand the relevance of test sets to clinical performance
  - Linkage to relevant (audit) data
  - Consider repercussions of test set results

Points to consider

- Include international perspective to compare across countries / screening settings
  - Acknowledge differences in screening practice
  - Identify (new) factors that influence performance
  - Study consistency of correlations worldwide
  - Vendor-specific test sets?

- Consider alternatives, such as mixing cases in into readers' screening practice (Soh et al, Clin Radiol 2012)
  - May overcome the artificial / laboratory setting
  - But will also present new challenges!

Thank you for your attention!