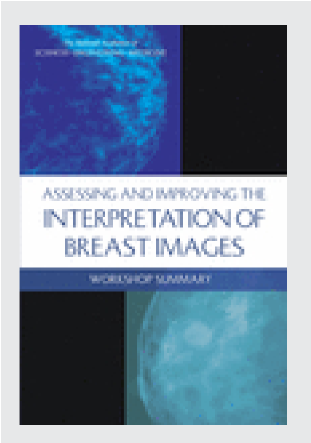

Assessing and Improving the Interpretation of Breast Images: Workshop Summary

DETAILS

94 pages | 6 x 9 | PAPERBACK | ISBN 978-0-309-37835-2

AUTHORS

Sharyl J. Nass and Margie Patlak, Rapporteurs; National Cancer Policy Forum; Board on Health Care Services; Institute of Medicine; The National Academies of Sciences, Engineering, and Medicine

ASSESSING AND IMPROVING THE INTERPRETATION OF BREAST IMAGES

WORKSHOP SUMMARY

Sharyl J. Nass and Margie Patlak, *Rapporteurs*

National Cancer Policy Forum

Board on Health Care Services

Institute of Medicine

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This project was supported by the American Cancer Society through a generous donation from the Longaberger Company's Horizon of Hope Campaign®. The views presented in this publication do not necessarily reflect the views of the organizations or agencies that provided support for the activity.

International Standard Book Number-13: 978-0-309-37835-2
International Standard Book Number-10: 0-309-37835-4

Additional copies of this workshop summary are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2015. *Assessing and improving the interpretation of breast images: Workshop Summary.* Washington, DC: The National Academies Press.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

WORKSHOP PLANNING COMMITTEE[1]

DIANA BUIST (*Chair*), Principal Investigator, Group Health Breast Cancer Surveillance

LORA BARKE, Radiologist, Radiology Imaging Associates, Invision Sally Jobe Breast Centers

PATRICIA A. CARNEY, Associate Director for Population Studies, Oregon Health & Sciences University Cancer Institute

PATRICIA GANZ, Distinguished Professor, Health Policy & Management and Medicine, University of California, Los Angeles (UCLA) Fielding School of Public Health, David Geffen School of Medicine at UCLA; Director, Cancer Prevention and Control Research, UCLA Jonsson Comprehensive Cancer Center

ABBE HERZIG, Patient Advocate

GRETA MASSETTI, Associate Director for Science, Division of Cancer Prevention and Control, Centers for Disease Control and Prevention

DIANA MIGLIORETTI, Dean's Professor in Biostatistics, Department of Public Health Sciences, University of California, Davis

BARBARA MONSEES, Evans Professor of Women's Health, Emeritus Chief, Breast Imaging Section, Washington University School of Medicine

TRACY ONEGA, Professor, Section of Biostatistics & Epidemiology, Dartmouth Medical School

ETTA D. PISANO, Vice President for Medical Affairs, Dean, Medical University of South Carolina

Project Staff

SHARYL J. NASS, Director, National Cancer Policy Forum

PATRICK ROSS, Research Assistant

HANNAH DURING, Senior Program Assistant

[1] Institute of Medicine planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published workshop summary rests with the workshop rapporteurs and the institution.

NATIONAL CANCER POLICY FORUM[1]

MICHAEL CALIGIURI (*Chair*), Chief Executive Officer, James Cancer Hospital and Solove Research Institute; Director, Ohio State University Comprehensive Cancer Center

PATRICIA A. GANZ (*Vice Chair*), Distinguished Professor, Health Policy & Management and Medicine, University of California, Los Angeles (UCLA) Fielding School of Public Health, David Geffen School of Medicine at UCLA; Director, Cancer Prevention and Control Research, UCLA Jonsson Comprehensive Cancer Center

AMY P. ABERNETHY, Chief Medical Officer and Senior Vice President for Oncology, Flatiron Health; Professor, Division of Medical Oncology, Duke University School of Medicine; Director, Center for Learning Health Care, Duke Clinical Research Institute

LUCILE ADAMS-CAMPBELL, Professor of Oncology, Associate Director for Minority Health and Health Disparities Research, Georgetown University Lombardi Cancer Center

KENNETH ANDERSON, Kraft Family Professor of Medicine, American Cancer Society Clinical Research Director, Jerome Lipper Multiple Myeloma Center, Harvard Medical School, Dana-Farber Cancer Institute

LOYCE PACE BASS, Health Policy Director, LIVESTRONG Foundation

MONICA BERTAGNOLLI, Professor of Surgery, Harvard University Medical School

OTIS BRAWLEY, Chief Medical Officer and Executive Vice President, American Cancer Society

CARLTON BROWN, Director of Professional Services, Oregon Nurses Association, and past president, Oncology Nursing Society

SERGIO CANTOREGGI, Chief Scientific Officer and Global Head of Research and Development, Helsinn Group

ROBERT W. CARLSON, Chief Executive Officer, National Comprehensive Cancer Network

[1] Institute of Medicine forums and roundtables do not issue, review, or approve individual documents. The responsibility for the published workshop summary rests with the workshop rapporteurs and the institution.

GREGORY CURT, Executive Director for External Relations in US Medical Affairs, AstraZeneca; Co-Chair, Life Sciences Consortium Task Force

WILLIAM S. DALTON, Chief Executive Officer, M2Gen Personalized Medicine Institute, Moffitt Cancer Center; Chair, American Association for Cancer Research Science Policy & Legislative Affairs Committee

GWEN DARIEN, Executive Vice President, Programs and Services, Cancer Policy Institute, Cancer Support Community

WENDY DEMARK-WAHNEFRIED, Associate Director for Cancer Prevention and Control, University of Alabama at Birmingham Comprehensive Cancer Center

JAMES DOROSHOW, Director, Division of Cancer Treatment and Diagnosis, Deputy Director, Clinical and Translational Research, National Cancer Institute

CAROL A. HAHN, Associate Professor of Radiation Oncology, Duke University Medical Center; Medical Director of Radiation Oncology, Duke Raleigh Hospital; Clinical Affairs and Quality Council Chair, American Society for Radiation Oncology

LORI HOFFMAN HŌGG, Veterans Health Administration National Oncology Clinical Advisor, Office of Nursing Services; Cancer Program Director, Albany Stratton Veterans Affairs Medical Center

SAMIR N. KHLEIF, Director, Georgia Health Sciences University Cancer Center, Georgia Regents University Cancer Center

LEE KRUG, Disease Area Head, Immuno-Oncology, Bristol-Myers Squibb

RICHARD A. LARSON, Hematology/Oncology Director, University of Chicago

MICHELLE M. LE BEAU, Arthur and Marian Edelstein Professor of Medicine, Director, Comprehensive Cancer Center, University of Chicago

SHARI LING, Deputy Chief Medical Officer, Center for Clinical Standards and Quality, Centers for Medicare & Medicaid Services

GRETA MASSETTI, Associate Director for Science, Division of Cancer Prevention and Control, Centers for Disease Control and Prevention

DANIEL R. MASYS, Affiliate Professor, Biomedical Informatics, University of Washington

MARTIN J. MURPHY, Chief Executive Officer, CEO Roundtable on Cancer

RICHARD PAZDUR, Director, Office of Oncology and Hematology Products, Food and Drug Administration

STEVEN PIANTADOSI, Phase One Foundation Endowed Chair and Director, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center

JENNIFER A. PIETENPOL, Director, Vanderbilt-Ingraham Cancer Center, Benjamin F. Byrd, Jr., Professor of Oncology, Professor of Biochemistry, Vanderbilt University

MACE L. ROTHENBERG, Chief Medical Officer & Senior Vice President, Clinical Development & Medical Affairs, Pfizer Oncology

AUGUST SALVADO, Senior Vice President, US Oncology Clinical Development & Medical Affairs, Novartis Oncology

ANDREW SCHIERMEIER, Senior Vice President, Head of Global Oncology, Merck Serono

RICHARD SCHILSKY, Chief Medical Officer, American Society of Clinical Oncology

DEBORAH SCHRAG, Chief, Division of Population Sciences, Professor of Medicine, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School

YA-CHEN TINA SHIH, Professor of Health Economics, Chief, Section of Cancer Economics and Policy, Department of Health Services Research, University of Texas MD Anderson Cancer Center

ELLEN V. SIGAL, Chair and Founder, Friends of Cancer Research

RALPH WEICHSELBAUM, Daniel K. Ludwig Professor and Chair, Department of Radiation Oncology; Director, Ludwig Center for Metastasis Research, The University of Chicago Medical Center

GEORGE J. WEINER, Director, Holden Comprehensive Cancer Center, University of Iowa, President, Association of American Cancer Institutes

National Cancer Policy Forum Staff

SHARYL J. NASS, Forum Director and Director, Board on Health Care Services

PATRICK ROSS, Research Assistant

HANNAH DURING, Senior Program Assistant

PATRICK BURKE, Financial Associate

Reviewers

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published workshop summary as sound as possible and to ensure that the workshop summary meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We wish to thank the following individuals for their review of this workshop summary:

Elena Elkin, Memorial Sloan Kettering Cancer Center
Louis Henderson, University of North Carolina
Ed Sickles, University of California, San Francisco
Dana Smetherman, Oschner Health Service

Although the reviewers listed above have provided many constructive comments and suggestions, they did not see the final draft of the workshop summary before its release. The review of this report was overseen by **Sue Curry,** University of Iowa. She was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the rapporteurs and the institution.

Acknowledgments

We thank the American Cancer Society for sponsoring for this project through a generous donation from the Longaberger Company's Horizon of Hope Campaign®.

Support from the many annual sponsors of the Institute of Medicine's National Cancer Policy Forum is crucial to the work of the Forum. Federal sponsors include the Centers for Disease Control and Prevention and the National Cancer Institute/National Institutes of Health. Nonfederal sponsors include American Association for Cancer Research, American Cancer Society, American Society for Radiation Oncology, American Society of Clinical Oncology, American Society of Hematology, Association of American Cancer Institutes, AstraZeneca, Bristol-Myers Squibb, Cancer Support Community, CEO Roundtable on Cancer, EMD Serono, Helsinn Healthcare SA, LIVESTRONG Foundation, National Comprehensive Cancer Network, Novartis Oncology, and the Oncology Nursing Society.

The Forum wishes to express its gratitude to the expert speakers whose presentations helped define the current challenges and opportunities for assessing and improving the quality of imaging interpretation in breast cancer screening and diagnosis. The Forum also wishes to thank the members of the planning committee for their work in developing an excellent workshop agenda.

Contents

ACRONYMS xix

INTRODUCTION 1

HISTORY OF MAMMOGRAPHY OVERSIGHT AND EFFORTS TO IMPROVE QUALITY 6

- Medical Audit, 7
- Pay-for-Performance, 13
- Breast Imaging Centers of Excellence, 13
- Continuing Medical Education, 13
- Digital Mammography, CAD, and 3D Mammograms, 15
- Ongoing Need for Quality Improvement, 16

CHALLENGES OF QUALITY INTERPRETATION 17

- Lack of Specialists, 18
- Lack of Experience, 18
- Lack of Skills-Building CME, 20
- Low Volume, 20
- Legal Challenges, 21
- Differential Access, 22

TRAINING AND EDUCATION 23

- Radiologist Training and Certification, 23
- Radiologic Technologist Training, Certification, and Licensing, 25

ASSESSING INTERPRETIVE PERFORMANCE 26

Audits of Interpretive Performance, 26

Accounting for Variability in Audit Metrics, 28

Other Factors That Can Affect Reader Metrics, 29

Criteria for Adequate Interpretive Performance, 30

Implementing Performance Criteria, 36

INTERVENTIONS FOR IMPROVING INTERPRETIVE PERFORMANCE 37

Test Sets, 37

Test Sets for Quality Assurance, 39

Lessons Learned from Test Set Use, 41

Correlation with Clinical Performance, 42

Challenges in Test Set Design, 43

Implementing Test Sets, 43

Assessing and Improving Mammography (AIM) Intervention, 44

MAMMOGRAPHY REGULATION AND QUALITY ASSURANCE IN OTHER COUNTRIES 44

RESEARCH NEEDS 47

POTENTIAL OPTIONS FOR ADDRESSING MAMMOGRAPHY QUALITY CHALLENGES 49

Training, 49

Mentoring and Coaching, 50

Reading One's Own Recalls, 51

Double Reading, 52

Volume Requirements, 53

Enhanced Audits, 54

Test Sets, 55

Seeding Positive Mammograms in Clinical Practice, 57

Performance-Based Incentives, 58

Prior Exams and Supplemental Images, 59

Communication with Consumers, 61

Federal Oversight of Breast Imaging, 62

WRAP-UP 63

REFERENCES 64

APPENDIX: Statement of Task and Workshop Agenda 69

Boxes, Figures, and Tables

BOXES

1 Suggestions Made by Individual Participants, 3
2 Basic Audit Definitions, 9
3 Basic Audit Calculations, 10
4 Explanations of the Themes That Categorize MISA Examination Questions, 39

FIGURES

1 National Mammography Database participating facilities, March 2015, 12
2 ACR Breast Imaging Centers of Excellence, 14
3 Variability in mammography interpretive performance in the United States, 16
4 Average performance by years of experience for screening mammography, 19
5 Analysis of first reader performance in the U.K. National Breast Screening Program, 47

TABLES

1 Breast Imaging Reporting and Data System (BI-RADS) Assessment Categories Standardized for All Modalities (2013), 8
2 Disparities in Breast Cancer Screening Access, Use, and Outcomes, 22
3 Final Cut-Points for Screening Mammography Using the Angoff Method, 31
4a Final Cut-Points for Diagnostic Mammography to Work Up Prior Abnormal Screening Exams Using the Angoff Method, 32
4b Final Cut-Points for Diagnostic Mammography to Work Up a Breast Lump Using the Angoff Method, 32
5 Combined Minimally Acceptable Performance Criteria for Radiologists with Complete Cancer Capture, 34
6 Analysis of First Reader Performance in the U.K. National Breast Screening Program, 46

Acronyms

ACR	American College of Radiology
ACS	American Cancer Society
AHRQ	Agency for Healthcare Research and Quality
ARRT	American Registry of Radiologic Technologists
BCSC	Breast Cancer Surveillance Consortium
BI-RADS	Breast Imaging Reporting and Data System
CAD	computer-aided diagnosis
CME	Continuing Medical Education
CMS	Centers for Medicare & Medicaid Services
DCIS	ductal carcinoma in situ
FDA	Food and Drug Administration
FN	false negative
FP	false positive
HIPAA	Health Insurance Portability and Accountability Act of 1996
IOM	Institute of Medicine

MISA	Mammography Interpretive Skills Assessment
MOC	maintenance of certification
MQSA	Mammography Quality Standards Act
MRI	magnetic resonance imaging
NCI	National Cancer Institute
NMD	National Mammography Database
NMQAAC	National Mammography Quality Assurance Advisory Committee
PERFORMS	Personal Performance in Mammography Screening
PPV	positive predictive value
RT	Registered Technologist
RVU	relative value unit
TN	true negative
TP	true positive

Workshop Summary

INTRODUCTION

Millions of women undergo screening mammography regularly with the hope of detecting breast cancer at an earlier and more curable stage. But the ability of such screening to accurately detect early cancers depends on the quality of mammography, including high-quality image acquisition and interpretation. To help ensure the quality of mammography, Congress passed the Mammography Quality Standards Act (MQSA) in 1994 and last reauthorized it in 2004. In advance of its expected reauthorization in 2007, Congress requested a consensus study from the Institute of Medicine (IOM) recommending ways to improve the quality of mammography, with an emphasis on image interpretation. The resulting report, *Improving Breast Imaging Quality Standards*, highlighted the need to decrease variability in mammography interpretation in the United States and identified gaps in the evidence needed to develop best practices (IOM, 2005). The consensus committee found that while the technical quality of mammography had improved since MQSA implementation, mammography interpretation remained quite variable, and that this variability limited the full potential of mammography to reduce breast cancer mortality by detecting breast cancers at an early stage.

Since 2005, a substantial new body of research pertaining to mammography has been published, including studies on reader volume, double readings, patient demographic characteristics, supplemental imaging, and other factors that can influence interpretive performance. Research has also examined criteria that could identify radiologists or facilities performing below minimum standards, and whether live instructor-led or self-paced interventions can improve that performance. This expanded evidence base has the potential to guide policies to improve the interpretation of mammograms.

To explore this evidence and its policy implications, the IOM's National Cancer Policy Forum, with support from the American Cancer Society (ACS), brought together experts and members of the public for the workshop,[1] "Assessing and Improving the Interpretation of Breast Images," which was held on May 12 and 13, 2015, in Washington, DC. At this workshop, clinicians and researchers, along with representatives from the Food and Drug Administration (FDA), National Cancer Institute (NCI), and patient advocacy organizations, discussed potential options for action to improve the quality of mammography interpretation. Topics discussed included

- challenges in the delivery of high-quality mammography, including a lack of mammography specialists and geographic variability in patient access to mammography;
- the impact of training and experience on interpretive performance;
- how best to measure interpretative performance and identify radiologists and facilities that could benefit from interventions to improve performance;
- various tools and interventions that could potentially be used to improve interpretation skills, such as self-tests, audits with feedback, and mentoring; and
- the impact of new technologies and supplemental imaging on interpretation of breast screening and diagnostic images.

This report is a summary of the presentations and discussions at the workshop. It is not a linear narrative of the presentations in each session, but rather is organized around the various themes discussed throughout the six sessions. A broad range of views and ideas were presented, and a summary of suggestions for potential solutions from individual participants is provided in Box 1. Additional details and context for these suggestions can be found

[1] The workshop was organized by an independent planning committee whose role was limited to the identification of topics and speakers. The workshop summary has been prepared by the rapporteurs as a factual account of what occurred at the workshop. Statements, recommendations, and opinions expressed are those of individual presenters and participants and are not necessarily endorsed or verified by the IOM. They should not be construed as reflecting any group consensus.

BOX 1
Suggestions Made by Individual Participants

Address Gaps in Federal Oversight of Breast Imaging

- Reauthorize the Mammography Quality Standards Act (MQSA) and publish revised regulations (including audit requirements) for public comment. (Helen Barr)
- Amend MQSA legislation or regulations to put more emphasis on image interpretation. (Barr)
- Continue discussions about the potential benefits of MQSA-like programs for other breast imaging modalities. (Barr)

Facilitate and Enhance Audits

- Define standard quality metrics to ensure comparisons of uniform measurements and use several metrics in audits to acquire a more complete picture of performance. (Diana Buist, Carl D'Orsi)
- Separate screening metrics from diagnostic metrics. (Buist, Robert Smith)
- Link more mammography facilities to tumor registries to assess important metrics such as true positives and negatives, and sensitivity and specificity. (Patricia Carney, D'Orsi, Barbara Monsees, Smith)
- Facilitate auditing by ensuring adequate long-term funding for the National Mammography Database (NMD) and the Breast Cancer Surveillance Consortium (BCSC). (Monsees, Smith)
- Encourage and incentivize more facilities to join the NMD. (Berta Geller)
- Devise measures that take into account biopsy capture rates and the bias that a lack of capture will have on cancer detection rates and other metrics. (Diana Miglioretti)
- Establish a collaboration between the American College of Radiology and the College of American Pathologists to link biopsy pathology reports to Breast Imaging Reporting and Data System classifications. (Brian Loy)
- Consider patient demographic characteristics when conducting audits and adjust accordingly. (Susan Harvey, Monsees, Tracy Onega, Matthew Wallis)
- Consider using test sets to assess sensitivity because readers have lower volumes in the United States. The audit may be sufficient to assess specificity. (Monsees)

continued

BOX 1 Continued

- Present audit data in cluster graphs or provide other visuals so it is easy to assess one's performance in relation to others and what areas need improvement. (Smith)
- Enhance transparency of audit data and methods to ensure that appropriate comparisons are being made across organizations and institutions. (Buist)
- Aggregate data over multiple years and consider confidence intervals when assessing interpretive performance metrics to account for the variability over time. (Miglioretti)
- Assess the accuracy of performance metrics with ratio reliability (the ratio of between-provider variation to total variation). (Rebecca Hubbard)

Address Research Needs

- Assess how long performance improvements last following interventions, such as test sets and mentoring. (Smith)
- Conduct research to determine the best ways to train physicians and determine what works best for lifelong learning. (Buist, Geller)
- Evaluate whether providing more flexibility in residency training to intensify breast imaging experience helps to improve proficiency. (Carney)
- Evaluate the impact of Centers of Excellence on mammogram interpretation and patient outcomes. (Buist)
- Conduct research to better understand the relevance of test sets to clinical performance. (Mireille Broeders)
- Undertake research and development on test sets, including identifying ranges for specialized test sets and best strategies for communicating results in a standardized fashion, and validating the value of test sets. (Smith)
- Develop a stronger evidence base to define who should have supplemental imaging and with what types of technologies. (Monsees)

Enhance Training, Continuing Education, and Maintenance of Certification

- Establish a supportive training environment and a special breast imaging curriculum for new radiologists who plan to enter the breast imaging workforce. (Carney, Monsees, Smith)
- Develop better training programs for mammography technologists, with a rigorous focus on measurement and evaluation of proper positioning. (Barr, Stephen Taplin)

- Enable more coaching of physicians during training and in clinical practice, and engage radiologists identified as high performers as mentors for performance improvement. (Carney, D'Orsi, Wallis)
- Link audits to mentoring if suboptimal performance is detected. (D'Orsi)
- The American Board of Radiology could specify that test sets are acceptable activities for Maintenance of Certification for radiologists. (Monsees)
- Develop test sets that adjust in real time the level of difficulty of questions and cases based on how well radiologists are performing as they make their way through the test cases (computierized adaptive testing). (Buist, Kelly Walborn)
- Document both the short- and long-term effects of various educational opportunities, such as selectorships at Centers of Excellence and whether various Continuing Medical Education programs and self-assessment tests improve outcomes. (Buist)

Improve Accuracy of Interpretation

- Require radiologists to do diagnostic work-ups for a minimum percentage of their own recalls. (Buist)
- Incorporate double reads in screening programs for radiologists with less experience or lower volume. (Smith, Wallis)
- Consider increasing the minimum interpretive volume in the United States. (Buist, Smith)
- Implement a minimum diagnostic interpretation volume requirement. (Buist)
- Seed a radiologist's clinical caseload with images of confirmed cancers. (Carney, Miglioretti, Smith)

Provide Incentives to Monitor and Improve Performance

- Reward quality performance by providing recognition and awards that mean something to the professional and patient community. (Dana Smetherman)
- Provide payment incentives for quality performance. (D'Orsi, Loy)
- Provide higher reimbursement rates to those who participate in the NMD or the BCSC and are regularly assessing their quality metrics. (Patricia Ganz)
- Create a culture of evaluation in radiology and other medical specialties in which it is routine to assess one's performance and a clear pathway for improvement is offered, if needed. (Smith)
- Incentivize participation in performance assessments through "safe harbor" provisions for those who participate in continuous quality improvement, including adhering to audits, regular reviews, and proficiency testing. (Lora Barke, Smith)

continued

BOX 1 Continued

Enhance Consumer Knowledge and Access

- Inform patients about quality metrics, in understandable language that is free of complex statistical concepts and terminology. (Pisano, Walborn)
- Develop a standard way of measuring all breast imaging practices and graphically depict results in a format that is easy for consumers to grasp. (Geller)
- Collect per-capita data on geographic access to mammography. (Onega, Smith)
- Use teleradiology to help alleviate the disparities in access to high-volume readers or new technologies. (Onega)
- Eliminate barriers that impede access to prior images. (Loy)

throughout the workshop summary. The workshop Statement of Task and agenda can be found in the Appendix. The speakers' presentations (as PDF and audio files) have been archived online.[2]

HISTORY OF MAMMOGRAPHY OVERSIGHT AND EFFORTS TO IMPROVE QUALITY

Breast imaging has been the focus of many debates over several decades, noted Diana Buist, senior scientific investigator for Group Health Breast Cancer Surveillance at the Group Health Research Institute, in her opening remarks at the workshop. She said most of the debate has centered around when we should start screening, how often we should screen, and when we should stop. But she said the goal of this workshop was to discuss ways to improve the quality of breast imaging. The quality and accuracy of mammography depend on both technical and human factors, but discussions at the workshop emphasized image interpretation.

In the 1980s it became apparent that mammography had serious quality issues, said Robert Smith, senior director of cancer control at ACS. This

[2] See https://www.iom.edu/Activities/Disease/NCPF/2015-MAY-12.aspx.

led the American College of Radiology (ACR) to create its mammography accreditation program and, along with the Centers for Disease Control and Prevention, to develop quality assurance standards. Then in 1992, Congress passed the MQSA, which went into effect in 1994. Regulations promulgated under MQSA put a primary emphasis on image quality, but also stipulated requirements pertinent to interpretive performance, including (1) medical audit; (2) requirements related to training, including initial training and Continuing Medical Education (CME); and (3) interpretive volume, including initial and continuing experience (minimum of 960 mammograms/2 years for continuing experience) (IOM, 2005).

After several studies done in the 1990s and early 2000s documented the wide variability in interpretive skills in mammography, ACR added skills assessments to its mammography accreditation program. In addition, an IOM committee of experts was formed to determine how to improve mammography quality further, in preparation for the anticipated 2007 reauthorization of MQSA. The IOM report that stemmed from this committee's deliberations was published in 2005 and, as reported by Etta Pisano, Dean Emerita and Distinguished University Professor, Medical University of South Carolina, made several recommendations related to improving mammography interpretation, including to

- revise and standardize the MQSA-required medical audit;
- facilitate a voluntary advanced medical audit with feedback;
- designate specialized Breast Imaging Centers of Excellence that undertake demonstration projects and evaluations; and
- study the effects of continuing medical education, reader volume, double reading, and computer-aided diagnosis (CAD) on the quality of mammogram interpretation (IOM, 2005).

"Many of the IOM recommendations about improving interpretation have been implemented, but by professional societies and not by the government. If there's a gap, it's in the assessment of whether they are effective," Pisano concluded.

Medical Audit

The medical audit originally required by MQSA regulations entailed recording all positive mammograms and biopsy results. The results of both mammogram interpretations and biopsies were to be analyzed and shared annually with the designated interpreting physician. But no specific metrics were required for that analysis. A number of changes were made to the medical audit with the publication of the final regulations in 1998, including defining a positive mammogram as suspicious or highly suggestive of malignancy (assessed as 4 or 5 on the Breast Imaging Reporting and Data System [BI-RADS] 4 or 5; see Table 1), requiring that facilities have a system for following up on positive mammograms, including obtaining biopsy pathology results, and for correlating the pathology results with the final assessment categories. The MQSA regulations also require audit data to be reviewed at least every 12 months. Compliance with the medical audit requirements are checked at the time of annual facility inspection.

TABLE 1 Breast Imaging Reporting and Data System (BI-RADS) Assessment Categories Standardized for All Modalities (2013)

Assessment	Management	Likelihood of Cancer
Category 0: Incomplete—need additional imaging evaluation and/or prior mammograms for comparison	Recall for additional imaging and/or comparison with prior examination(s)	Not applicable
Category 1: Negative	Routine screening	Essentially 0% likelihood of malignancy
Category 2: Benign	Routine screening	Essentially 0% likelihood of malignancy
Category 3: Probably benign	Short-interval (6-month) follow-up or continued surveillance imaging	>0% but ≤2% likelihood of malignancy
Category 4: Suspicious (also 4A/B/C for mammography and ultrasound)	Tissue diagnosis	>2% but <95% likelihood of malignancy
Category 5: Highly suggestive of malignancy	Tissue diagnosis	≥95% likelihood of malignancy
Category 6: Known biopsy-proven malignancy	Surgical excision when clinically appropriate	Not applicable

SOURCES: D'Orsi presentation, May 12, 2015; http://www.acr.org/Quality-Safety/Resources/BIRADS (accessed August 16, 2015).

The 2005 IOM committee thought these requirements were too vague, according to Pisano. Thus, the committee recommended including in the medical audit-specific metrics such as the proportion of BI-RADS 4 and 5 ratings that lead to a diagnosis of breast cancer (positive predictive value 2 or PPV_2; see Boxes 2 and 3), the cancer detection rate per 1,000 women, and a measurement of the abnormal interpretation rate, that is, those readings that lead to additional imaging or biopsy (recalls for additional imaging rate plus overall biopsy numbers). The IOM committee also recommended auditing screening exams separately from diagnostic mammograms, allow-

BOX 2
Basic Audit Definitions

Sensitivity: Percentage of cancer detected from all cancers

Specificity: Percentage of negative cases identified when no cancer is present

Recall rate: Percentage of screens given BI-RADS 0 (additional imaging evaluation needed)

False-negative result: A test result that incorrectly indicates that breast cancer is not present when in fact it is present

False-positive result: A test result that incorrectly indicates that breast cancer may be present when in fact it is not

Abnormal interpretation rate: Percentage of all positive exams/all exams

Accuracy: Percentage of cancer and negative cases identified from all cases

Positive Predictive Value_1 (PPV_1): Percentage of screening exams with a positive interpretation and cancer diagnosed within 1 year

Positive Predictive Value_2 (PPV_2): Percentage of all positive exams with a biopsy recommended (BI-RADS 4/5) and cancer diagnosed within 1 year

Positive Predictive Value_3 (PPV_3): Percentage of biopsies done with a positive interpretation (BI-RADS 4/5*)* and a known diagnosis of cancer in 1 year

Cancer detection rate (per 1,000): Number of cancers detected per 1,000 women

Percentage of minimal cancer: Percentage of all cancers detected that are ≤1 cm or ductal carcinoma in situ

SOURCES: D'Orsi presentation, May 12, 2015; IOM, 2005.

BOX 3
Basic Audit Calculations

Sensitivity = TP/TP + FN
Specificity = TN/TN + FP
Accuracy = TP + TN/TP + TN + FP + FN
PPV$_1$ = TP/TP + FP (percentage of positive screens [BI-RADS 4/5] with a cancer diagnosis within 1 year)
PPV$_2$ = TP/TP + FP (percentage of positive exams [BI-RADS 4/5] with a *recommendation* for biopsy and a cancer diagnosis within 1 year)
PPV$_3$ = TP/TP + FP (percentage of known biopsies *done* for patients with BI-RADS 4/5 with a cancer diagnosis within 1 year)

NOTE: FN = false negative; FP = false positive; PPV = positive predictive value; TN = true negative; TP = true positive.
SOURCE: D'Orsi presentation, May 12, 2015.

ing physicians to combine their outcomes if they work at more than one facility, and verifying data collection and analysis during an FDA inspection of a facility (IOM, 2005).

Helen Barr, Director, Division of Mammography Quality Standards, FDA, pointed out that shortly after the IOM report was published in 2005, she and her former colleague Charles Finder drafted amended regulations to address some of the IOM recommendations, particularly with regard to those focused on further enhancing the medical audit. Those amended regulations were submitted in 2007 but still have not been published for public comment. MQSA was not reauthorized when it expired in 2007, perhaps due to other legislation taking precedence, Barr said. However, Congress has continued to fund the program without that authorization, and FDA continues to certify and inspect facilities and collect fees to do so. But reauthorization opens up the opportunity for the statute to be amended, Barr stressed.

Pisano noted that the 2005 IOM report also specified that a voluntary advanced medical audit with feedback should include

- collecting patient characteristics and tumor staging;
- creating a central data and statistical center to collect and analyze the data;
- providing feedback to interpreting physicians;
- developing, implementing, and evaluating self-improvement methods; and
- reporting aggregate data to the public.

According to Pisano, none of these recommendations have been fully implemented, except creating a central data and statistical center, which was met with the creation of ACR's National Mammography Database (NMD).[3] The NMD was established in 2009 and complements NCI's Breast Cancer Surveillance Consortium (BCSC),[4] a collaborative network established in 1994 consisting of seven mammography registries with linkages to tumor and/or pathology registries. (Two of the 7 have stopped contributing data and are no longer active, but their data are still included in the BCSC data archives and in many analyses.)

Consisting of 275 registered sites with 162 contributing data, the NMD is a registry for breast imaging that allows facilities and physicians to monitor and improve quality using standardized data elements and measures consistent with BI-RADS, reported Carl D'Orsi, director of breast imaging research, Emory Healthcare. The database has information from more than 9 million exams, with good representation across the country and across practice types and locations, he said. Data collected include patient demographic characteristics, height, weight, and personal and family history of breast cancer. The NMD also collects exam information, including the date of the exam, physician and facility identifying codes, breast density, assessment category, and management recommendation. In addition, data are included on outcomes, such as biopsy procedure date and result, and for breast cancers detected, tumor size, nodal status, and tumor stage. The NMD data submission is automated, with data sent directly from certified vendors or through certification of home-grown software. The NMD is expected to expand to include magnetic resonance imaging (MRI) and ultrasound data by late 2015. Most states have a facility that participates in the NMD (see Figure 1).

[3] See http://www.acr.org/Quality-Safety/National-Radiology-Data-Registry/National-Mammography-DB (accessed July 10, 2015).

[4] See http://breastscreening.cancer.gov/about (accessed July 10, 2015).

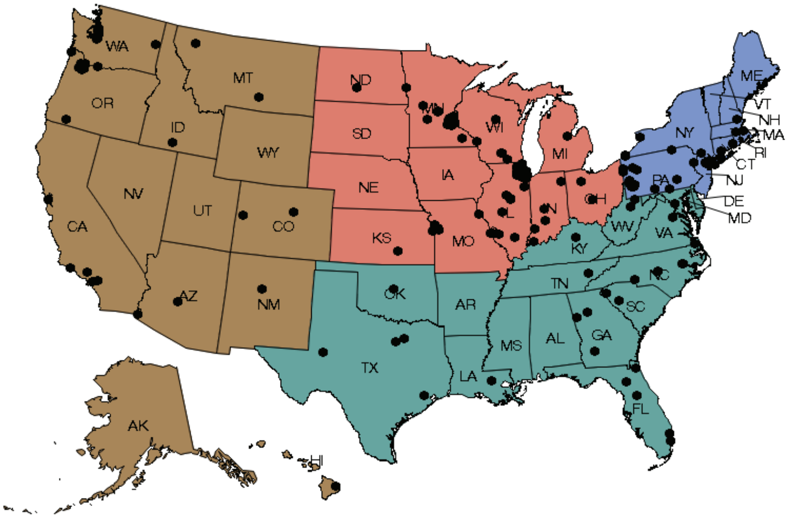

FIGURE 1 National Mammography Database participating facilities, March 2015.
SOURCE: D'Orsi presentation, May 12, 2015.

D'Orsi pointed out that the data collected in 2013 indicate the NMD correlates well with the BCSC, having similar recall rates, cancers detected, and PPV2 rates. Participating facilities have regular audits of their data, and both facilities and physicians are given feedback in performance charts and graphs, as well as conference calls about what the audits reveal. "If you belong to the NMD, you get your stats compared to the rest of the people in the program and compared to the BCSC data," said Barbara Monsees, Emeritus Chief, Breast Imaging Section, Washington University School of Medicine.

But unlike the BCSC, the NMD is currently not linked to tumor registries, so there is no way to calculate sensitivity and specificity rates from its data, Monsees noted. "One of our top priorities on our wish list is to have links to tumor registries," she stressed, adding that the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule[5] impedes such links. "We have limitations on the types of data that can be collected and interrogated," she said.

[5] Federal regulations promulgated under the Health Insurance Portability and Accountability Act of 1996 to protect the privacy of individually identifiable health information. See http://www.hhs.gov/ocr/privacy/index.html (accessed July 20, 2015).

Pay-for-Performance

The 2005 IOM report also recommended developing incentives, such as pay-for-performance, for those physicians that opt to participate in advanced audits and are shown to be meeting performance criteria. Pisano noted that although some payers have implemented pay-for-performance mammography programs, they are not uniformly available.

Breast Imaging Centers of Excellence

The IOM consensus committee also called for Breast Imaging Centers of Excellence, which would participate in both basic and advanced audits, and could also undertake studies on the influence of high-volume reading or double reading on interpretation accuracy. "They were seen as test centers, not just demonstration projects," Pisano pointed out, although the Centers were not intended to conduct clinical trials. The Centers of Excellence were also imagined as places where physicians could receive training and be linked to facilities that provided comprehensive multidisciplinary breast cancer care. Centers of Excellence could also provide regional mammography interpretation in areas lacking mammography experts, and be incentivized by pay-for-performance metrics.

Although no federal action was taken in response to this recommendation, in 2007 ACR started its own Breast Imaging Centers of Excellence program to complement the Breast Cancer Centers of Excellence program.[6] There are currently more than 1,200 such breast imaging centers scattered throughout the country, as shown in Figure 2. Wyoming is the only state that lacks a Breast Imaging Center of Excellence.

To achieve a designation as a Center of Excellence, facilities have to earn accreditation in all of ACR's voluntary breast imaging programs and modules, including mammography, ultrasound, stereotactic biopsy, and, by 2016, MRI as well.

Continuing Medical Education

MQSA regulations require radiologists who interpret mammograms to have CME relevant to mammography. To help satisfy that requirement and to further improve mammography interpretation, Pisano reported that

[6] See http://www2.nqmbc.org (accessed September 10, 2015).

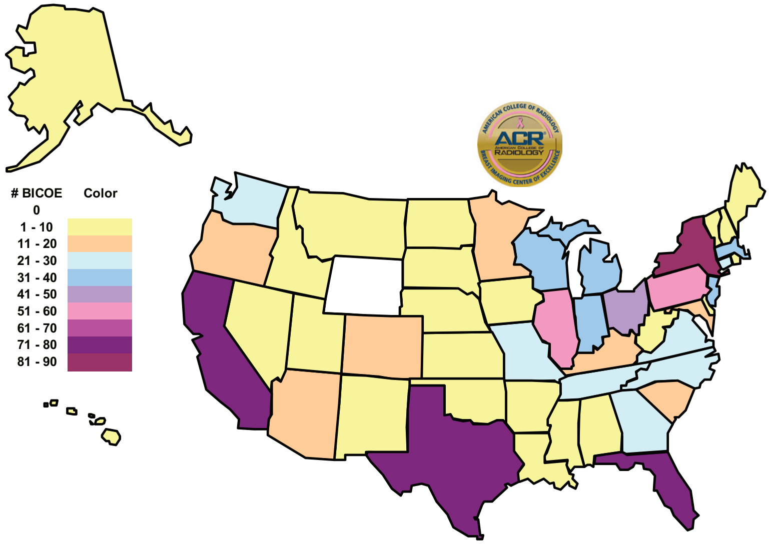
# BICOE
Color
0
1 - 10
11 - 20
21 - 30
31 - 40
41 - 50
51 - 60
61 - 70
71 - 80
81 - 90
AMERICAN COLLEGE OF RADIOLOGY
ACR
AMERICAN COLLEGE OF RADIOLOGY
BREAST IMAGING CENTER OF EXCELLENCE

FIGURE 2 ACR Breast Imaging Centers of Excellence. As of May 4, 2015, there was a total of 1,234 Breast Imaging Centers of Excellence.
SOURCE: Pisano presentation, May 12, 2015.

ACR established breast imaging boot camps, as well as a Mammography Case Review, which is an online self-paced review of 118 breast imaging cases that can be done for CME credit. This collection of difficult as well as easy cases also can be used to satisfy volume requirements. As Pisano noted, "This test set is more valuable than reading 480 consecutive cases at a breast imaging center because with the latter, you might only see one or two cancers, but the test cases include cancers, and all the abnormal cases are pathology proven, whereas if you're in a practice all you have is the opinion of the person sitting next to you about whether the case is positive or not, so you rise to the level of that person training you as opposed to the known truth."

Pisano pointed out that the aim of the boot camp program is to improve radiologists' interpretive skills, but is not intended to be used as a screening tool to assess which physicians should improve their performance. "I would assume most radiologists want to do a good job and that they will voluntarily seek out more CME in mammography if they do poorly with the test cases, but there is no formal mechanism to get them to stop reading mammograms" she said.

The Society for Breast Imaging also now provides at its annual symposia a screening self-assessment case set imported from the United Kingdom known as Personal Performance in Mammography Screening (PERFORMS), which is discussed further in the section "Test Sets for Quality Assurance."

Digital Mammography, CAD, and 3D Mammograms

Several participants noted that the transition to digital images, which occurred primarily after the 2005 IOM report, has also improved mammogram interpretations. Pisano and Monsees noted that the wider recording latitude and the elimination of film processors in digital mammograms has substantially reduced the variability in the technical quality of mammograms previously seen in film. Pisano added that the medical literature indicates that digital mammograms are easier to read, while Monsees noted that quality control is easier and more streamlined with digital mammography, and that digital mammograms are easier to transfer and track compared to film.

With the advent of digital mammograms, most radiologists also now use CAD, which Pisano said can increase the cancer detection rate in women with dense breasts, but added there is conflicting evidence on whether CAD improves interpretations. "CAD does do a good job flagging microcalcifications on screening mammograms," and improves detection of ductal carcinoma in situ (DCIS) and of smaller breast cancers, Monsees said. She cited one study that found that use of CAD during screening mammography among Medicare enrollees is associated with increased DCIS detection, the diagnosis of invasive breast cancer at earlier stages, and increased diagnostic testing among women without breast cancer (Fenton et al., 2013). But Buist stressed that "the evidence around CAD does not demonstrate improved performance," including the findings of what she said was a more definitive paper that was expected to be published shortly (Lehman et al., under review).

Monsees also reported on breast tomosynthesis, a newer imaging technology that enhances mammography by allowing the radiologist to see slices through the breast. It is sometimes referred to as 3D mammography, although it should not be, because it is not really a 3D examination. Digital tomosynthesis is increasingly being used for breast cancer screening. Citing both prospective and retrospective studies, Monsees said research indicates that this technology can improve cancer detection rate, decrease recall rates, and improve screening performance in all but fatty breasts (Ciatto et al.,

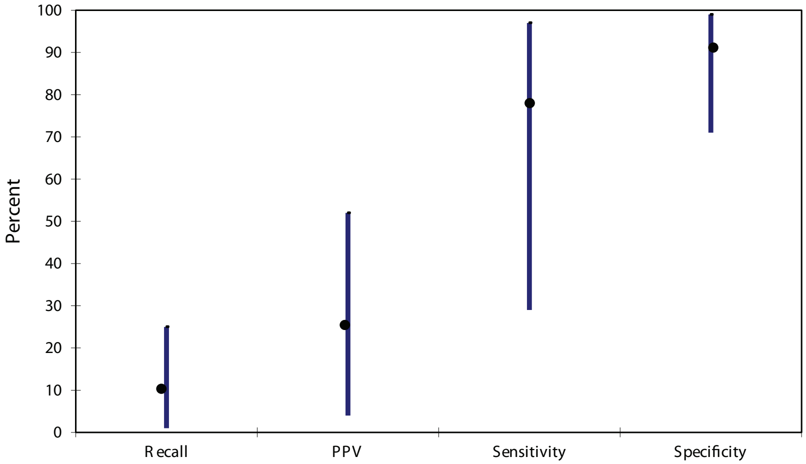

FIGURE 3 Variability in mammography interpretive performance in the United States.
NOTE: PPV = positive predictive value.
SOURCE: Buist presentation, May 12, 2015

2013; Friedewald et al., 2014; Haas et al., 2013; Rose et al., 2013; Skaane et al., 2013, 2014). But she noted there are only data for prevalent screens and not for incident screens.[7]

Ongoing Need for Quality Improvement

Despite all these efforts to improve the interpretation of mammograms, there is an ongoing need for better quality, several participants noted. Buist presented data showing the tremendous variability in the performance of U.S. radiologists interpreting mammograms, with sensitivity rates varying from around 30 percent to more than 90 percent, and PPV rates from less than 10 percent to just more than 50 percent (see Figure 3). Pisano suggested that when it comes to mammogram interpretive quality, "the bar is set too low" in part because "we don't have a system in this country to do anything about this."

Patricia Ganz, distinguished professor, Health Policy & Management, University of California, Los Angeles, stressed that it is detrimental to

[7] Incident screens are screening tests performed at regular intervals after an initial (prevalent) screen for a breast cancer.

encourage women to have regular mammograms if they are not going to be accurate. "If we don't have a quality product then just checking the box and saying it's being done doesn't really help women in this country," she said. Matthew Wallis, director of the Cambridge and Huntington Breast Screening Service, added that "If you are recalling, you've got to be able to see and cope with the tears of distress that are associated with the damage you cause when you write to women saying 'please come back.' Anybody who says recall is not a stressful process does not live in my world."

Ganz also stressed that, "The risks of misdiagnosis or a harm from an overdiagnosis or recall are huge at the population level. It is going to cost the health care system a lot." Smith agreed and said, "If you can improve quality and reduce avoidable recalls, you're going to save money." He added that enormous societal costs are avoided if breast cancer is identified and treated early, including the costs of job absenteeism due to cancer treatment or while caring for a relative with advanced cancer, as well as the costs of disability payments and the loss of a valued employee. "Cancer is enormously expensive to the workplace, and prevention and early detection represents a tiny fraction of the monthly bill for health insurance. So we need to communicate to employers the value of a good, high-quality program." Susan Harvey, director, Johns Hopkins Breast Imaging Section, agreed, adding that improving the quality of mammography interpretation would also reduce costs from reexcisions of tumors, noting that at Hopkins, the reexcision rate is 30 percent and involves multiple trips to the operating room, general anesthesia, and time off from work.

Wallis stressed that "You've got to identify the poor performers because they're probably doing harm to women," while others emphasized that radiologists want to do the best job possible. "We all want to figure out the best way to do a better job," said Lora Barke, a radiologist from Radiology Imaging Associates. Smith added, "We have the opportunity to address a very important challenge, which is ensuring that women getting mammograms can have the confidence that they're going to be accurate."

CHALLENGES OF QUALITY INTERPRETATION

Several workshop speakers and participants discussed challenges to accurately interpreting mammograms that need to be overcome. These challenges include a lack of mammography specialists; a lack of experience, especially among radiologists practicing in low-volume clinics; malpractice concerns; and differential access to quality facilities.

Lack of Specialists

Monsees reported that some facilities have general radiologists reading all their mammograms, whereas others have radiologists who specialize in breast imaging. A kind of hybrid also exists in which general radiologists read screening mammograms, but there is centralized interpretation of diagnostic imaging, work-ups, and biopsies by breast imaging specialists or vice versa. She noted that the digital transition has facilitated centralized interpretations of mammograms and that many multioffice practices now employ them.

Although there have been fellowships for breast imaging since 1985 (as of 2007, 55 institutions offered a fellowship program in breast imaging [Baxi et al., 2009]), there is a lack of radiologists who specialize in breast imaging, several speakers noted. ACR data indicate that only 647 radiologists devote themselves solely to breast imaging, compared to more than 9,000 radiologists who report spending some time reading mammograms, and the more than 7,000 who report spending some time doing breast imaging (ACR, unpublished data). Ganz stressed that more specialization is needed in radiology and other medical fields because "There is too much information and we can't know it all. We have to be realistic and change what the expectations are for physicians. We have to say 'do what you are most comfortable doing and you can't do everything.'"

Some participants discussed whether radiologic technologists should also be specialized, with several speakers pointing out that the positioning of the breast during mammography can influence the accuracy of a radiologist's interpretation. "Dedicated technologists bring to the table better positioning, better compression and that makes us better breast imagers," Monsees said. "The days of general technologists who do a few chest X-rays, some CT scans, and some mammograms should be in the past."

Lack of Experience

A lack of experience correlates with less than optimal interpretation of mammograms, Buist stressed, presenting data which show that recall rates decrease as years in practice increase, with the inverse being true for PPV_1 rates (Miglioretti et al., 2009) (see Figure 4).

Patricia Carney, associate director for population studies, Oregon Health & Science University, noted that this paper suggests radiologists are not clinically ready to work independently until they have been practicing

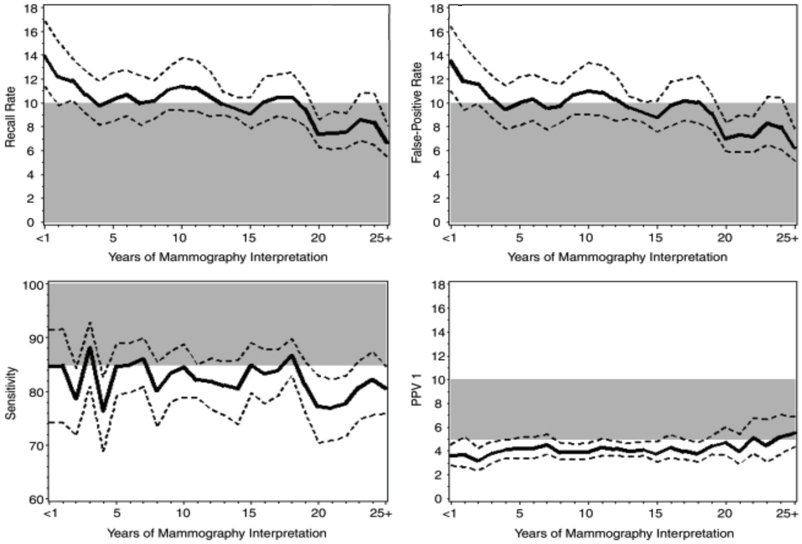

FIGURE 4 Average performance by years of experience for screening mammography.
NOTE: Solid lines show population-average performance of screening mammography. Dashed lines show 95 percent confidence interval. Shaded area represents the desirable goals for performance of interpretation as defined by the Agency for Healthcare Research and Quality. The study analyzed data from the Breast Cancer Surveillance Consortium (BCSC). To assess performance, the authors included screen-film screening mammograms interpreted by a participating radiologist at a BCSC facility during the study period for all women ages 18 and older (excluding those with with breast augmentation or a history of breast cancer).
SOURCES: Buist presentation, May 12, 2015; Miglioretti et al., 2009.

for 5 years, at which point their performance metrics in mammography tend to meet standards set by the Agency for Healthcare Research and Quality (AHRQ).

Debra Monticciolo, professor of radiology, section chief, breast imaging, Texas A&M College of Medicine, responded that the more experience the better the performance, but that this is true for any area of radiology, as well as for surgery and other medical specialties. “I don’t think that means you don’t certify them because the process right now does prepare radiologists very well to do their work,” she said. She added that radiologists who opt to do selectorships or fellowships in breast imaging will have more experience in a concentrated period of time. Monsees agreed and said, “If you shine a light on any part of medicine you would say that when a person goes out into practice, on day one they are not at their top. It's just that we're shining a light on this area of radiology because we have so much data, but I'm going to guess that we're no different than other medical specialties."

Lack of Skills-Building CME

Once radiologists are in practice, they are expected to continue their education through CME programs, said Berta Geller, research professor, Office of Health Promotion Research, College of Medicine, University of Vermont. She noted that although MQSA regulations require radiologists to have CME related to mammography, it does not specify what type of CME they are required to fulfill. Radiologists could satisfy their CME requirements by only taking classes aimed at increasing knowledge, but not at building their mammogram interpretation skills. Even when the focus of CME is improving such skills, it may not be effective at doing so, Geller explained. She cited several studies that found skills-building CME programs often increase sensitivity, but not specificity or other performance metrics (Adcock, 2004; Berg et al., 2002; Carney et al., 2012; Geller et al., 2014; Linver et al., 1992; Scott, 2006; Urban et al., 2007). Dana Smetherman, vice chair, Department of Radiology, Oschner Health Service, added that CME is seen as a bridge to quality, but it is provided to physicians through local institutions or through national societies. Although there is oversight and accreditation for CME, "there is really not a tremendous amount of instruction about how to work with adult learners," she said.

Low Volume

Because of the low probability of finding a cancer (less than 1 percent) in screening mammograms, radiologists who do not have large practices may not accrue adequate experience detecting breast tumors. The volume of mammograms read by radiologists correlates with their interpretive performance, several studies indicate. Buist presented a compilation of results of several studies and showed that although most studies do not find a statistically significant association between volume and sensitivity, there is stronger evidence that there is a statistically significant link between increased volume and lower false-positive rates (Hofvind et al., 2008; IOM, 2005; Perry et al., 2008; Roberge, 2007). A study she conducted using data from the BCSC found that low volume was significantly linked to a higher false-positive rate, a lower cancer detection rate, and a lower sensitivity among radiologists who mostly read screening as opposed to diagnostic mammograms (Buist et al., 2014). A subanalysis found no consistent association between volume and diagnostic performance, although the highest false-positive rates were among radiologists for whom diagnostic exams composed fewer than 20 percent of their caseloads (Haneuse et al., 2012).

Volume requirements for radiologists vary from program to program across Canada, said Isabelle Théberge, vice president of scientific affairs, National Public Health Institute in Quebec. She reported that compared with radiologists who always maintain a volume of at least 500 mammograms per year, those with less than that volume experienced a 20 percent reduction in sensitivity and a 91 percent increase in false-positive rates (Théberge et al., 2014). Measuring interpretive accuracy as the ratio of sensitivity to false-positive rate, Théberge determined that interpretive accuracy increased with each volume increase of 100 mammograms annually, with the greatest gains observed among radiologists reading less than 3,000 mammograms annually. She said these results indicate that raising the volume of mammograms read by radiologists could help to minimize false-positive rates without changing sensitivity. Théberge's study helped convince Canada's Ministry of Health and the Quebec Association of Radiologists to gradually increase the volume of mammography interpretation requirements for radiologists, she noted. The latter has increased the volume threshold to 750 mammograms annually, with the threshold being raised to 1,000 by January 2016. But Smith and others noted that it is difficult for many U.S. radiologists, especially those in rural practices, to meet such high volume requirements.

Legal Challenges

Monsees noted that performance expectations are high for radiologists interpreting mammograms in the United States, and that radiologists frequently are sued for malpractice if they fail to identify a cancer in a mammogram. Concerns about potential malpractice suits could foster higher recalls, she said. D'Orsi agreed that malpractice concerns are a challenge for the evaluation culture needed to support improved performance. He noted that there has not been sufficient tort reform, and consequently malpractice concerns "are the baby elephant in the room."

Differential Access

Tracy Onega, associate professor, Section of Biostatistics & Epidemiology, Dartmouth Medical School, reported on the variable use and access women have to quality mammography facilities. Geographic access does not seem to limit mammography use for most women, but some populations, such as Native American and rural women, may have travel times greater than 30 minutes to access mammography services (Onega et al., 2014) (see Table 2).

In addition, geographic distribution may limit women's access to other types of breast imaging, such as MRI. Onega cited a study that found that half of breast imaging facilities took nearly a decade to make the transition to digital mammography (Miglioretti et al., 2009) and noted that MRI breast imaging is also slowly diffusing into clinical practice. "When new technologies come on board there might be quite a lag until we can achieve equal distribution and that can widen disparities," Onega said. She also stressed that geographic access does not necessarily correlate with use, as women with low incomes tend to be less likely to report a recent mammogram compared with women with higher incomes (Miller et al., 2012).

Smith noted that although the spatial distribution of mammography

TABLE 2 Disparities in Breast Cancer Screening Access, Use, and Outcomes

% Women with Travel Time > 30 Min. to Closest Mammography		Women 40+ Yrs. with Mammography in Past 2 Years (BRFSS)		% Late-Stage Breast Cancer (Stage III or IV) at Diagnosis	
White	12.6%	White	75.4%	White	7.6%
Black	6.4%	Black	78.6%	Black	11.2%
Asian	2.2%	Asian	73.7%		
Native Amer.	39.6%	Native Amer.	63.9%	Breast cancer mortality rates	
Urban	0.5%	≥$75,000/yr.	83.8%	(per 100,000)	
Rural	27.9%	<$35,000/yr.	68.1%	White	22.7
				Black	30.8
				Hispanic	14.8
Onega et al., 2014		Miller et al., 2012		DeSantis et al., 2014	

NOTE: BRFSS = Behavioral Risk Factor Surveillance System.
SOURCES: Onega presentation, May 12, 2015; DeSantis et al., 2014: Miller et al., 2012; Onega et al., 2014.

facilities may be geographically even, population differences may impede access to certain facilities because the demand for mammograms is greater than the number of facilities and personnel needed to meet that demand. He suggested collecting per-capita data on geographic access to mammography. Onega agreed and said she hopes to provide those data in the future.

In addition, even though mammography facilities are evenly distributed spatially in this country, the quality of those facilities is uneven, studies indicate. One study found considerable variation in sensitivity, PPV_1 and PPV_2, at the facility level (Taplin et al., 2008). Another study found no significant differences in the sensitivity rates of screening facilities serving vulnerable versus non-vulnerable women, but did find that the former tend to have significantly greater specificity. However, false-positive rates for diagnostic mammograms were higher at facilities serving vulnerable women (Goldman et al., 2008, 2011). Another study found that facilities serving the most vulnerable populations were significantly less likely to detect tumors with a good prognosis and significantly more likely to detect tumors with a poor prognosis (Goldman et al., in press). "There's mixed evidence of differential quality by sociodemographic characteristics at the facility level," Onega summarized.

Harvey noted that in the breast imaging program she directs for Johns Hopkins, mammography is performed at six facilities, ranging from inner-city Baltimore sites to wealthy suburban areas to a rural site on the eastern shore of Maryland. She has found that interpretive data vary dramatically from site to site in this program, even though the same radiologists work at all sites. For example, at the site in the city serving an underresourced population, the recall rate is 16 percent and the cancer detection rate is 8 per 1,000 women. But when the same group of radiologists read in the suburban site, the recall rate is 9 percent and the cancer detection rate is 3 per 1,000 women. "We're not smarter or dumber at one facility versus another or providing different quality by site, but rather we're seeing very different populations," she stressed.

TRAINING AND EDUCATION

Radiologist Training and Certification

Debra Monticciolo provided an overview of radiologists' training, certification, and maintenance requirements. Board certification is a marker for high-quality care, Monticciolo said, and is achieved by taking a core exam and a certifying exam. Prior to taking the American College of Radiology Core and Certifying Exams, physicians must have 1 year of clinical medical training in a variety of specialties, as well as 4 years of diagnostic radiology training for those who choose this aspect of radiology as their specialty. (Mammography falls under the diagnostic radiology specialty.) Diagnostic radiology training includes a minimum of 3 months focused on breast imaging. Residents must read 240 breast imaging cases within a 6-month period during the last 2 years of their residency program for initial certification.[8]

The core exam, which is computerized and image rich, is taken after the third year of residency. Residents must pass all 18 of the core exam's subspecialty categories and modalities, one of which is breast imaging. Major areas covered in this category are regulations, screening, diagnosis, pathology, imaging findings, interventions, and physics.

Fifteen months after completing their residency, radiologists can take the certifying exam. They must pass all five sections of this exam. The American Board of Radiology requires two of the sections to be on non-interpretive skills and essentials of radiology, while the candidate for certification chooses the remaining three sections from a list of 12 topic areas, one of which is breast imaging.

"It is possible for a radiologist to be board certified without testing in the breast imaging module," Monticciolo noted, although the non-interpretive skills section contains topics pertinent to breast imaging, such as breast screening, recall rates, and radiation safety. "So even if you are not opting for a breast module, you're still going to see breast imaging questions on the certifying exam," Monticciolo stressed.

Radiologists who achieve certification after 2002 must also meet maintenance of certification (MOC) requirements. The proportion of radiologists that must meet MOC requirements will continue to rise over the next few decades as the number of radiologists certified before 2003 decline in practice. The MOC requirements for diagnostic radiology consist of four parts: (1) professional standing, which mainly involves ensuring state licensing; (2) lifelong learning and self-assessment, which are CME requirements; (3) cognitive expertise, which is demonstrated via testing; and (4) participation in practice quality improvement. The exam for cognitive expertise is required every 10 years and includes a non-interpretive standards module as well as three additional modules chosen by the radiologist.

[8] See theabr.org (accessed August 16, 2015).

Radiologic Technologist Training, Certification, and Licensing

Louise Henderson, assistant professor of radiology at the University of North Carolina, reported that the American Registry of Radiologic Technologists (ARRT) is responsible for testing, certifying, and registering radiologic technologists to promote a high standard of patient care. The ARRT awards the Registered Technologist (RT) title, and ensures continued education and ongoing compliance by requiring annual registration of the RT certificate. RT certification is voluntary, but most employers, state licensing agencies, and federal regulators view the ARRT credential as an indicator that the technologist has met recognized national standards for medical imaging, Henderson said.

Educational requirements of the ARRT mammography certification include completing 25 supervised mammography exams and performing an additional 75 exams focused on patient preparation and education as well as the mammographic procedure. As part of their education requirements, technologists must also participate in performance evaluation and recording of quality control tests, and review at least 10 mammogram exams with an MQSA-qualified interpreting physician who evaluates their technique and positioning, and assesses their knowledge of breast anatomy and pathology. The technologist also has to pass an exam that assesses the knowledge and skills typically required of entry-level mammography technologists. Henderson added that there is an ethics requirement that states technologists must "be a person of good moral character and must not have engaged in conduct that is inconsistent with the ARRT rules of ethics."

Once certified by the ARRT, the technologist maintains credentials by renewing annually as well as by taking continuing education credits every 2 years. Technologists also have to meet state law requirements to be licensed to practice within the state. As of May 2015 in the United States, 35 states use the ARRT exam scores in licensing decisions.

Mammography technologists must also meet MQSA-specific requirements, including having a full license to perform radiologic procedures issued by the state or certification from an FDA-approved certifying agency. They must also have at least 40 contact hours of documented training specific to mammography under the supervision of a qualified instructor. This includes training in breast anatomy and physiology, position and compression, quality assurance and quality control techniques, and imaging of patients with breast implants. They must perform at least 25 exams under direct supervision and have at least 8 hours of training in each mammography modality that will be used. Technologists must also have performed at least 200 mammograms in the 24 months prior to a facility's annual MQSA inspection, and have taught or completed at least 15 education units in mammography during the 36 months prior to the facility's annual MQSA inspection.[9]

ASSESSING INTERPRETIVE PERFORMANCE

Audits of Interpretive Performance

Several workshop participants emphasized the importance of ensuring that audit metrics are comparable and described how various performance metrics interact with each other, how confidence intervals and reliability measures can help account for insufficient data, and how patient demographic characteristics and other factors can influence performance metrics.

D'Orsi stressed that "You don't know if interpretation is improving unless you measure it." Carney added that "Audits are really important because they help radiologists understand that there might be a difference between how they think they're doing and how they're actually doing."

But for audits to be meaningful and effective, it is critical that they consistently measure the same things and account for patient demographic characteristics or other factors that might influence interpretive outcomes, several speakers noted. "You can't count apples and oranges and then call them all apples," Wallis said. For example, some screening programs include DCIS in their cancer detection rates or in their percentage of minimal cancers detected rates, while others only count invasive cancers. In addition, some audits lump together diagnostic and screening mammograms when assessing performance while others separate them. Wallis reported that initially the United Kingdom included DCIS in its overall cancer detection rate and thought its rate was comparable to that of other nations' screening mammography programs. It was only after discovering that these countries were not including DCIS in the detection rates that U.K. program leaders realized that their interpretive performance needed to be improved, Wallis noted. Buist called for transparency in audit data and methods to ensure appropriate comparisons are being made across organizations and institutions.

[9] See www.acr.org/~/media/ACR/Documents/Accreditation/Mammography/Forms/PersonnelForms/PersonnelRequirements.pdf (accessed August 16, 2015).

D'Orsi said that the NMD includes both invasive cancers and DCIS in cancer detection rates that are reported back to participating facilities and radiologists. However, the NMD also provides a CDR for invasive cancers only, which is one of the metrics that CMS has established as part of pay-for-performance. This combination also enables facilities to calculate the CDR for DCIS, although the actual calculation is not provided by the NMD.

Several metrics must also be used in audits to acquire a more complete picture of performance, D'Orsi and others stressed. He pointed out that a radiologist can claim 99.9 percent accuracy after reading 3,000 mammograms as negative, even though 7 breast cancers are eventually detected within 1 year of screening. That is because one way of defining accuracy is as the combined measures of true positives (TPs) and true negatives (TNs) over this sum and the addition of false positives (FPs) and false negatives (FNs) (accuracy = TP+TN/TP+TN+FP+FN). So in this case, the radiologist had a TN of 2993, a TP of 0, an FP of 0, and an FN of 7, which gives the accuracy of 99.9 percent (0+2,993/0+2,993+0+7). "This illustrates that it's very important not to look at a single metric in isolation because it will give you a skewed result and it doesn't really mirror what's going on in the real world," D'Orsi said.

As another example, he pointed out that there tends to be an inverse relationship between false-positive and false-negative rates, as well as between sensitivity and specificity. As one rate goes up, the other rate goes down. For example, as the false-positive rate is decreased, specificity tends to go up, but the false-negative rate tends to rise as well. To help make sense of this, D'Orsi presented an analogy in which the false-positive rate can be considered as the "money paid" to detect breast cancer, and the cancer detection rate is "how much stuff—breast cancers detected—that you bought with that money." The minimal cancer detection rate is a measure of "what *kind* of stuff you bought with that money," he added, continuing the analogy. He noted that when evaluating a facility or a person, an audit essentially determines what the person or facility accomplished, in terms of cancers detected, and how much it cost, in terms of false positives and false negatives, and that these are related variables.

A number of different measures can help determine interpretive performance, as defined in Boxes 2 and 3. Although sensitivity and specificity may provide the most accurate information, D'Orsi noted that most facilities do not have access to tumor registries and thus are unable to determine the TN and TP rates needed to calculate sensitivity and specificity. "You really have to be attached to a tumor registry to be able to ask, did this woman who had a negative exam really have cancer in a year or not?" D'Orsi said. Instead facilities can use PPV in their audits, which he claims is a better measure because it determines not just the accuracy of interpretations, but what they cost, in terms of FPs (PPV = TP/TP+FP). But he noted that a high PPV$_1$ rate can stem from a low false-positive rate due to a low recall rate. "The FP money you're spending is too low so you're only detecting low-hanging fruit—the large tumors. That's why the type of malignancy you're detecting is also very important," D'Orsi said. "The higher the PPV number doesn't mean the better because you have to think of what you're getting for what you pay," he added.

Diana Miglioretti, Dean's Professor in Biostatistics, Department of Public Health Sciences at the University of California, Davis, also stressed that "the false-positive rate and sensitivity are correlated and the more you recall, the more cancers you are going to detect." Buist too pointed out the importance of considering several measures when determining interpretive performance because studies show that high performers on one measure are not necessarily high performers on another. "There are some radiologists who have low false-positive rates that also have low sensitivity, but there are others that have high sensitivity and low false-positive rates," she said.

Accounting for Variability in Audit Metrics

Other presenters stressed the importance of considering confidence intervals and other measures of variability when assessing metrics, to account for the variability over time that can occur in interpretive performance. Miglioretti noted that in 1 year's time, a radiologist's performance can vary considerably and that metrics that aggregate data over 3 years tend to be more accurate (Burnside et al., 2014). Wide confidence intervals are often needed due to the small volume of mammograms read by most radiologists and the rarity of breast cancer, she added. "You can't just look at the value by itself. You need to look at the variability of the precision in that value before you classify a radiologist as an adequate or inadequate performer," she said. She added that confidence intervals can be adjusted to account for the variability of the population being screened.

Rebecca Hubbard, associate professor of biostatistics at the University of Pennsylvania School of Medicine, suggested refining audits by considering measures of reliability. She noted that an individual radiologist's variable performance can make it difficult to determine whether he or she is performing adequately because this person's variability may overlap with the variability seen between adequately performing radiologists and those that could improve their performance. "Ideally, in order to identify those radiologists who aren't performing at the standard we would like to see, we would hope that there is enough variation among radiologists that it exceeds the variation of the individual radiologist so we can tell the difference between those who are doing well and those who are doing poorly," she said.

She noted that a large amount of variation in a provider's performance may be due to reading a small volume of mammograms each year. She suggested assessing the accuracy of performance metrics with what she called the ratio reliability, which is defined as the ratio of between-provider variation to total variation, which is the sum of between-provider and within-provider variability. "The more the total variability is explained by differences among radiologists, the more likely we are going to be successful at being able to differentiate between poor performers and good performers," she said. A reliability ratio of 0.9 is generally considered necessary for "high stakes" profiling, such as metrics that will be in the public domain, she said.

Other Factors That Can Affect Reader Metrics

A few participants suggested considering patient demographic characteristics when conducting audits and adjusting them accordingly. Cancer detection rates will vary depending on the age of the population being screened and how frequently they are screened, and other measures will vary according to genetic, ethnic, or sociodemographic characteristics of the population, as noted by Wallis and Onega. "Who your population is should be weighed heavily," Monsees said. Harvey added, "If we're going to make judgments about facilities or physicians, we have to get that granularity in the audits." Pisano noted that some radiologists tend to consider the pretest probability that a given population will develop an aggressive cancer and that influences their interpretation of mammograms. She said African American women tend to have more aggressive cancers so she tends to recommend doing more biopsies on their breast lesions because the consequences of not detecting a breast cancer may be greater for these women than for other ethnicities. "I'm more likely to be more aggressive with these patients, but that doesn't mean it's poor quality to have more false positives in that population if more aggressive advanced tumors are likely to be detected," she said.

Smith also pointed out that both the technique and the judgment of radiologic technologists can influence the accuracy of radiologists' interpretation of a mammogram, with some technologists recalling patients before the radiologist has evaluated their exams. Henderson discussed studies indicating that the technologist had significant effects on radiologists' sensitivity, specificity, recall rates, and cancer detection rates for both film and digital screening mammograms, whereas for diagnostic mammograms, the technologist significantly influenced radiologist interpretive performance for film, but not for digital images (Henderson et al., 2015a,b). Smith suggested considering the influence of the technologist when auditing radiologists. Although D'Orsi agreed that was a valid point, he said in practice it probably would be difficult to disentangle the effects of a technologist on a radiologist's metrics.

Criteria for Adequate Interpretive Performance

One session of the workshop focused on ways to identify radiologists and facilities that might benefit from interventions aimed at improving mammogram interpretive accuracy. This session explored possible criteria and cut-points for low performance, challenges in using those cut-offs for quality assurance purposes, and ways to measure facility versus radiologist interpretive performance.

Carney began this session by noting the significant variability of the interpretive acumen of radiologists in mammography, with their sensitivity varying between 75 and 95 percent and the specificity ranging between 83 and 98.5 percent (IOM, 2005). She noted that a report from AHRQ in 1994 defined 85 percent sensitivity as a desirable goal for radiologists interpreting mammograms, but as she and others pointed out, one criterion is not sufficient to determine quality (Bassett et al., 1994). In addition, cut-points are necessary to identify those needing additional training, she said.

Carney and a group of mammography experts convened by NCI and the ACS developed such cut-points for interpretive performance for both screening and diagnostic mammography using the Angoff method, which is the most commonly used method for determining educational performance standards. It is used to board certify and license practicing physicians (Carney et al., 2010). The cut-points they defined for screening and diagnostic mammography are shown in Tables 3, 4a, and 4b.

The experts then used current BCSC data to determine what percentage of the BCSC radiologists would fall in the low performance range using their criteria. For the screening sensitivity cut-point of less than 75 percent, about 18 percent of the BCSC radiologists fell into the low performance range. Specificity, recall rate, PPV1 and PPV_2, each had an upper cut-point as well as a lower one to ensure high sensitivity was not gained at the expense of a high recall rate that did not productively identify cancers. About half the BCSC radiologists fell into the screening low performance ranges for recall rate, and nearly a third were in the low performance range for PPV1 and PPV_2, Carney noted. The cancer detection cut-point was defined as fewer than 2.5 cancers per 1,000 exams, and 28 percent of the BCSC radiologists fell below this rate.

A simulation using the cut-points determined that if radiologists in the low performance range moved into the acceptable range, an additional 14 breast cancers per 100,000 women screened would be detected, and there would be a reduction in the number of false-positive exams of 880 per 100,000 women screened, Carney reported. Simulation of what would occur in diagnostic mammography indicated that an additional 86 cancers would be detected for every 100,000 women worked up for an abnormal screening result, along with a reduction in the number of false-positive examinations of 1,067 per 100,000 women. For work-up of a breast lump, an additional 335 cancers would be diagnosed per 100,000 women, with a reduction in the number of false-positive examinations of 634 per 100,000 women.

Carney noted that the normative data used to determine cut-points was based on at least 30 cancer interpretations for sensitivity and 1,000 interpretations for the other performance measures, but these numbers may be too

TABLE 3 Final Cut-Points for Screening Mammography Using the Angoff Method

Measure	Low Performance Range	Percentage of the BCSC Radiologists in Low Performance Range
Sensitivity	<75	18.0%
Specificity	<88 or >95	47.7%
Recall rate	<5 or >12	49.1%
PPV_1	<3 or >8	38.4%
PPV_2	<20 or >40	34.0%
Cancer detection rate	<2.5/1,000	28.4%

NOTE: BCSC = Breast Cancer Surveillance Consortium; PPV = positive predictive value.
SOURCES: Carney presentation, May 12, 2015; Carney et al., 2010.

TABLE 4a Final Cut-Points for Diagnostic Mammography to Work Up Prior Abnormal Screening Exams Using the Angoff Method

Measure	Low Performance Range	Percentage of the 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%
Cancer detection rate	<20/1,000	23.2%

NOTE: BCSC = Breast Cancer Surveillance Consortium; PPV = positive predictive value.
SOURCES: Carney presentation, May 12, 2015; Carney et al., 2010.

TABLE 4b Final Cut-Points for Diagnostic Mammography to Work Up a Breast Lump Using the Angoff Method

Measure	Low Performance Range	Percentage 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%
Cancer diagnosis rate	<40/1,000	19.7%

NOTES: BCSC = Breast Cancer Surveillance Consortium; PPV = positive predictive value.
SOURCES: Carney presentation, May 12, 2015; Carney et al., 2010.

small in most radiologists' practices to gather stable performance estimates. Carney also stressed that the single measure of sensitivity used for each radiologist may not discriminate among interpreting physicians because tumor size and type can vary, and many facilities do not have the capability to determine the sensitivity and specificity because they lack access to tumor registry data. Carney also stressed that a major limitation for this analysis of performance cut-points is that the performance measures were examined independently even though they are interrelated.

Miglioretti reported that she worked with several experts to devise cut-points that combine the criteria described by Carney and her colleagues (Miglioretti et al., 2015). For BCSC facilities and a few others connected to tumor registries enabling calculations of sensitivity and specificity, the researchers created joint cut-points for these two measures. This joint analysis specified that for radiologists with a sensitivity greater than 80 percent, a specificity of greater than 85 percent is acceptable. For those with sensitivities between 75 and 79 percent, specificities between 88 and 97 percent were acceptable. The combined sensitivity and specificity criteria cut-points enabled higher false-positive rates (up to 15 percent) for radiologists with the high sensitivity grade of 80 percent or greater. Subsequent simulations showed that 69 percent of BCSC radiologists met these revised combined criteria, as opposed to the 51 percent who met the original sensitivity and specificity criteria when they were not combined (see Table 5).

Recognizing that many facilities are not connected to tumor registries, the authors also developed criteria that combined recall rate, cancer detection rate, and PPV rate for those facilities able to track cancers detected in positive mammogram exams. For these combined criteria, a broader range of recall rates was allowed for radiologists with higher cancer detection rates. The percentage of radiologists who met these combined criteria was 62 percent, compared to 40 percent of radiologists who met the original cut-points prior to the combined analysis.

The experts also developed wide confidence intervals for the cancer detection and recall rates for radiologists with low-volume practices, because as Miglioretti noted, “Large volumes are needed to confidently assess a person’s cancer detection rate. We might need to combine data over multiple years in order to get enough confidence [on] whether a person is performing adequately.” Radiologists whose metrics fell completely within the confidence intervals were classified as having acceptable performance, while those whose metrics fell completely outside the confidence intervals were considered to have inadequate performance. Radiologists whose metrics fell both within and outside the acceptable zones were classified as those with uncertain performance.

With such wide confidence intervals, many radiologists will fall “into the gray zone” in which their true performance is uncertain and cannot be confidently determined, Miglioretti noted. For her analysis, the standard 95 percent confidence intervals were used, but she noted that such a high degree of confidence may not be needed to identify adequately performing

TABLE 5 Combined Minimally Acceptable Performance Criteria for Radiologists with Complete Cancer Capture

Criteria	Sensitivity		Specificity	% of the BCSC Radiologists Who Met Criteria
Original	≥75%		88-95%	51%
Updated criteria 1	≥80%	and	≥85%	62%
Updated criteria 2	75-79%	and	88-97%	7%

NOTE: BCSC = Breast Cancer Surveillance Consortium.
SOURCES: Miglioretti presentation, May 12, 2015; Miglioretti et al., 2015.

radiologists, whereas "you might want to be really confident that someone is inadequate before you tell them to get additional training."

In contrast to Miglioretti's three-part classification scheme using confidence intervals, Hubbard suggested a binary one in which radiologist performance is classified as inadequate if it falls above or below the confidence intervals and all others are classified as adequate. She agreed with Miglioretti that the confidence interval approach is quite flexible and can be tuned depending on the degree of precision required.

Hubbard conducted simulations of her binary approach using representative BCSC data and Medicare claims data and the guideline threshold cancer detection rate of 2.5 per 1,000 women and a recall rate of 12 percent. She found that her binary approach in both types of simulations worked well for the recall rate criteria because recalls are relatively common, but was not as precise for the cancer detection rate, which is based on a rarer outcome. This was true even when wide confidence intervals were used. "For cancer detection rate, the profiling measures using point estimates were working reasonably well but not fabulously. We were definitely making some notable errors," she said. Hubbard added that misclassification of results in the Medicare database also makes it challenging to use claims data to estimate radiologists' performance with greater precision. "When we introduced the additional error due to misclassification of the outcome, the sensitivity [of detecting inadequate performers] was obviously unacceptably low," she said.

Stephen Taplin, deputy associate director of the Healthcare Delivery Research Program at NCI, noted that Hubbard's data indicate that the recall rate is the most precise way to detect radiologists with outlier performance, yet others have indicated that using only one measure will not reveal true performance. Hubbard responded that the recall rate should be considered in context with other information, such as cancer detection rate or sensitivity, which are more difficult to reliably estimate, especially with only 1 year's worth of data. She also stressed that just because recall rate can be estimated well and is more reliable from a statistical standpoint does not mean it is a better tool to measure performance.

Wallis noted that Miglioretti's simulation to determine adequate performers is based on cancer detection rate data from radiologists performing a minimum of nearly 3,000 mammograms. With a 2.4 per 1,000 cut-point for the cancer detection rate, "you are going to have to wait almost 6 years before you can confidently identify an underperformer," he pointed out, which Miglioretti agreed is unacceptable, but might be possible if there was seeding of positive cancers in the exams radiologists read in practices. This concept is discussed further in the section "Seeding Positive Mammograms in Clinical Practice."

Bryan Loy, vice president, Oncology, Laboratory, and Personalized Medicine, at Humana, raised the issue of how to evaluate performance at the facility level, noting that consumers choose the facility, not the specific provider, when deciding where to have their mammography performed. Miglioretti responded that one can ask a facility about its recall rate and cancer detection rate, which they are required by MQSA to document. However, they may not respond to such a request unless enough consumers demand it and threaten to go to another facility unless they receive that information, she said. But a starting point would be to have the facilities themselves paying closer attention to their audit data, compare them to benchmarks, and hopefully decide to improve their metrics if they are not up to standards, she said.

As for the problem of audits and metrics providing incomplete or insufficient performance information, especially for radiologists who interpret low volumes of mammograms, cancer survivor and patient advocate Kelly Walborn pointed out that as a patient, she has had to make life-changing decisions with "an incomplete puzzle of information, so you can rest assured that I as a patient expect the same from you."

Implementing Performance Criteria

Once performance criteria are established, implementation should involve a clear plan of action for those radiologists identified as having inadequate performance metrics, several participants said. Such a pathway will make the criteria more acceptable to radiologists, especially those found to be in the low performance range, Smetherman noted. Carney suggested coaching or mentoring individuals who don't meet performance criteria. Pisano responded that such coaching tends to occur in large academic practices. For example, at her own institution, a radiologist's recommendation for a 6-month follow-up of diagnostic mammography has to be agreed upon by consultation with other radiologists in the practice. "We review every pathology report. Every biopsy you do, you review as a group and you learn from that. There are enough of us that we coach each other," she said.

Wallis pointed out that in the United Kingdom, inadequate performance is remedied with mentoring, and there is a clear procedure and government funding for that mentoring, which is described more fully in the section "Mammography Regulation and Quality Assurance in Other Countries." Nationally funded mentoring is also linked to audits in the Netherlands, said Mireille Broeders, assistant professor of clinical epidemiology at the Dutch Reference Center for Screening, and Ruud Pijnappel, radiologist at University Medical Center Utrecht. They said at least one radiologist from the team that is audited meets with a pair of expert radiologists as part of the audit session to discuss the audit results. At that session, 40 cases with interval cancers are reviewed and they discuss whether the cases should have been recalled. "It's good to have this conversation with peers in order to see if you can improve and do better next time," Pijnappel said. Other members of the radiology team also often attend these discussions because they appreciate the opportunity to learn more, Broeders added.

Buist reported that Group Health Cooperative in Washington state voluntarily used the criteria developed by Miglioretti and her colleagues, and they restructured their mammography program to improve performance "which speaks to its relevance," she said, adding that implementing criteria might vary from organization to organization unless there is a regulatory requirement for it.

INTERVENTIONS FOR IMPROVING INTERPRETIVE PERFORMANCE

Broeders said that radiologists may not seek needed training, citing one study indicating that radiologists do not accurately estimate how well they read mammograms (Cook et al., 2012). In addition, both Broeders and Smith pointed out that most radiologists have few opportunities to receive feedback on their performance. Broeders said this is unfortunate because "feedback can point out areas where they can improve and where training could focus." Two ways radiologists can receive feedback on their performance is by learning the results of medical audits, as previously discussed, or by using test sets.

Test Sets

Smith began the session on test sets by delineating their strengths and limitations compared to those of audits. He noted that audits provide feedback on practice patterns and outcomes and provide the opportunity to measure performance against a gold standard and compared to peers. Audits also can indicate possible corrective actions. But given the infrequency with which radiologists encounter malignant lesions in their mammography reads, audits can take several years to identify with certainty poor or falling performance, or improving performance. In addition, the general outcome measures of audits reveal little about the specific areas that need improvement. For example, a radiologist might do well interpreting mammograms of non-dense breasts and may just need further training in interpreting mammograms of dense breasts. Also some audit outcomes, such as specificity and sensitivity, are not easily measured due to lack of links to cancer registries. The impact of audits on performance also has not yet been fully assessed, Smith noted.

Test sets can overcome some of the limitations of audits by providing radiologists with a large number of normal and abnormal exams that can be read in a single setting or in multiple settings in a much shorter period of time than it would take to encounter them in an actual practice. Many test sets provide immediate feedback and detail how the radiologist's interpretation differs from those of experts. Test sets also provide an opportunity to set reference standards, Broeders and Smith pointed out, and to measure performance against that standard or to assess performance on new imaging technology. In addition, test sets can reveal changes in performance in response to an intervention, and can provide performance feedback based on the appearance of a mammogram, the types of lesions present, or other specific factors not fine-tuned in an audit. Finally, test sets can indicate whether recalls are appropriate and provide sensitivity and specificity values based on true negatives and positives rather than a consensus measure of performance judged by peers.

Monsees reiterated that it is easy for radiologists to obtain their recall rates, even if they operate at low volume, whereas it is more difficult to assess cancer detection rates. Thus, she suggested that radiologists could use their audit data to determine their specificity and rely on test sets more to determine their sensitivity and detection rate.

The ACR developed one of the first breast imaging test sets, called Mammography Interpretive Skills Assessment (MISA). This self-assessment test was not designed to be imposed on radiologists to measure performance, but rather to be a voluntary assessment that could provide tutorial assistance to those interpreting mammograms. The latest version of this test set is digital (the original used film mammograms) and covers various breast imaging platforms, including digital mammography, ultrasound, MRI, and tomosynthesis.[10] Skills assessed in the test set encompass detection, validation, analysis, management, and image quality (see Box 4). Multiple questions are asked for each of the 28 mammography cases in the test set.

The Netherlands also developed and tested a self-test aimed at offering individual feedback to Dutch screening radiologists and to identify areas for further training, Broeders reported. In this test set, a number of questions are asked for each case, including the location of the most suspicious lesion, its laterality and type, and a rating of the confidence of the radiologist's suspicion that the lesion is malignant. The test taker is immediately provided with the results of the test set and can review cases and compare his or her results with those of an expert panel.

Broeders also reported on an international mammography test set called Assessment of International Mammography Screening Skills. The test set is designed to measure proficiency and provide immediate feedback to participating radiologists with the aim of being useful internationally across many different types of settings. The pilot test set is composed of cases of women between the ages of 40 to 79, includes prior images, and has a mixture of difficult, moderate, and easy cancer cases. No interval cancers

[10] See http://www.acr.org/Education/e-Learning/Mammo-Case-Review (accessed July 16, 2015).

BOX 4
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 the quadrant.
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.

SOURCE: Smith presentation, May 12, 2015.

are included, and for the normal cases there is a mixture of difficulty as well as 2-year follow-up data to ensure they are truly normal exams. Teaching points for the test set are currently being developed that include discussion of how appropriate recalls are based on setting. Radiology experts from four countries contributed the cases used in the test set.

Pisano also noted that there are also test sets designed for use in generating data for FDA device review and approval. One study in which she is involved uses a breast imaging test set called DEMOS, which has 50,000 cases, including 365 cancers.

Test Sets for Quality Assurance

Broeders and Smith reported that some test sets were developed to help ensure quality in national screening programs, including the PERFORMS test set that debuted in the United Kingdom in 1991 along with its national mammography screening program, the test set called BREAST, which is used in Australia and New Zealand, a test set used to qualify radiologists to read mammograms in British Columbia's provincial mammography screening program, and a test set developed in Italy.

PERFORMS is designed to be an educational self-assessment and training program for professionals interpreting mammograms. Radiologists working in the United Kingdom's mammography screening program are expected to take PERFORMS or a similar test as part of their quality assurance obligations, Wallis reported. PERFORMS has a challenging set of 60 cases and does not provide prior images for any of them. The test set is compiled from cases submitted by U.K. radiologists, most of whom send in their most difficult cases, Wallis noted. All radiologists who take the PERFORMS test are given feedback based on expert opinion, including which of their recalls were correctly identified as cancers. The feedback is provided directly after taking the test as well as at a later point in time after all radiologists required to take the test have done so. This provides the opportunity for radiologists to learn where they stand in comparison to their peers.

Radiologists whose performance scores are considered outliers are asked to repeat the test set. In addition, medical and personnel directors of screening facilities are informed of the results. Since this system was set up 5 years ago, no radiologist has been an outlier in two consecutive test sets, Wallis noted. Smith raised the question, what is the point of PERFORMS if no one fails it? "Is PERFORMS really necessary for regular readers?" he asked. Wallis responded affirmatively, noting that "it is a good way of seeing, in a short period of time, a reasonable number of cancers, and it reminds you of the sorts of things you have difficulty with." He added that most radiologists treat PERFORMS as a competitive game. "It's a good way of learning as a group. We all enjoy doing it," he said.

BREAST was developed more recently in 2011. It also has 60 cases that do not include prior images, and provides immediate feedback and annotated images so the test taker can see the lesion location he or she has selected compared to that selected by an expert panel.

British Columbia's test set has 120 cases, 40 of which are invasive cancers (15) or DCIS (25). The test set also includes 40 non-cancer abnormal cases, and at least one-third of the cases are of dense breast tissue. This test set is only taken once to qualify radiologists as mammography readers when they begin working in the province's mammography screening program. Subsequently, bimonthly review of all screen-detected cancers and all interval cancers as well as an annual audit review of individual and program data are deemed sufficient to measure continuing performance and to qualify for the program.

A proficiency test set developed by the Italian agency responsible for training radiologists in mammography had 17 cancer cases and 133 normal mammograms, with a reference standard to achieve at least 80 percent sensitivity and a recall rate of equal to or less than 15 percent. Only half of radiologists taking the test had satisfactory performance for both sensitivity and specificity, leading the agency that created the test set to conclude that although there was some degree of correlation between clinical experience and performance on the test set, "experience indicators are not sufficient in themselves to accredit radiologists to read screening mammograms." The agency recommended a proficiency test, such as the test set they devised, as part of the accreditation process for radiologists, Smith reported.

Lessons Learned from Test Set Use

Several participants discussed lessons learned from experience to date with test sets that could inform future test set design, implementation, and use.

The experience with the Dutch test set revealed the importance of having location sensitivity measured, according to Broeders, because for some cases there was a large degree of variability in location interpretation, with some radiologists identifying lesions that were not in close proximity to those identified by the experts, and sometimes even identifying lesions in the opposite breast. Use of the Dutch test also underlined the need to have fixed test dates and locations to avoid technical user problems that occurred when participants downloaded the test cases to their workstations at multiple locations, Broeders reported. Fixed test dates and locations would also prevent possible discussions among those completing the test, which was frequently observed. "Rather than taking it as a self-assessment, they sat together and discussed cases sometimes," Broeders noted.

Smith discussed a study that identified several factors that affect the validity of test sets, including the nature and extent of scrutiny of a test-taker's actions, the artificiality of the environment, the oversimplification of responses, and the prevalence of abnormalities (Soh et al., 2012). These researchers noted that the reading environment in practice differs from the environment in which the test sets are performed, and the implications for correct and incorrect judgment is different. "When you know there are many more cancers, you are more likely to be spooked by these test cases than you would be reading them in clinical practice," Smith noted. There also can be overly simplified judgments in test cases because of a lack of prior images or features of the image that prompt questions that cannot be answered. "That limits your ability to give the answer that you might otherwise have given with additional information," Smith said.

A study of the Dutch test set found that overall it was well received, with 80 percent of radiologists opting to take the test (Timmers et al., 2014). But Broeders said, "You wonder who are the 20 percent who opted not to do the test set. Are these newly trained radiologists who feel a bit shy and not too confident? Or are they the very experienced radiologists who think they don't need to take the test? This is something we would like to learn more about."

Carney noted that a lack of time prevents many radiologists from participating in test set assessments. Boredom is another factor, she added. "They get bored because it's very tedious to sit and read 110 cases in our final test set," she said. D'Orsi responded that another reason for a lack of participation in self-assessments is because "people don't want to find out if they're not good. They want to keep that idea that they read well and they're afraid that when they do a study that is relatively objective, they're going to find out they're not good." But rather than make radiologists feel inadequate if they do not perform well in test sets, he suggested affirming that they are doing "pretty well but we can help you do better," he said.

Correlation with Clinical Performance

There is mixed evidence on whether performance on test sets correlate with clinical performance. A study cited by Smith found that clinical performance was significantly better than reader performance in a laboratory setting (Gur et al., 2008). In contrast, Smith noted an Australian study on BREAST indicated that test set performance correlated well with clinical performance, particularly on measures of sensitivity, recall rate, and detection rate of small cancers, but did not correlate well with specificity, probably due to differences between the settings of clinical practice and testing and because the radiologists were informed the test set was embedded with cancers (Soh et al., 2015). Broeders cited one study indicating no correlation between clinical performance and test accuracy, as well as a more recent study that found the PERFORMS assessment broadly reflected clinical performance, with moderate correlations in the positive direction (Rutter and Taplin, 2000; Scott et al., 2009).

Miglioretti commented that a lack of correlation between test set performance and clinical practice might be because "we are taking a very poorly estimated measure of clinical practice based on the detection of only 5 to 10 cancers and then trying to correlate it with a test set with many more cancers."

Challenges in Test Set Design

Smith noted a number of challenges in developing test sets, including uncertainty about the optimal number of cases to include and the optimal ratio of normal and cancer cases, as well as the appropriate mix of difficulty for the cases. In addition, there is also debate about whether the positive exams should be confirmed with biopsy results or determined by expert consensus. Other unanswered questions about test sets include who should be tested and how often, how often the images should be refreshed to avoid recognition of images seen previously, and how performance on test sets should be evaluated to improve test set composition.

Broeders added that if we want test sets to better resemble the clinical screening situation, they should include prior mammograms. Test sets could also be designed to improve agreement on certain lesion types and could be developed for technologists as well as radiologists, she noted. Buist suggested test sets might be more beneficial if, like many computer games, they adjust in real time the level of difficulty of questions and cases based on how well radiologists are performing as they make their way through the test cases, a testing approach known as computerized adaptive testing. Walborn concurred, noting that the current generation of physicians has grown up with the computerized gaming culture and are used to simulations, in which they compete against themselves or their peers and have challenges ramped up according to their performance.

Implementing Test Sets

Several participants stressed the need for more clarity on how test sets are best implemented. For example, should they be voluntary or mandatory? Monsees and Smith suggested it might be best to start with test sets being voluntary and perhaps mandating them later. Monsees noted that mammography accreditation by ACR was initially voluntary, and later became mandatory under MQSA. She pointed out that there was substantial voluntary participation prior to the mandate. Smith added that “test sets can be enormously beneficial, but we need to do considerably more research and development on test sets, including identifying ranges for specialized test sets, best strategies for communicating results in a standardized fashion, and we need to validate the value of test sets.”

Assessing and Improving Mammography (AIM) Intervention

Geller reported on the AIM intervention, which is a large-scale, international collaboration that applied the latest educational theories and findings in the design and testing of a CME program for radiologists specializing in mammography. For example, the CME was designed to be interactive, relevant, and based on a needs assessment, and with an active, self-directed approach to learning. Radiologists in one arm of the study received skills training via a DVD that was self-paced, could be completed over several sessions, and could be done at home or at the office. The DVD provided immediate feedback on whether their interpretations were correct, as well as extra cases for additional practice. In the other arm of the study, radiologists received a live seminar in which experts tailored teaching points to the specific questions of the participants and gave them immediate feedback on any improvement noted during the seminar. With the live seminar, there was also immediate collation of results so radiologists could see how their responses to questions compared to others. Both the DVD and the live seminar had the same teaching cases.

The researchers compared the individual performance on test sets before and after the teaching program and found that although more radiologists made positive changes after the live seminar compared to the DVD arm, the magnitude of those changes was not statistically significant. When the DVD intervention arm was compared to a control group, sensitivity and PPV significantly improved, but not specificity. When the live seminar group was compared to a control group, there were no significant differences except for a decrease in specificity. Radiologists in the live seminar group more frequently reported an intention to change their clinical practice as a result of the intervention compared to the DVD group. The majority of participants in both study groups believed the interventions were a useful way to receive CME mammography credits (Carney et al., 2013).

MAMMOGRAPHY REGULATION AND QUALITY ASSURANCE IN OTHER COUNTRIES

Nationally funded, population-based mammography screening programs are common in Europe, and some are more highly regulated than mammography in the United States. Wallis gave a detailed presentation on the United Kingdom's national mammography screening program, noting that it has a centralized process for quality assurance, and radiologists in the program are regulated by the General Medical Council, which has statutory duties requiring reporting of suboptimal performance. Each region in the United Kingdom has its own quality assurance system that must meet age-based standards established nationally at the onset of the program. National and regional data are collected each year and auditing is done regularly at least every 3 years to ensure standards are met. "We measured how we were doing against the standard and we kept raising the bar because we got better," Wallis noted. Oversight in the U.K. mammography program also entails an element of public health evaluation.

In 1998, after Wallis and his colleagues showed that practice reading mammograms was the major determinate of performance for cancer detection and recall rates, the United Kingdom instituted double reading in its program. This team approach substantially improved the cancer detection rate, with a modest increase in the recall rate (Blanks et al., 1998).

Wallis stressed the importance in the U.K. system of radiology professional representatives who provide peer support, are trained as quality assurance teams, and are supported with national funds. From the data collected, underperforming teams are identified and analyzed to make sure they are true underperformers and not radiologists whose volumes were insufficient to reliably assess their performance. The analysis also considers if the underperformance is clinically important. If there is a serious problem, the quality assurance team will provide a mentoring service with feedback. Such mentoring involves meeting with the local team, analyzing their practice, discussing some of their cases, and indicating possible interventions for improvement (see Table 6).

For example, if readers with low performance scores have high recall rates combined with moderate to high cancer detection rates, their recalls are reviewed with a mentor, whereas those with low recall and low cancer detection rates are encouraged to increase their recalls. Those with high recall rates combined with low cancer detection rates are given more intensive mentoring, which might include reviewing all their interpretations, and having them retrain with PERFORMS. Mentoring typically involves twice-weekly sessions for 2 or 3 weeks, during which cases are reviewed and discussed.

First reader performance analyses based on 3-year data and at least 3,000 mammograms are also conducted and graphically depicted with recall rates on one axis and cancer detection rates on the other so that individual radiologists can easily see their reading style and what aspects of their interpretations need improvement (see Figure 5). Harder to comprehend statisti-

TABLE 6 Analysis of First Reader Performance in the U.K. National Breast Screening Program

Low Recall with High CDR	High Recall with High CDR
Possible actions • No actions needed • Consider whether there are any possible learning points from their film reading method	Possible actions • Review false positive recalls
Low Recall with Low CDR	**High Recall with Low CDR**
Possible action • Increase recall rate? • Avoid other similar readers • Do not arbitrate alone • Review missed cancers	Possible actions • Review missed cancers • Review false positive recalls • Potential training issue

NOTE: CDR = cancer detection rate.
SOURCE: Wallis presentation, May 13, 2015.

cal measures such as sensitivity and specificity are not used because "most radiologists fell asleep in statistics classes," Wallis said. The analysis of cancer detection and recall rates can suggest radiologists who read similarly and should possibly not be paired together for double readings. For example, a radiologist with low cancer detection and recall rates may be paired with one with a higher rate so that together they interpret closer to the mean. "We've demonstrated that this works," Wallis said.

Wallis also stressed that an advantage of the United Kingdom's quality assurance system is that it offers radiologists specific pathways for improving their performance and mentoring activities to support the changes needed. "We don't just hand bits of information back and then ignore them," he said. More than 95 percent of the time, radiologists can access women's prior exams, he said.

Broeders noted that the Dutch mammography screening program also provides national funding for quality assurance. Radiologists do not necessarily read their own recalls and they are provided with prior exams when interpreting screening mammograms.

When making comparisons between the quality assurance programs of different countries, it is helpful to note that some programs take the "carrot approach" and try to encourage underperformers to improve their performance, while others use more of the "stick approach" and try to eliminate

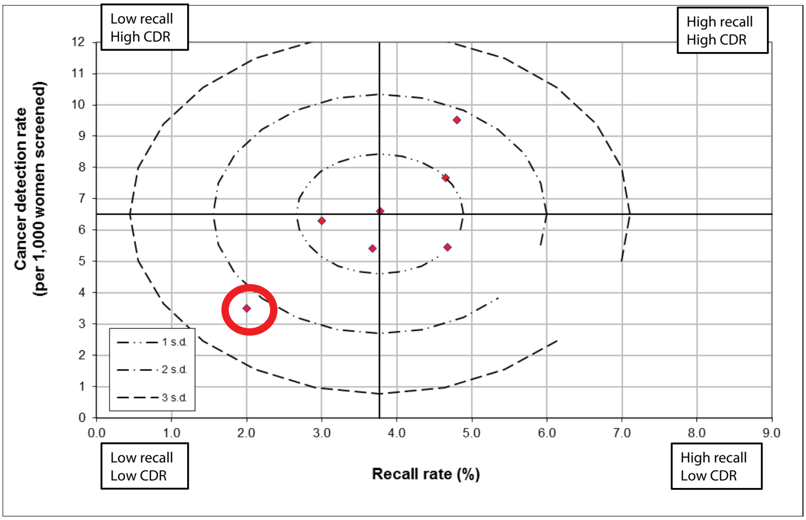

FIGURE 5 Analysis of first reader performance in the U.K. National Breast Screening Program.
NOTE: CDR = cancer detection rate.
SOURCES: Wallis presentation, May 13, 2015; West Midlands Quality Assurance Reference Centre, Public Health England.

underperformers, Wallis stressed. He said the traditional stick approach to quality assurance is based on the assumption that there is a bell curve to performance, and quality can be assured by removing the small numbers of outliers at the end of this bell curve. However, experience with mammography interpretive performance in the U.K. program indicates that performance does not conform to a bell shape, but rather "there's a long tail of underperformance at [one end], so if you can move up the performance of the ones at the bottom, the whole system moves up a lot better," Wallis said. "It's a repeating cyclical process because if you always keep knocking off the bottom by making them better, you move the whole thing sequentially along each time around. So you've got to be good at identifying the low performers."

RESEARCH NEEDS

Many participants discussed research needs to better understand how best to improve mammogram interpretation, including research on various interventions, quality criteria, educational methods, and best practices and organizational strategies. D'Orsi stressed that increased federal regulation requires good evidence from research studies—data that show various interventions can improve performance. From a payer's standpoint, Loy suggested that research was needed to better define quality criteria and to develop more universally accepted standards. He stressed that inadequate performance cannot be ascertained from claims data, even with the advent of electronic medical records. "As a payer, [I would like] to be able to say to folks, 'this is the measuring stick and we know all these metrics work in concert and we can demonstrate improvement,'" he said. Without such metrics, mammography will be evaluated as a commodity. "If we've got a commodity, then the conversation will not be about who's the best, but who's the cheapest and we'll find ourselves paying for the cheapest," Loy emphasized.

Smith suggested assessing how long performance improvements last following interventions, such as test sets and mentoring. "Do people fall back into their same old reading patterns if they don't have the support of ongoing audits, reviews, and support? What is required to maintain that improved performance?" he asked.

Geller suggested conducting research to determine the best ways to train physicians. Carney added that some evidence suggests that a reinforcing element, sustained over time, is needed for training to be effective in the long term. She stressed that "there are qualitative aspects of understanding training and behavior changes that we have not touched. I would love to be a fly on the wall when mentoring is done to see exactly what the key components are that really produce change. We haven't gone there." Carney also suggested that researchers could evaluate whether changing residency training to provide people with more flexibility to intensify their breast imaging helps improve proficiency. "I want to know if these mini-fellowships, these areas of concentration make a difference," she said. "This educational research . . . matters to the payers, the professions, and the public. The educational sciences have been growing in leaps and bounds and medicine needs to keep up with it." Buist added, "We need to understand what works best for life-long learning." She added that it would be important to evaluate the impact of Centers of Excellence on mammogram interpretation and patient outcomes.

Finally, several participants suggested doing global studies to identify international best practices that promote quality mammogram interpretations.

POTENTIAL OPTIONS FOR ADDRESSING MAMMOGRAPHY QUALITY CHALLENGES

Presenters and other participants suggested many possible ways to address mammography quality challenges in the areas of education and training, peer support and mentoring, auditing and self-assessments, financial incentives, government regulations, and communications with consumers.

Training

Wallis estimated that it takes at least 9 months to get a radiology Fellow in training to the point where that they can accurately interpret mammograms on their own without consulting with someone more experienced. "There's a difference between confidence and competence and it all takes time. Training is really important," he said. The 2005 IOM study also recognized that residency training alone is not sufficient to develop good interpretive skills in mammography, Geller noted (IOM, 2005). Carney pointed out that radiology residency training is designed to be broadly comprehensive and is not tailored to the type of practice residents expect to have. She suggested that those who know they want to focus on mammography should have the flexibility to receive more training in this area. She also noted two studies showing that fellowship training could definitely improve the interpretive performance of radiologists in mammography (Elmore et al., 2009; Miglioretti et al., 2009).

Smith expressed concern that a survey undertaken several years ago found that radiology residents did not perceive mammography as an engaging or exciting specialty (Bassett et al., 2003). To encourage more radiologist residents to specialize in mammography, he stressed the need for a supportive environment for learning how to read mammograms. He also said it was important to make sure that both technologists and radiologists in mammography do not feel isolated in their practices. Smith suggested "investing in a new generation of readers" who are encouraged to specialize in mammography by targeting promotion of this career path to ACR's Young and Early Career Physician Sections and its Residents and Fellows section. The Society of Breast Imaging could also create a subgroup for young breast imagers "and provide care and feeding to that group at regular meetings," he said. Monsees concurred, noting that the Society of Breast

Imaging is primarily an educational organization that would naturally gravitate toward helping its younger and less experienced members.

D'Orsi added that mammography is completely different from other forms of biomedical imaging that radiologists routinely interpret because it is not focused on anatomy so much as discerning signs of a cancer from background signals. "It's a signal-to-noise problem and we haven't focused on providing that kind of education, so it is really alien to a lot of people who have only 3 months of mammography in their training after having 3.5 years in anatomy-based imaging," he said. Smith suggested developing a special breast imaging curriculum for new radiologists entering the mammography workforce, with emphasis on the value of tracking their performance and participating in the NMD. He also noted that studies suggest it might be useful to structure training programs so that new radiology residents can work up their own recalls and receive structured feedback on them.

Taplin suggested developing enhanced training for mammography technologists and perhaps being more rigorous about how proper positioning is measured and evaluated. Pisano noted that the ACR has defined measures for image quality that depend on positioning. These measures include the amount of pectoral muscle in the image, whether there is any motion, and whether the entire breast is included. Barr added that most of the mammography facilities that do not meet the requirements of an FDA inspection fail because of inappropriate positioning. "That's the number one problem when clinical images are reviewed for facilities that fall into what we call serious risk to human health—it's almost always related to positioning," she said, and suggested working with the American Society of Radiologic Technologists or the ARRT to ensure positioning in both training and in practice over time. Proper position technique could potentially be a requirement for maintaining certification, she noted.

Mentoring and Coaching

Carney suggested more coaching of physicians during training as well as in clinical practice. She noted that her institution had reorganized so that every medical student has a coach who helps him or her throughout the medical school experience. "Tiger Woods and Michael Jordan have coaches, so why don't we have coaches in medicine? We've got to get over this idea that the way you train physicians is to have them spend their entire time in their own heads," she said.

Carney also suggested engaging radiologists identified as high performers as mentors for facilities. She noted that this approach was used by the Northern New England Cardiovascular Group. They identified which surgeons in the group had the best outcomes and sent them to mentor other centers whose performance was not as high. The result was improved outcomes for these centers. But Harvey noted that "physicians are fiercely independent, and that independence is part of our training," so they may be reluctant to work in teams or to be coached.

Reading One's Own Recalls

Buist noted that only about 40 percent of radiologists work up more than 50 of their own recalls annually. However, a study she conducted found that radiologists who work up a large number of their own recalls tend to have better performance metrics, such as sensitivity and cancer detection rates, than those who work up fewer than 25 per year (Buist et al., 2014). She suggested requiring radiologists to do the diagnostic work-ups for a minimum percentage of their own recalls.

Buist pointed out that reading one's own recalls should be feasible for most practices because the use of digital imaging enables radiologists to read and interpret diagnostic images remotely for their recalled patients. However, a survey by Smetherman of members of the Society for Breast Imaging found that only about 17 percent of the responding radiologists who read mammograms reported they did any remote reading of diagnostic mammograms.

Barke noted that although it might be feasible to read one's own recalls, for quality purposes it is often better if someone with expertise in the lesion in question does the work-up of the recall. Harvey agreed, but noted that in her practice, all radiologists review their own recalls even if they do not do the actual diagnostic interpretation.

Wallis pointed out that during fellowship training in the United Kingdom, radiologists do the work-ups for their own recalls. In addition, potential recalls are often discussed with colleagues in the practice. "If you can ask a friend on a quiz show, why can't that work in radiology?" he asked. He suggested that if an individual has a very high recall rate, the practice could require that his or her recalls be discussed twice weekly with an expert in the practice. "This would be something relatively easy to implement if you can persuade the good readers—low-recall readers with reasonable sensitivity—to do that," he said.

Double Reading

Wallis said the success of the U.K. national mammography program is partly due to having all mammograms double read, and several participants suggested implementing double reading at U.S. mammography facilities. However, for economic reasons, double reading is rarely done now in the United States, and some participants considered this approach infeasible. But Smith suggested double readings could be required for a short period of time until documentation shows that radiologists are performing adequately, at which time they could start doing single reads, perhaps using CAD as a sort of second reader once they are more experienced. D'Orsi agreed this would be a useful approach, but he emphasized that it is dependent on accurately measuring interpretive performance, "which we still have issues with here in the United States compared to the United Kingdom," he said. It would require setting up cut-points to determine who would be the trainee and who would be the trainer, as well as cut-points for identifying when the trainee has moved into the proficient range.

Barke noted that double reading takes more time and "patients want answers rather quickly." A way to streamline double reading so it can be done expeditiously would be required in this country, she said. Monsees added that the workforce is currently not robust enough to support double readings in this country, although it is done informally in many practices for difficult cases. Radiologists in solo practices, however, do not have this option for double reads.

Smith noted that technologists with extensive training and experience could be used as double readers. Henderson concurred, reporting that mammography technologists have functioned as prereaders or double readers in several screening studies, most of which were done in Europe, where technologists received special training on how to interpret mammograms. These studies found that use of the trained technologists as prereaders or double readers increased cancer detection rates without significantly increasing recall or false-positive rates (Bassett et al., 1995; Haiart and Henderson, 1990; Pauli et al., 1996; van den Biggelaar et al., 2009; Wivell et al., 2003). However, Monsees noted that it would be hard under MQSA to engage technologists or other non-radiologists in image interpretation because MQSA specifies one has to be a physician to read mammograms.

Volume Requirements

Buist noted that volume requirements (the number of mammograms radiologists are required to interpret per year) vary from country to country, and a consistent relationship between volume and recall or cancer detection rates has not been observed across those programs, perhaps because so many other variables differ among programs. However, a study she did using U.S. data from the BCSC (Buist et al., 2011) showed the following correlations:

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

Buist said these data suggest that "we should consider increasing minimum interpretive volume in the United States and also include a minimum diagnostic interpretation requirement." Smith agreed, noting that new volume requirements could be addressed with amendments to MQSA regulations. However, Monsees noted that the National Mammography Quality Assurance Advisory Committee (NMQAAC) developed the initial volume requirements under MQSA and purposely made them on the low side so as not to exclude too many radiology practices that would not be able to meet them. Presumably, the NMQAAC could devise new volume requirements that could be incorporated into MQSA regulations if there was enough evidence to support those new requirements, she said.

Buist added that about $1.6 billion in annual screening costs are due to false positives (not including the time, travel, and anxiety costs), and increasing the mammography volume requirements from the current 480 screens per year to 1,000 screens per year would decrease the costs associated with false positives by about $59 million per year.

Onega said a study she conducted found that facilities with the highest volume of mammograms are significantly more likely to detect tumors with a good prognosis than those with low volumes (Onega et al., 2015). She also suggested teleradiology might help alleviate the disparities in access to facilities with sufficient volumes or new technologies. Such remote interpretation of images enabled by the advent of digital technologies decouples the reading of the image from the acquisition of the image. This new "force in radiology" may help readers maintain minimum volume thresholds and quality, Onega pointed out. Teleradiology might also help mitigate variations in the use of advanced imaging technologies as supplemental imaging techniques slowly diffuse into practice, she said.

Enhanced Audits

Smith reiterated the call for a more meaningful MQSA-mandated audit, by separating screening from diagnostic exams, for example. Barr stressed that although more specific metrics could be required as part of the MQSA-mandated medical audit, "how that audit is used in the facility and how it produces increased quality is something out of our control." In addition, she noted that currently FDA only inspects to see if volume requirements are met by facilities, and volume statistics are not collected for individual physicians.

Several participants noted that audits could be improved if more mammography facilities were linked to tumor registries to assess important metrics such as true positives and negatives, sensitivity and specificity. "If we want to be able to give a radiologist feedback on every critical measure, we have to have links to registries," Smith said. "Why can't we have the same access to data as our colleagues in Europe do, where it makes it so much easier for them to measure performance?" he asked. Wallis noted that the United Kingdom had to create "Cancer Intelligence Units" to enable cancer registries to provide data in a more timely fashion. "In the old days, the cancer registry was a repository for information that never left and there's no point in having a library when you can't borrow the book," he said.

Auditing also could be improved by ensuring long-term and adequate funding for the NMD and the BCSC, Smith suggested. "It's an absolute shame that we do not have these health services registries set up as enduring registries and that they have to be treated as ongoing grant programs," he said. Agreeing with Smith, who called the BCSC a national asset, Monsees added, "If the NMD or the BCSC are national treasures then we have to find federal funding for them." Monsees suggested considering merging the two databases and exploring whether it would be better to support the National Mammography Database and link it to tumor registries in all 50 states, or instead link each mammography facility to a tumor registry. Both options would require a significant amount of funding, she noted.

Geller suggested that professional societies and payers could encourage more facilities to join the NMD. She noted that having facilities link to tumor registries can be difficult because many states have laws, in addition to federal regulations (e.g., the HIPAA Privacy Rule), that may restrict sharing of identifiable patient information. Instead she suggested expanding the databases of the BCSC and including more data on cancer detection and recall rates.

Miglioretti added that the BCSC and the NMD could collaborate to understand limitations in metrics such as cancer detection rate because the capture of biopsy information varies so much by facility. She suggested devising measures that take into account biopsy capture rates and the bias that a lack of capture will have on cancer detection rates and other measures, noting a site might appear to have a very low cancer detection rate because it is not capturing a lot of the biopsy results that follow a positive exam. "There are opportunities for collaboration between the BCSC and NMD and understanding the strengths and limitations due to not having the cancer linkage," Miglioretti said.

Loy suggested a collaboration between ACR and the College of American Pathologists to link biopsy pathology reports to BI-RADS classifications. Pisano responded that although there is no official collaboration between the two organizations, "almost all really strong breast imaging centers have a strong collaboration with pathology because otherwise you can't really correlate and decide whether a core biopsy needs to be repeated or an open biopsy needs to be done. So that's happening in the trenches." She added that recent data indicate that expert pathologists have as much as a 15 percent disagreement rate on whether lesions should be classified as DCIS versus invasive tumors or atypical hyperplasia (Elmore et al., 2015). "Pathologists also have too much variability in their interpretive performance, which is a whole other problem we have," she said. Smetherman noted that although it is still in its infancy, the National Accreditation Program for Breast Centers[11] is starting to define metrics that are multispecialty. They are developing a Breast Cancer Quality Improvement Program that will be similar to the Colonoscopy Quality Improvement Program.

Test Sets

Several participants suggested greater use of test sets for self-assessment and quality improvement. As Smetherman noted, "No radiologist is doing a suboptimal job by choice, but rather because either they don't realize they are doing a suboptimal job or they realize that but don't know how to

[11] See https://www.facs.org/quality-programs/napbc (accessed July 21, 2015).

improve." Carney noted that a study she did found that radiologists who accepted an invitation to participate in a self-assessment test set and subsequent CME activity were more likely to be in an earlier stage of practice, with less than 10 years of experience. "There is a perception that the older folks who have been in practice a long time don't need to do these activities anymore," Carney said. But this may not be the case because when she did a test set assessment as part of a CME activity with breast pathologists, she found the biggest predictor of lower performance on the test set was advancing age.

Ganz added that, "Physicians are pretty stubborn. We think the way we know how to do it is the best way. But if we practice as a group and feel confident about sharing our uncertainties, we are going to become more proficient." But she added that it may be challenging for radiologists to accept that they have a low performance measure and to work toward performance improvement because mammography may be a small part of their practice and they may not want to invest the time.

Smith called for creating a culture of evaluation in radiology and other medical specialties so that it is considered routine to assess one's performance, and a clear pathway for improvement is offered, if needed. Smetherman agreed, noting that it is a failure of the current educational system that there is an assumption that once radiologists are board certified, they are performing adequately. Carney noted that radiology could learn from the American College of Cardiology, which has started to test cardiologists on actual angiograms that are uploaded onto a website for maintenance of certification. "Now they have to show how good they are doing, which means they should know how good they are doing. I am hopeful that we are also headed in a direction where there is that culture of evaluation of loading something up and sharing how good you are, having new indexes calculated in a meaningful way that will be much more accurate than taking a certification exam," she said.

Monsees added that test sets are a good way for radiologists to see how accurately they are able to detect cancer and should be used more to assess performance. Audits can also provide useful feedback and D'Orsi suggested that audits be linked to mentoring if suboptimal performance is detected. Smith added that audits should present data in cluster graphs or provide other visuals so it is easy to assess one's performance in relation to others and to see what areas need improvement. Both he and Wallis noted that audit results that just provide a summary of statistics are not as helpful to radiologists.

Several participants suggested that self-improvement audits, and test set assessments should be encouraged if not mandatory. To incentivize such participation, Smith suggested defining safe harbor provisions for those who participate in continuous quality improvement, including adhering to audits, regular reviews, and proficiency testing. He said such a safe harbor from medical malpractice would be similar to that provided for manufacturers of childhood vaccines by the National Childhood Vaccine Injury Act of 1986, which led to the creation of the National Vaccine Injury Compensation Program.[12] If a similar program for mammograms was enacted, the federal government could still compensate women who had cancers missed in their mammograms, but the radiologists practicing in safe harbor would not be sued for malpractice, he said. Barke seconded the idea that such a "carrot" for self-assessment be offered to radiologists. Monsees also suggested that the American Board of Radiology could specify test sets and other quality improvement and self-assessment projects as part of MOC for radiologists.

Seeding Positive Mammograms in Clinical Practice

Some participants suggested that seeding a radiologist's clinical caseload with images of confirmed cancers could enable radiologists to more quickly gain expertise in identifying cancers and hopefully improve their cancer detection rates. Miglioretti pointed out that the Transportation Security Administration seeds luggage screening with computer-generated images of guns to improve screeners' detection rates for weapons. Smith added that just as one is likely to "overcall" cancers in a test environment where one knows the number of positive mammograms has been increased, one is also likely to "undercall" cancers in clinical settings in which breast cancers are relatively rare. But he also pointed out that it is not yet known what the prevalence of the abnormal results should be to ensure a highly accurate process of interpretations. Geller also cautioned that such seeding might create new challenges and questions that need to be answered. "It's another road we need to further explore before we decide which turn to take," she said.

Seeding could also enable more accurate assessments of sensitivity, Miglioretti and Carney noted. Smith pointed out a study showing that when positive mammograms were seeded in the clinical setting, participants missed 30 percent of the cancers, but when interpreting the same images in a high prevalence setting (50 percent positive), participants missed just 12

[12] See http://www.hrsa.gov/vaccinecompensation/index.html (accessed July 21, 2015).

percent of the same cancers (Evans et al., 2013). This led the researchers to suggest that a method of seeding clinical practice with cases for which "gold standard" truth is known could provide a relatively unobtrusive mode for individual or group assessment and could address some of the variability in estimates of rates of missed cancer.

Performance-Based Incentives

Smith suggested that breast imaging centers should facilitate the provision of feedback on performance to radiologists, adopt new strategies to improve performance, and reward quality performance. Referring to the RVU (relative value unit) method of determining the value and financial compensation for various physician activities, Monsees added, "The quality of how we perform is important to reward. We live in a very RVU-based society so it's important to figure how the quality and performance metrics fit into that RVU system." Smetherman suggested rewarding quality performance with "bragging rights, awards, and recognitions that mean something to the community, not only to professional organizations. Make benchmarks matter when we have conversations out in the community."

Loy suggested payment incentives for quality performance and noted that payers are increasingly moving away from fee-for-service payment models and instead using bundled payments, with a set number of dollars for each patient's episode of care. This population-based payment method can incorporate value based on quality and offers financial incentives for high performers or evidence of improvement in the quality of care. Susan Dentzer, senior health policy advisor at the Robert Wood Johnson Foundation, noted that the Centers for Medicare & Medicaid Services (CMS) has initiated bundled-payment care initiatives. Smetherman added that ACR has developed some bundled-payment proposals that could also include payment incentives for quality performance or a structure that would encourage quality care. D'Orsi suggested that CMS could set the standard that others are likely to follow by defining and rewarding quality care for Medicare patients.

But some participants also added a note of caution about restructuring mammography practices and how they are reimbursed. "As a payer, we're agreeing that more integration is probably better, and double reads, for example, would probably be desirable, but that doesn't come without a price and somebody has to finance that," Loy said. "We have to figure out what it's worth in terms of investing in reengineering a practice and the value we are getting back in terms of quality, patient convenience, and satisfaction. We have to be reasonably sure that we're going to improve quality or that quality is not going to suffer because of a new payment module," Loy added.

Ganz suggested payers could provide higher reimbursement to those who participate in the NMD or the BCSC and are regularly assessing their quality metrics. Pamela Wilcox, assistant executive director, American College of Radiology, noted that part of the CMS pay-for-performance initiatives recognize qualified clinical data registries and that eight measures within the NMD enable facilities to meet the reporting requirements needed to receive a bonus payment. In addition, CMS is convening a working group with other payers to ensure that measures across payers are consistent so physicians do not have to report different metrics for different payers. "It's going in that direction and if people don't want to get a penalty, they're going to have to report data," she said.

Prior Exams and Supplemental Images

Several presenters noted the importance of having access to prior mammograms to properly interpret a current exam. Loy suggested breaking down any existing barriers that impede access to prior exams. "Let's integrate information that we already have or should have at our disposal so we can make use of it," he said. Supporting this suggestion, Kathryn Pearson-Peyton, chief medical officer, Mammosphere, noted a study by Sickles indicating that when prior mammograms are absent, recall rates are 260 percent higher. She added that a study by Burnside also showed that when prior mammograms are available in the screening setting, the breast cancers detected are more likely to be at an earlier stage, before any spread to lymph nodes (Burnside et al., 2002). Pearson-Peyton said she is working to form a nonprofit organization to network all MQSA-certified facilities so that mammograms can easily be shared among them. The goal is to use cloud storage to create a mammography image database that can be accessed through the Internet.

Several workshop participants also noted that supplemental imaging, such as ultrasound or MRI, might also aid interpretation of difficult cases, such as dense breasts, which can obscure tumors in mammograms. Monsees pointed out that MQSA regulations require providers to send women a lay summary of the results of their exam, but does not require reporting of breast density status. However, 22 of 50 states have passed laws that women need to be notified if their mammograms reveal they have dense breast tissue, with some indicating ultrasound or other supplemental screening. A few states also have laws that mandate coverage for supplemental screening. However, she said it's not yet clear which women are most likely to benefit from supplemental imaging.

She stressed that ultrasound images can sometimes help improve interpretation of mammograms of dense breasts. "Some things are much easier to see on ultrasound," she said. Studies done in Connecticut, which mandates supplemental ultrasound for women with dense breasts, found that ultrasound detects additional cancers compared to mammography alone (Hooley et al., 2012; Parris et al., 2013; Weigert and Steenbergen, 2012). However, the U.S. Preventive Services Task Force concluded that there is insufficient evidence that ultrasound images can improve breast cancer screening (USPSTF, 2009). In addition, a trial conducted by the American College of Radiology Imaging Network showed that adding ultrasound to mammography increased the number of cancers found, but also generated more unnecessary biopsies. A very small percentage of those biopsies detected cancer compared to biopsies done following mammography alone (Berg et al., 2008).

Monsees said that supplemental MRI is better at detecting additional cancers in dense breasts than supplemental ultrasound, citing a study showing that about 15 more cancers were detected per 1,000 women by MRI in women who had already been screened by mammography and ultrasound (Berg et al., 2012). She added that a new technology known as fast MRI can enable a breast exam in just 3 minutes. One study found that fast MRI detected more breast cancers than standard mammography (Kuhl et al., 2014).

Monsees concluded that there are limits to standard mammography and that supplemental imaging with ultrasound, MRI, or other complementary technologies could improve cancer detection, but she emphasized again that it is not yet known who should have supplemental imaging and with what types of technologies. She added that the new breast imaging modalities are a mixed blessing because although they can help work up recalls and detect more cancers, they add costs and take more time to perform with a workforce already short on time. "It gives a better work product and better answers, but it is definitely more time consuming," she said.

Communication with Consumers

Several participants suggested informing women who undergo mammography about quality metrics related to interpretation. Informed women would lead to more consumer demand for better quality, Pisano said. "We need health care consumers to start asking about outcome measures. That way there will be more interest among health care providers in reaching a standard. The consumer has to demand quality," she said. Wallis agreed, saying, "The only way things will improve is when consumers ask for performance information." He added that in the United Kingdom, performance data for surgical groups are published nationally in newspapers.

But Monsees expressed reservations about whether consumers will be asking the right questions, and whether outcomes data on the Internet would be unacceptable to many physicians and might lead them to leave the profession. Loy noted that consensus is still lacking on what the criteria are for adequate or better than adequate quality, so it is not clear what the consumer could demand. "I don't think even I know what 'bad' is in order to tell someone to go somewhere else," he said. Monsees added that consumers tend to be more concerned about quality when it comes to diagnosing suspicious lesions than in screening. She noted that women will often seek out an academic facility for a second opinion if their community facility finds something of concern on their mammogram. "People need to look for where there is good quality from the beginning. They do know there is a difference, but they don't think it matters for screening unless they have something wrong with them," she said.

Walborn stressed that communication to consumers should use language that is understandable and is free of complicated statistical concepts and terminology. Geller agreed and suggested developing a standard way of measuring all practices and graphically depicting results in a format that is easy for consumers to grasp. For example, the Canadian government has a graphic depicting 1,000 women that is color coded to show how many women's cancers were missed in screening, how many women's biopsies were negative, etc. "I don't think the population understands what false positives are, and it would be great to come up with some sort of patient education tool that was simple to understand and that every practice could provide," Geller said.

Federal Oversight of Breast Imaging

Barr pointed out several "gaps" in MQSA, which originally was created because of concerns about radiation dose and image quality. There was a focus on training for the technologists and medical physicists to provide a quality image and ensure proper dose, and it was assumed that a quality image would ensure quality mammography; the quality of the interpretation was not given precedence, she said. "We're sitting here today because we see there is a big gap that isn't covered by MQSA and the time has come for a complete sea change, overall," she said. "There are certain things we could do in MQSA, but I'm wondering if they're Band-Aids and we really need to refocus MQSA on other aspects. Has MQSA done what it's supposed to do and now there's something else that needs to be done?" Barr asked. For example, she suggested that a new law could require the oversight and mentoring akin to what is done in the United Kingdom.

New technologies have also been developed since MQSA was enacted. Onega noted that mammography is rapidly transitioning from 2D to 3D with the adoption of tomosynthesis and MQSA does not account for this. Barr agreed, adding that X-ray imaging of the breast is all that is included in MQSA oversight currently, not other types of imaging, such as ultrasound and MRI. She thought new amendments to MQSA would not be sufficient to cover oversight of these new technologies. "We would need a fresh piece of paper if we move beyond X-ray imaging."

Smith asked what would be required to create a new statute. Barr responded that ACR and other professional societies and patient advocacy groups should voice the need for new regulation and spark the interest of members of Congress. "Once you say, 'this is what we want' and they buy into it, then the experts in bill writing will decide whether it requires new statutory language or can be built into the existing statute," Barr said. Noting that there is currently a record number of women in Congress, Dentzer added that "the time is probably ripe to bring this issue back to the fore."

However, Pisano offered a voice of dissent for more government regulation, noting that physicians are "fiercely independent." She added, "I'm not saying we can't find ways to improve, but those have to come from the profession itself for them to be effective. . . . We are the poster children on how to improve our practice and that was all due to voluntary programs. MQSA was laid upon a voluntary program. Medicine is in crisis right now and people are feeling beleaguered by regulation so we've got to find a way to motivate radiologists within their own practice."

WRAP-UP

In closing remarks at the workshop, Buist reiterated several options for improving the interpretation of mammograms. She noted that the opportunity for using the quality measures discussed at the workshop may increase with the continued implementation of the Affordable Care Act and pay-for-performance incentives that are increasingly being used by payers and accountable care organizations.

But she added that there is also a need for greater awareness of the limits of these measures and how they should be combined in meaningful ways because one performance metric is not sufficient to assess quality. She said it was also important to consistently measure performance to ensure that we are comparing apples to apples, especially if patients will use quality metrics when comparing and choosing practitioners or mammography facilities. Differences in patient populations and other confounders in such comparisons should also be taken into account, she noted.

Buist also emphasized that rather than taking the stick approach to improve performance, as has been common in this country, it might be best to follow a carrot approach similar to the U.K. National Breast Screening Program, which recognizes high performers and aims to lift others into the high performance range with mentorship and support.

Buist also stressed that quality improvement depends on collection and sharing of data. As more data become available, determining how best to use the data will be critical, she said. At the same time, recognition must be given to the fact that more than just data are needed to implement changes for quality improvement. In addition, she emphasized the need to document both the short- and long-term effects of various educational opportunities, such as selectorships at Centers of Excellence, and to determine whether various CME programs and self-assessment tests improve outcomes and for how long.

In closing, Buist pointed out that since the advent of MQSA, mammography has been at the forefront in medicine for assessing and ensuring quality performance, and what has been learned from that experience could be applied to other areas of medicine, including lung and colon cancer screening programs. “What we are doing today is really relevant to virtually every field of medicine,” Buist concluded.

REFERENCES

Adcock, K. A. 2004. Initiative to improve mammogram interpretation. *Permanente Journal* 8(2):12-18.

Bassett, L., R. Hendrick, and T. Bassford. 1994. Clinical practice guideline number 13: Quality determinants of mammography. AHCPR publication 95-0632. Rockville, MD: U.S. Department of Health and Human Services, Agency for Healthcare Policy and Research. *Public Health Service* 83.

Bassett, L. W., A. J. Hollatz-Brown, R. Bastani, J. G. Pearce, K. Hirji, and L. Chen. 1995. Effects of a program to train radiologic technologists to identify abnormalities on mammograms. *Radiology* 194(1):189-192.

Bassett, L. W., B. S. Monsees, R. A. Smith, L. Wang, P. Hooshi, D. M. Farria, J. W. Sayre, S. A. Feig, and V. P. Jackson. 2003. Survey of radiology residents: Breast imaging training and attitudes. *Radiology* 227(3):862-869.

Baxi, S. S., L. Liberman, C. Lee, and E. B. Elkin. 2009. Breast imaging fellowships in the United States: Who, what, and where? *American Journal of Roentgenology* 192(2):403-407.

Berg, W. A., J. D. Blume, J. B. Cormack, E. B. Mendelson, D. Lehrer, M. Böhm-Vélez, E. D. Pisano, R. A. Jong, W. P. Evans, and M. J. Morton. 2008. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA* 299(18):2151-2163.

Berg, W. A., C. J. D'Orsi, V. P. Jackson, L. W. Bassett, C. A. Beam, R. S. Lewis, and P. E. Crewson. 2002. Does training in the breast imaging reporting and data system (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology* 224(3):871-880.

Berg, W. A., Z. Zhang, D. Lehrer, R. A. Jong, E. D. Pisano, R. G. Barr, M. Böhm-Vélez, M. C. Mahoney, W. P. Evans, and L. H. Larsen. 2012. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA* 307(13):1394-1404.

Blanks, R., M. Wallis, and S. Moss. 1998. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: Results from the uk national health service breast screening programme. *Journal of Medical Screening* 5(4):195-201.

Buist, D. S., M. L. Anderson, S. J. Haneuse, E. A. Sickles, R. A. Smith, P. A. Carney, S. H. Taplin, R. D. Rosenberg, B. M. Geller, and T. L. Onega. 2011. Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology* 259(1):72-84.

Buist, D. S., M. L. Anderson, R. A. Smith, P. A. Carney, D. L. Miglioretti, B. S. Monsees, E. A. Sickles, S. H. Taplin, B. M. Geller, and B. C. Yankaskas. 2014. Effect of radiologists' diagnostic work-up volume on interpretive performance. *Radiology* 273(2):351-364.

Burnside, E. S., E. A. Sickles, R. E. Sohlich, and K. E. Dee. 2002. Differential value of comparison with previous examinations in diagnostic versus screening mammography. *American Journal of Roentgenology* 179(5):1173-1177.

Burnside, E. S., Y. Lin, A. Munoz Del Rio, P. J. Pickhardt, Y. Wu, R. M. Strigel, M. A. Elezaby, E. A. Kerr, and D. L. Miglioretti. 2014. Addressing the challenge of assessing physician-level screening performance: Mammography as an example. *PLoS ONE* 9(2):e89418.

Carney, P. A., E. A. Sickles, B. S. Monsees, L. W. Bassett, R. J. Brenner, S. A. Feig, R. A. Smith, R. D. Rosenberg, T. A. Bogart, and S. Browning. 2010. Identifying minimally acceptable interpretive performance criteria for screening mammography. *Radiology* 255(2):354-361.

Carney, P. A., L. Abraham, A. Cook, S. A. Feig, E. A. Sickles, D. L. Miglioretti, B. M. Geller, B. C. Yankaskas, and J. G. Elmore. 2012. Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. *Academic Radiology* 19(9):1114-1120.

Carney, P. A., A. Bogart, E. A. Sickles, R. Smith, D. S. Buist, K. Kerlikowske, T. Onega, D. L. Miglioretti, R. Rosenberg, and B. C. Yankaskas. 2013. Feasibility and acceptability of conducting a randomized clinical trial designed to improve interpretation of screening mammography. *Academic Radiology* 20(11):1389-1398.

Ciatto, S., N. Houssami, D. Bernardi, F. Caumo, M. Pellegrini, S. Brunelli, P. Tuttobene, P. Bricolo, C. Fantò, and M. Valentini. 2013. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (storm): A prospective comparison study. *The Lancet Oncology* 14(7):583-589.

Cook, A. J., J. G. Elmore, W. Zhu, S. L. Jackson, P. A. Carney, C. Flowers, T. Onega, B. Geller, R. D. Rosenberg, and D. L. Miglioretti. 2012. Mammographic interpretation: Radiologists' ability to accurately estimate their performance and compare it with that of their peers. *American Journal of Roentgenology* 199(3):695-702.

DeSantis, C., J. Ma, L. Bryan, and A. Jemal. 2014. Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians* 64(1):52-62.

Elmore, J. G., S. L. Jackson, L. Abraham, D. L. Miglioretti, P. A. Carney, B. M. Geller, B. C. Yankaskas, K. Kerlikowske, T. Onega, and R. D. Rosenberg. 2009. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology* 253(3):641-651.

Elmore, J. G., G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, and S. J. Schnitt. 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* 313(11):1122-1132.

Evans, K. K., R. L. Birdwell, and J. M. Wolfe. 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.

Fenton, J. J., G. Xing, J. G. Elmore, H. Bang, S. L. Chen, K. K. Lindfors, and L.-M. Baldwin. 2013. Short-term outcomes of screening mammography using computer-aided detectiona population-based study of Medicare enrollees. *Annals of Internal Medicine* 158(8):580-587.

Friedewald, S. M., E. A. Rafferty, S. L. Rose, M. A. Durand, D. M. Plecha, J. S. Greenberg, M. K. Hayes, D. S. Copit, K. L. Carlson, and T. M. Cink. 2014. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA* 311(24):2499-2507.

Geller, B. M., A. Bogart, P. A. Carney, E. A. Sickles, R. Smith, B. Monsees, L. W. Bassett, D. M. Buist, K. Kerlikowske, and T. Onega. 2014. Educational interventions to improve screening mammography interpretation: A randomized controlled trial. *American Journal of Roentgenology* 202(6):W586-W596.

Goldman, L. E., J.-P. H. Sebastien, D. L. Miglioretti, K. Kerlikowske, D. S. Buist, B. Yankaskas, and R. Smith-Bindman. 2008. An assessment of the quality of mammography care at facilities treating medically vulnerable populations. *Medical Care* 46(7):701.

Goldman, L. E., R. Walker, D. L. Miglioretti, R. Smith-Bindman, and K. Kerlikowske. 2011. Accuracy of diagnostic mammography at facilities serving vulnerable women. *Medical Care* 49(1):67.

Gur, D., A. I. Bandos, C. S. Cohen, C. M. Hakim, L. A. Hardesty, M. A. Ganott, R. L. Perrin, W. R. Poller, R. Shah, and J. H. Sumkin. 2008. The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 249(1):47-53.

Haas, B. M., V. Kalra, J. Geisel, M. Raghu, M. Durand, and L. E. Philpotts. 2013. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. *Radiology* 269(3):694-700.

Haiart, D., and J. Henderson. 1990. A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist: Results and implications. *British Journal of Clinical Practice* 45(1):43-45.

Haneuse, S., D. S. Buist, D. L. Miglioretti, M. L. Anderson, P. A. Carney, T. Onega, B. M. Geller, K. Kerlikowske, R. D. Rosenberg, and B. C. Yankaskas. 2012. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. *Radiology* 262(1):69-79.

Henderson, L. M., T. Benefield, J. M. Bowling, D. D. Durham, M. W. Marsh, B. F. Schroeder, and B. C. Yankaskas. 2015a. Do mammographic technologists affect radiologists' diagnostic mammography interpretative performance? *American Journal of Roentgenology* 204(4):903-908.

Henderson, L. M., T. Benefield, M. W. Marsh, B. F. Schroeder, D. D. Durham, B. C. Yankaskas, and J. M. Bowling. 2015b. The influence of mammographic technologists on radiologists' ability to interpret screening mammograms in community practice. *Academic Radiology* 22(3):278-289.

Hofvind, S., P. M. Vacek, J. Skelly, D. L. Weaver, and B. M. Geller. 2008. Comparing screening mammography for early breast cancer detection in vermont and norway. *Journal of the National Cancer Institute* 100(15):1082-1091.

Hooley, R. J., K. L. Greenberg, R. M. Stackhouse, J. L. Geisel, R. S. Butler, and L. E. Philpotts. 2012. Screening us in patients with mammographically dense breasts: Initial experience with Connecticut Public Act 09-41. *Radiology* 265(1):59-69.

IOM (Institute of Medicine). 2005. *Improving breast imaging quality standards*. Washington, DC: The National Academies Press.

Kuhl, C. K., S. Schrading, K. Strobel, H. H. Schild, R.-D. Hilgers, and H. B. Bieling. 2014. Abbreviated breast magnetic resonance imaging (MRI): First postcontrast subtracted images and maximum-intensity projection—a novel approach to breast cancer screening with MRI. *Journal of Clinical Oncology* 32(22):2304-2310.

Lehman, C. D., R. Wellman, D. M. Buist, K. Kerlikowske, A. N. Tosteson, and D. L. Miglioretti. Under review. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*.

Linver, M., S. Paster, R. Rosenberg, C. Key, C. Stidley, and W. King. 1992. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology* 184(1):39-43.

Miglioretti, D. L., C. C. Gard, P. A. Carney, T. L. Onega, D. S. Buist, E. A. Sickles, K. Kerlikowske, R. D. Rosenberg, B. C. Yankaskas, and B. M. Geller. 2009. When radiologists perform best: The learning curve in screening mammogram interpretation. *Radiology* 253(3):632-640.

Miglioretti, D. L., L. Ichikawa, R. A. Smith, L. W. Bassett, S. A. Feig, B. Monsees, J. R. Parikh, R. D. Rosenberg, E. A. Sickles, and P. A. Carney. 2015. Criteria for identifying radiologists with acceptable screening mammography interpretive performance on basis of multiple performance measures. *American Journal of Roentgenology* 204(4):W486-W491.

Miller, J. W., J. B. King, D. A. Joseph, L. C. Richardson, and Centers for Disease Control and Prevention. 2012. Breast cancer screening among adult women—Behavioral Risk Factor Surveillance System, United States, 2010. *Morbidity and Mortality Weekly Report* 61(Suppl):46-50.

Onega, T., R. Hubbard, D. Hill, C. I. Lee, J. S. Haas, H. A. Carlos, J. Alford-Teaster, A. Bogart, W. B. DeMartini, and K. Kerlikowske. 2014. Geographic access to breast imaging for US women. *Journal of the American College of Radiology* 11(9):874-882.

Onega, T., L. E. Goldman, R. Walker, D. L. Miglioretti, D. M. Buist, S. Taplin, B. Geller, D. Hill, and R. Smith-Bindman. 2015. Facility mammography volume in relation to breast cancer screening outcomes. *Journal of Medical Screening*. Epub ahead of print.

Parris, T., D. Wakefield, and H. Frimmer. 2013. Real world performance of screening breast ultrasound following enactment of Connecticut Bill 458. *The Breast Journal* 19(1):64-70.

Pauli, R., S. Hammond, J. Cooke, and J. Ansell. 1996. Radiographers as film readers in screening mammography: An assessment of competence under test and screening conditions. *British Journal of Radiology* 69(817):10-14.

Perry, N., M. Broeders, C. De Wolf, S. Törnberg, R. Holland, and L. Von Karsa. 2008. European guidelines for quality assurance in breast cancer screening and diagnosis—summary document. *Annals of Oncology* 19(4):614-622.

Roberge, D. 2007. Provider's volume and quality of breast cancer detection and treatment. *Breast Cancer Research & Treatment* 105(2):117-132.

Rose, S. L., A. L. Tidwell, L. J. Bujnoch, A. C. Kushwaha, A. S. Nordmann, and R. Sexton Jr. 2013. Implementation of breast tomosynthesis in a routine screening practice: An observational study. *American Journal of Roentgenology* 200(6):1401-1408.

Rutter, C. M., and S. Taplin. 2000. Assessing mammographers' accuracy: A comparison of clinical and test performance. *Journal of Clinical Epidemiology* 53(5):443-450.

Scott, H. J., and A. G. Gale. 2006. Breast screening: PERFORMS identifies key mammographic training needs. *British Journal of Radiology* 79(Spec. No. 2):S127-S133.

Scott, H. J., A. Evans, A. G. Gale, A. Murphy, and J. Reed. 2009. The relationship between real life breast screening and an annual self assessment scheme. Paper read at SPIE Medical Imaging. Buena Vista, FL. February 7.

Skaane, P., A. I. Bandos, R. Gullien, E. B. Eben, U. Ekseth, U. Haakenaasen, M. Izadi, I. N. Jebsen, G. Jahr, and M. Krager. 2013. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology* 267(1):47-56.

Skaane, P., B. Osteras, E. Eben, and R. Gullien. 2014. Comparison of digital mammography (FFDM) and FFDM plus digital breast tomosynthesis in mammography screening for cancer detection according to breast parenchyma density. Paper read at Radiological Society of North America, Chicago, IL.

Soh, B., W. Lee, P. Kench, W. Reed, M. McEntee, A. Poulos, and P. Brennan. 2012. Assessing reader performance in radiology, an imperfect science: Lessons from breast screening. *Clinical Radiology* 67(7):623-628.

Soh, B. P., W. B. Lee, C. Mello-Thoms, K. Tapia, J. Ryan, W. T. Hung, G. Thompson, R. Heard, and P. Brennan. 2015. Certain performance values arising from mammographic test set readings correlate well with clinical audit. *Journal of Medical Imaging and Radiation Oncology* 59(4):403-410.

Taplin, S., L. Abraham, W. E. Barlow, J. J. Fenton, E. A. Berns, P. A. Carney, G. R. Cutter, E. A. Sickles, and J. G. Elmore. 2008. Mammography facility characteristics associated with interpretive accuracy of screening mammography. *Journal of the National Cancer Institute* 100(12):876-887.

Théberge, I., S.-L. Chang, N. Vandal, J.-M. Daigle, M.-H. Guertin, É. Pelletier, and J. Brisson. 2014. Radiologist interpretive volume and breast cancer screening accuracy in a canadian organized screening program. *Journal of the National Cancer Institute* 106(3):djt461.

Timmers, J., A. Verbeek, R. Pijnappel, M. Broeders, and G. Den Heeten. 2014. Experiences with a self-test for dutch breast screening radiologists: Lessons learnt. *European Radiology* 24(2):294-304.

Urban, N., G. M. Longton, A. D. Crowe, M. J. Drucker, C. D. Lehman, S. Peacock, K. A. Lowe, S. B. Zeliadt, and M. A. Gaul. 2007. Computer-assisted mammography feedback program (CAMFP): An electronic tool for continuing medical education. *Academic Radiology* 14(9):1036-1042.

USPSTF (U.S. Preventive Services Task Force). 2015. *Draft recommendation statement: Breast Cancer: Screening*. http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementDraft/breast-cancer-screening1 (accessed September 3, 2015).

van den Biggelaar, F., K. Flobbe, J. van Engelshoven, and N. de Bijl. 2009. Pre-reading mammograms by specialised breast technologists: Legal implications for technologist and radiologist in the netherlands. *European Journal of Health Law* 16(3):271-279.

Weigert, J., and S. Steenbergen. 2012. The connecticut experiment: The role of ultrasound in the screening of women with dense breasts. *The Breast Journal* 18(6):517-522.

Wivell, G., E. Denton, C. Eve, J. Inglis, and I. Harvey. 2003. Can radiographers read screening mammograms? *Clinical Radiology* 58(1):63-67.

Appendix

Statement of Task and Workshop Agenda

STATEMENT OF TASK

An ad hoc committee will plan and host a 1.5-day public workshop to examine the evidence regarding interpretive performance in breast cancer screening. The workshop will feature panel discussions and invited presentations from experts in the interpretation of screening mammography to assess the state of the evidence, identify remaining gaps, and examine potential opportunities for advancing research and practice.

Participants will be invited to discuss topics related to

- the relationship between reader volume and interpretive performance;
- the relationship between screening and diagnostic volume, and the impact of working up one's own recalled cases on interpretative performance;
- performance criteria to identify radiologists who might benefit most from interventions;
- use of test instruments to assess interpretive performance; and
- tools and interventions to improve interpretive performance.

The committee will develop the agenda for the workshop sessions, select and invite speakers and discussants, and moderate the discussions. An individually authored workshop summary of the presentations and discussions at the workshop will be prepared by a designated rapporteur in accordance with institutional guidelines.

WORKSHOP AGENDA

May 12, 2015

7:45 am **Registration**

8:00 am **Welcome from the National Cancer Policy Forum and the American Cancer Society**
Patricia Ganz, University of California, Los Angeles, Vice Chair, National Cancer Policy Forum
Robert Smith, American Cancer Society

Overview and Goals of the Workshop
Diana Buist, Group Health Research Institute
Planning Committee Chair

Overview of the 2005 IOM Consensus Report
Improving Breast Imaging Quality Standards
Etta Pisano, Medical University of South Carolina

8:45 am **Session 1: Challenges in the Delivery of High-Quality Mammography**
Moderator: Barbara Monsees, Washington University School of Medicine

Overview of Current Challenges
Barbara Monsees, Washington University School of Medicine

Geographic Access, Equity, and Impact on Quality
Tracy Onega, Dartmouth Medical School

Audits and the National Mammography Database
Carl D’Orsi, Emory Healthcare

Group Discussion

10:30 am **Break**

10:45 am **Session 2: Training/Experience and Interpretive Performance**
Moderator: Diana Buist, Group Health Research Institute

U.S. and International Variation in Volume and Performance Measures
Diana Buist, Group Health Research Institute

Interpretive Volume and Accuracy in Canada
Isabelle Théberge, National Public Health Institute, Quebec

The Role of Specialist Radiology Technologists
Louise Henderson, University of North Carolina School of Medicine

Residency Requirements and Board Certification and Maintenance
Debra Monticciolo, Texas A&M College of Medicine

Group Discussion

12:30 pm **Lunch Break**

1:15 pm **Session 3: Identifying Radiologists and Facilities That Would Benefit from Intervention**
Moderator: Diana Miglioretti, University of California, Davis

Diagnostic and Screening Aspects
Patricia Carney, Oregon Health & Science University and the Knight Cancer Institute

Joint Criteria and Confidence Interval-Based Approaches
Diana Miglioretti, University of California, Davis

Identifying Facilities Through CMS Data
Rebecca Hubbard, University of Pennsylvania School of Medicine

Group Discussion

3:00 pm **Break**

3:15 pm **Session 4: Test Instruments to Assess Interpretive Performance: Challenges and Opportunities**
Moderator: Lora Barke, Radiology Imaging Associates

Overview of Test Set Design and Use
Robert Smith, American Cancer Society

Overview of International Test Sets
Mireille Broeders, Dutch Reference Center for Screening

Panel Discussion:
Session speakers and Matthew Wallis, Cambridge and Huntington Breast Screening Service

Group Discussion

4:30 pm **Wrap-Up Day 1 and Adjourn**

May 13, 2015

8:00 am **Registration**

8:30 am **Session 5: Tools and Interventions to Improve Interpretive Performance**
Moderator: Patricia Carney, Oregon Health & Science University Cancer Institute

Educational Interventions to Improve Screening
Berta Geller, University of Vermont

Thresholds for Performance
Matthew Wallis, Cambridge and Huntington Breast Screening Service

Panel Discussion:
Session speakers and
Mireille Broeders, Dutch Reference Center for Screening
Carl D'Orsi, Emory Healthcare

Group Discussion

10:00 am **Break**

10:15 am **Session 6: Reactor Panel: Potential Solutions to Current Challenges**
Moderator: Susan Dentzer, Robert Wood Johnson Foundation

Panelists:
Lora Barke, Radiology Imaging Associates
Helen Barr, Food and Drug Administration
Bryan Loy, Humana
Barbara Monsees, Washington University School of Medicine
Dana Smetherman, Ochsner Health System
Robert Smith, American Cancer Society
Kelly Walborn, Patient Advocate

Group Discussion

11:45 am **Workshop Wrap-Up**
Diana Buist, Group Health Research Institute

12:00 pm **Adjourn**