# Best Ways to Provide Feedback to Radiologists on Mammography Performance

**Erin J. Aiello Bowles**[1] and **Berta M. Geller**[2]

[1]Group Health Center for Health Studies, 1730 Minor Ave, Ste. 1600, Seattle, WA 98101. Address correspondence to E. J. Aiello Bowles (aiello.e@ghc.org)

[2]Office of Health Promotion Research, Department of Family Medicine, Department of Radiology, and the Vermont Cancer Center, University of Vermont, Burlington, VT

## Abstract

**Objective—**The Mammography Quality Standards Act of 1992 required a minimum performance audit of radiologists performing mammography. Since then, no studies have evaluated radiologists' perceptions of their audit reports, such as which performance measures are the most or least useful, or what the best formats are to present performance data.

**Materials And Methods—**We conducted a qualitative study with focus groups and interviews of 25 radiologists currently practicing mammography. All radiologists practiced at one of three sites in the Breast Cancer Surveillance Consortium (BCSC). The discussion guide included open-ended questions to elicit opinions on the following subjects: the most useful performance outcome measures, examples of reports and formats that are easiest to understand (e.g., graphs or tables), thoughts about comparisons between individual-level and aggregate data, and ideas about additional performance measures they would find useful. All discussions were tape-recorded and transcribed. We developed a set of themes and used ethnographic software to qualitatively analyze and extract quotes from transcripts.

**Results—**Radiologists thought that almost all performance measures were useful. They particularly liked seeing individual data presented in graphic form with a national benchmark or guideline for each performance measure clearly marked on the graph. They appreciated comparisons between their individual data and their peers' data (within their facility or state) and requested comparisons with national data (such as the BCSC). Many thought customizable, Web-based reports would be useful.

**Conclusion—**Radiologists think that most audit statistics are useful; however, presenting performance data graphically with clear benchmarks may make them easier to understand.

## Keywords

audit reports; mammography performance; provider feedback; quality; radiology

---

Mammography remains the best available breast cancer screening technology, but U. S. radiologists vary widely in how accurately they interpret mammograms [1–3]. Although some radiologists specialize in breast imaging, radiology generalists interpret more than two thirds of mammograms in the United States [4]. The effectiveness of mammography depends on the ability to perceive mammographic abnormalities and to interpret these findings accurately; both tasks are quite challenging and require ongoing education to improve interpretive skills, or even to maintain them. For this reason, many countries with organized screening programs have developed audit and feedback systems based on their national data to help radiologists assess and improve their skills [5,6]. In the United States, the Mammography Quality Standards Act of 1992 (MQSA) required a minimal audit of U. S. radiologists who interpret

mammography [7–9]. But unlike other countries, the United States does not offer standard audit or feedback systems to radiologists to help them evaluate and improve their skills.

As first envisioned, the MQSA audit was to serve as a teaching tool and performance summary for each radiologist. The National Cancer Institute–funded Breast Cancer Surveillance Consortium (BCSC) was originally started to meet the MQSA mandate to establish a breast cancer screening surveillance system [10]. The BCSC published benchmarks for both screening and diagnostic mammography based on the performance of community radiologists in 2005 and 2006 [2,3]. Benchmarks let radiologists compare their performance with that of others and to accepted practice guidelines; but we are unaware of any studies that have assessed whether radiologists find these published benchmarks useful as guidelines for their own performance.

Studies in mammography and other areas of medicine have shown that audit feedback has indeed improved performance [5,11–15]. Providing benchmarks or guidelines has also been shown to improve effectiveness of physicians' performance in ambulatory settings [16]. The recent Institute of Medicine (IOM) report, *Improving Breast Imaging Quality Standards* [17], recommended increasing the number of audit measures MQSA requires so that all practicing radiologists who interpret mammograms have this tool for improvement. However, neither the IOM report nor any other feedback study has given any indication as to whether the additional audit measures will be useful or the best format to present these measures to radiologists. In addition, the IOM report does not address instructions or education that clinicians may need to fully understand rates, risks, and proportions and make them useful. To our knowledge, there are no published articles on working with radiologists to develop these tools.

The purpose of this study was to examine radiologists' understanding of audit measures and to explore various formats for presenting audit data. We conducted focus groups and interviews with radiologists who interpret mammograms at one of three BCSC sites to collect qualitative data. We conducted this study within the BCSC because all five sites currently provide detailed individual-level or facility-level audit reports to their participating radiologists. Our goal was to understand and recommend the best ways to provide feedback to radiologists on their mammography performance.

## Materials and Methods

### Study Setting

The BCSC is funded by the National Cancer Institute to study the effectiveness of mammography in community practice [10]. There are five sites across the nation that collect breast imaging data from facilities within their region, along with patient demographic and risk factor information. Four of the five sites send annual outcome audit reports to participating radiologists; the fifth site sends only facility reports. Institutional review board approval was obtained at three sites that participated in this study: Group Health, an integrated delivery system based in western Washington; Vermont Breast Cancer Surveillance System, a statewide mammography registry; and the New Hampshire Mammography Network, an almost statewide mammography registry.

Recruitment letters were mailed or e-mailed to a convenience sample of radiologists from three sites. The convenience sample consisted of radiologists who had shown enthusiasm for improving their audit reports in prior meetings. We offered an honorarium and a meal as incentives for participation in the focus group. We invited radiologists to attend a focus group at a specific date and time. If the radiologist wanted to participate but the date of the focus group was not convenient, we invited him or her to be interviewed at a mutually agreed on date. From September 2007 through February 2008, we conducted five focus groups and two

in-depth interviews with a total of 25 radiologists representing five radiology practices in three states. The backgrounds of the radiologists varied widely because we recruited from small rural community practices, larger hospitals, and an integrated delivery system.

### Development of Focus Group Materials

We reviewed mammography outcome audit reports from the five current BCSC sites plus one former site as well as from four members of the International Cancer Screening Network (United Kingdom, Tasmania, Canada, and Israel). We compared each report with the American College of Radiology (ACR) Basic Clinical Relevant Mammography Audit for completeness and consistency of definitions, and with each other for presentation formats [18]. The performance measures and definitions that are included in the ACR minimum audit are outlined in Table 1. From this review we created a discussion guide and sample audit reports to use in the focus groups and interviews with radiologists. The discussion guide consisted of several topic headings including outcome measurements (questions about understanding the measures, what are the most and least useful, should any measures be reported by covariates, number of measures desired); review of different formats for data (ease and accuracy of interpreting data, like and dislike about format, time trends, cumulative or annual reporting periods); comparison data (desire for comparisons, with whom); and finally, a more detailed review of the two measures liked best.

### Data Collection and Analysis

All the sessions were audio taped and transcribed verbatim, deleting all names. After each focus group or interview, the authors met to review what was learned. From this we modified both the discussion guide and the example audit reports. We used qualitative data management programs to analyze the transcripts (NVivo, Atlas). These software programs allow the user to organize and code the text of transcripts into themes and to easily extract quotes for an analysis. Each author independently analyzed all transcripts twice. Concepts and themes were agreed on before analyses, with the plan to add others as they emerged. The authors discussed the independent analyses and came to consensus when there were different interpretations of the data. We classified the results into major and minor themes.

## Results

Below we describe the focus group and interview results according to the major themes: radiologists' understanding of performance measures; comparison data; graphic data displays; and customizable Web reports. We provide quotes from the focus groups or interviews that illustrate each theme. We also summarize the feedback on each audit report example.

### Radiologists' Understanding of Performance Measures

Almost all radiologists thought all of the performance measures recommended by the ACR were important to their practice (see Table 1 for a list with definitions). Recall rate, sensitivity, and a list of known false-negative examinations were identified as being the most important. Although they did not recommend deleting any from the reports, there were a few they thought could use additional explanation. Radiologists did not understand the differences between PPV1 (positive predictive value 1), the percentage of all positive screening examinations that result in a cancer diagnosis within 1 year; PPV2, the percentage of all screening or diagnostic examinations recommended for biopsy or surgical consultation that result in a cancer diagnosis within 1 year; and PPV3, the percentage of all biopsies done after a positive screening or diagnostic examination that result in a cancer diagnosis within 1 year. Many wanted to know which positive examinations were really true-positives, but they did not know this was their PPV when they saw it. In addition, they did not understand how results were attributed to

radiologists when different radiologists performed the initial screening and diagnostic follow-up mammograms.

"I never know if it's just the one that I biopsied; let's say 'XXXX' recommended a biopsy and I performed it—does that go on MY statistic or does it go on her statistic?"

### Comparison Data

All radiologists agreed that peer comparison data were helpful in understanding how well they were doing and identifying areas for improvement. We describe three types of comparisons that radiologists thought were particularly useful: comparisons with peers, comparisons with national guidelines, and comparisons over time.

**Comparisons with peers—**In general, radiologists like to compare themselves with their group practice, region, or state, and many suggested they would like to compare themselves with national data (such as the BCSC). They wanted to see how their individual performance compared with their colleagues' performances.

"[I like] comparing it to my colleagues rather than the state because I think we generally have higher recall rates than the rest of the state. So I want to make sure I am in line with my group. It is interesting, though, how we are doing compared with the rest of the state too, so that has directly impacted my practice."

"I think any comparison that allows us, any format that allows us to compare our performance to each other and to other, other institutions, other facilities…is very helpful."

**Comparisons with national guidelines—**Radiologists thought that providing national benchmarks within the audit report would allow them to immediately identify areas that need improvement. They thought that benchmarks could provide a way for them to see whether they were "veering off from typical practice standards." We tested various formats for providing benchmarks that we refined throughout the focus groups and interviews. Some radiologists preferred to see a line on a graph representing a benchmark.

"…you are below, meeting, or above the expectation…it's not a specific number."

Others thought that the specific benchmark number provided more information.

"[Providing a number] would actually be better than the meeting/exceeding [standards], because then it's imprinted in your mind what the expectation is. What the benchmark is. I think that actually is the best model."

**Comparisons over time—**Radiologists liked to compare their own performance from year to year but also recognized that some measures may be unstable with only 1 year of data. When shown a graph of performance measures year by year, radiologists liked how easy it was to see changes. However, several radiologists suggested cumulative measures over time might be more representative of their performance.

"This I think is a little bit problematic just because for some of the radiologists in our group who read so few mammograms the sensitivity may actually be affected in any given year…and I'm assuming isn't the cumulative…."

### Graphic Data Displays

We included a few tabulated formats and many different graphic formats to display performance data. Radiologists overwhelmingly preferred the graphic displays because they consider themselves "visual people." To them, the graphs were easier to understand and

interpret than the tables. One radiologist stated that "a sheet of numbers basically will make people's eyes glaze over."

> "You are dealing with a group of radiologists who are very spatial…and visually oriented,…that may have something to do with it."

### Customizable Web Reports

Discussion during the focus groups and interviews led to some unanticipated findings. Several people suggested that they would like to access their reports on the Internet to reduce paper. Web-based reports could allow them to customize how and when they wanted to see their reports. For example, they could select varying time frames, different graphic formats, different comparison groups, or cumulative rather than yearly data.

> "You could have all your statistics on a secure Website with a password. You don't have to even see papers."

### Feedback on Audit Report Examples

We summarize the radiologists' reactions to the audit report examples we presented in the focus groups and interviews. The examples we showed are described in detail in Table 2 and illustrated in Figure 1. They included a scattergram, a vertical bar graph with benchmarks, a single-page table, a $2 \times 2$ table, a color-coded chart, and color-coded horizontal bar graphs. Table 2 also includes what the radiologists liked and disliked about each format. Below, we review the examples in the order that the radiologists ranked them, most useful to least useful.

**Recall by sensitivity scattergram—**This graph showed how sensitivity (on the *y*-axis) increased with recall rate (on the *x*-axis) [19]. Each radiologist was shown as an individual point with a regression line to summarize the relationship between recall and sensitivity. Two vertical lines on the graph indicated the lowest range of recall rates that still resulted in the highest sensitivity. All radiologists liked this graph, although for some it was not intuitive and took more explanation. They interpreted the two vertical lines as national benchmarks rather the range of recall rates that produced the best sensitivity.

> "This is very helpful. It gives the range where you should be. The circle is excellent. It does show…the lowest certain recall rate. There is really a fall-off in what you're producing for the benefit of the patient. Over here, your call back rate…doesn't add anything to the patient's value, to the value of the patient, of the exam to the patient."

> "I think just visually that is easier for me to get a sense of where my performance is compared to everybody else's."

**Vertical bar graph with benchmark—**We presented a vertical bar graph with screening sensitivity (on the *y*-axis) over time (past 5 years on the *x*-axis) with different bars for the individual radiologist, his or her primary practice region, and the entire registry. We added a single line to the graph to indicate 85% sensitivity (the BI-RADS recommended benchmark), and we color-coded the sensitivity values green or red depending on whether they were above or below the benchmark. This graph was the second favorite of the radiologists. Based on the feedback we received in early focus groups, we tested horizontal bar graphs and line graphs showing the same data in later focus groups. However, most radiologists still preferred the vertical format. We also added a fourth bar to represent sensitivity values for the entire BCSC; most radiologists favored this addition. They also thought that the color-coding of sensitivity values was unnecessary.

> "I think (in the bar graph) you got all the data in one look… well, first you have the idea of just where you stand in terms of the height of the histogram, and then you can

look at the individual numbers and compare, and you know which since they are colored; and you know if you are specifically interested in looking at your facility and your colleagues, you can do it and when you look at overall…. I feel it's easy to identify who you are… and you can see easily your peaks and where you are toward this consortium."

**Single-page table—**This table included all performance measures on a single page. It showed cumulative data for the entire period the radiologist participated in the registry compared with the entire state. Radiologists liked that it was all on one sheet, included all the ACR recommended performance measures, and included statewide comparison data, but they missed the graphic display. This table included screening mammograms in one column and all mammograms (screening plus diagnostic) in another. Radiologists commented that they would prefer to see screening and diagnostic mammography separately.

"I really like [the table] because it has the yearly and the cumulative. And then if you add few line graphs to that… That might be useful for the radiologist themselves, but I think this is a great sheet."

**2 × 2 table—**This was a standard $2 \times 2$ table showing the number of true-positive, false-negative, false-positive, and true-negative examinations. Several radiologists stated that this was a familiar way to look at performance data.

"It is the same one we did in med school so it is easy to understand."

Some found the four cells were not helpful without the formulas that allow them to calculate sensitivity and specificity performance measures. Others stated that even with the formulas, they would not use them to calculate sensitivity and specificity themselves. Radiologists did not like that this table had no comparison data or benchmarks. Some preferred to look at these data (number and percentage of true- and false-positives and -negatives) in tabulated form as in the single-page table.

**Color-coded chart—**This was the only example we showed that compared performance across facilities. The chart used color codes to represent whether the facilities achieved (green) or failed (red) various standards over the past year. Radiology chiefs in particular thought that this could provide a quick assessment to identify radiologists or facilities with poor performance in order to provide coaching. Others thought the format was better suited for facilities rather than for individuals because it did not provide actual performance rates.

"If I had to make a presentation to the board of trustees or somebody else like that, this is a very visual document that everybody can see literally from across the room."

"I went to a pass–fail honors medical school. It was disconcerting to know that you were in a group but not know where you were in the group."

**Color-coded horizontal bar graph—**This bar graph summarized outcome measures associated with the detriments (e.g., recall and biopsy rates) and benefits (e.g., cancer detection and small invasive cancer rates) of screening mammography. The bars were horizontal and stratified outcome measures by first and subsequent screens. Each bar was color-coded so that a radiologist could see if he or she achieved (green) or failed (red) targets for various standards. The target was 100% for each measure because measures were calculated as proportions; this was confusing to the radiologists and different from the way most were used to seeing their performance data.

"I find this confusing. Because this is less than 100 and over 100 and the fact that you could be red on one side of the line and green on the other side of the line."

Radiologists suggested that the targets need to change relative to the performance measure. For instance, if recall is presented, the target range could be 8–20%. Some people found it easier to understand than the color-coded chart but suggested changing the measure to show how far off they were from the goal. We did not show this graph to the last two focus groups because no one in the previous groups liked this way of displaying performance data.

## Discussion

Audit reports are a common method of providing feedback for quality improvement in professional practice and health care services. Several reviews have shown that they result in small to moderate effects on performance in clinical practice and are more effective when baseline performance is at a low level, when combined with educational meetings, or when administered by an authoritative figure [12,13,20]. Our study provides insight into what radiologists understand and want in their mammography audit reports. Not surprisingly, visual display of data was important, as well as comparing oneself with peers and national benchmarks. What was surprising was the radiologists' desire to use the Internet to access their reports and customize the presentation of their performance data.

In a qualitative study of clinical and administrative staff at eight U. S. hospitals, the authors elucidated seven key themes related to the effectiveness of performance feedback data [11]. One of the themes, in particular, resonated with our study: "Benchmarking improves the meaningfulness of the data feedback." We discussed several different ways of providing benchmarks, and many radiologists requested that this be included on their reports. Targets for mammography performance were created using the opinion of experienced radiologists and were subsequently published in the ACR BI-RADS manual [18]. These audit measures include recommended ranges and minimums for performance [18]. In addition, the BCSC has published articles that provide benchmarks for screening and diagnostic mammography based on national data [2,3]. Because there may be subtle differences in the way outcome data are calculated (adjusted or unadjusted) and variations in definitions used to determine a positive and negative examination (initial or final BI-RADS assessment), it may be difficult for individual radiologists to make an exact comparison with the published benchmarks. Several reports have emphasized that comparing individual data with benchmarks makes performance data more clinically useful [16,21]. The radiologists in our study clearly agreed.

A second major finding of our study was that radiologists preferred and found it easier to interpret graphic data displays than tabular. This is not surprising because radiologists consider themselves "visual" individuals and find it easier to see trends in performance on a graph than in a table. However, we are unaware of any reports or guidelines that support this intuitive finding. In fact, one review that addressed the format of feedback reports was limited by the fact that many published studies do not provide sufficient descriptions or examples of their reports [20]. In our study, radiologists appeared to prefer two types of graphs: one that describes the relation between performance measures (e.g., recall by sensitivity scattergram); and one that compared trends in performance over time and between individuals and their peers (e.g., vertical bar graph). Having comparison data to review were important to our radiologists so that they could see whether they were on par with their colleagues. Peer-comparison feedback has been shown to be effective in promoting adherence to guidelines in other studies [22], but we are unaware of any studies that have described the best format for providing comparison data. Future studies should explore additional ways to present these data that might improve on these examples.

The MQSA audit was designed to serve as a teaching tool and summary for each radiologist [2]. Many of the early performance targets were developed on the basis of the evaluation of outcomes from small groups of radiologists with a special interest in breast imaging [17,18].

The MQSA requires radiology facilities to collect data for a minimal audit that includes following up positive mammographic examinations until resolution, but it does not suggest how to use audit data to help improve performance. Because the interpretive skills of radiologists vary widely, studies have been developed to test methods to improve these skills [5,6,15]. Several studies have included audit feedback as part of multicomponent interventions to enhance radiologists' interpretive performance [5,6,15]. In a study by Perry [5], radiologists certified in the United Kingdom National Health Program underwent training that included a 2-week multidisciplinary course with a specialist, training in high-volume screening sites, and three sessions per week of interpreting screening mammograms. Radiologists also attended routine breast disease–related meetings and received personal and group audit reports, including data on cancer detection rate, recall rate, and PPV2. With all these combined activities, performance indices improved: recall rate dropped from 7% to 4%, and the detection rate of small invasive cancer increased from 1.6 to 2.5 per 1,000. Unfortunately, the specific contribution of audit reports versus other aspects of the intervention is unknown.

In the recent IOM report, *Improving Breast Imaging Quality Standards* [17], the number-one recommendation was to revise and standardize the medical audit component of the MQSA to make it more meaningful and useful [17]. The authors suggested these audits be enhanced to include (stratified by screening and diagnostic examinations) PPV of a recommendation for biopsy (PPV2); cancer detection rate per 1,000 women; and abnormal interpretation rate, sometimes called recall rate. Audit reports for most of the radiologists in our study already included PPV2, cancer detection rate, and abnormal interpretation rate, and our study participants agreed that these are valuable measures.

Several radiologists from different focus groups in our study strongly favored providing access to performance data via the Internet. In a discussion about format (in which they did not agree on the best format), radiologists suggested that performance data on the Web be customized by the radiologists into a format of their preference. We are unaware of any studies that have tested Web-based applications for providing mammography performance data. Developing and testing a standardized way of presenting audit data will provide evidence and models for MQSA to make these data more meaningful. Including radiologists in the development and critical review of these audit reports will make the end product more relevant, useful, and acceptable to any possible new MQSA rules.

Our qualitative study had several limitations. Qualitative studies are designed to understand the depth and breadth of the issues and not to be experimental. Our study had a limited number of respondents who included radiologists specifically motivated to provide feedback toward improving their audit reports. Therefore, as with all qualitative studies, we have some selection bias in our sample that limits the generalizability of our findings. In addition, all radiologists who participated in this study practice at facilities that collect and provide mammography outcome data to the BCSC in a standard format. The results may not be generalizable to radiologists or facilities that do not routinely collect mammography data or do not have the ability to compare their data with those of other facilities. We also did not collect demographic data on the radiologists in our study and therefore could not make comparisons with radiologists who did not participate.

However, our study also has several strengths. We were able to include radiologists from a wide variety of clinical settings, including small rural community practices, larger hospitals, and an integrated delivery system. We were also able to test a wide variety of reports and data formats currently in use by facilities in the United States and other parts of the world. As far as we know, no previous studies have evaluated specific data formats or the best ways to graphically present performance data.

Audit reports are a valuable tool to radiologists who strive to improve their performance. Our focus groups and interviews clearly showed that all performance statistics are important, but the way data are presented to radiologists can improve the value of the reports. The inclusion of graphic data displays, benchmarks, and comparison data may be easily done by the BCSC and other facilities that collect mammography outcome data and provide regular audit reports. Future studies may evaluate whether Web-based applications can further improve the value of audit reports to radiologists, in addition to making these reports more cost-effective and easier for facilities to provide.

## Acknowledgments

## References

1. Barlow WE, Lehman CD, Zheng Y, et al. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. J Natl Cancer Inst 2002;94:1151–1159. [PubMed: 12165640]

2. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006;241:55–66. [PubMed: 16990671]

3. Sickles EA, Miglioretti DL, Ballard-Barbash R, et al. Performance benchmarks for diagnostic mammography. Radiology 2005;235:775–790. [PubMed: 15914475]

4. Lewis RS, Sunshine JH, Bhargavan M. A portrait of breast imaging specialists and of the interpretation of mammography in the United States. AJR 2006;187:W456–468. [PubMed: 17056875]1167; [web]

5. Perry NM. Interpretive skills in the National Health Service Breast Screening Programme: performance indicators and remedial measures. Semin Breast Disease 2003;6:108–113.

6. van der Horst F, Hendriks J, Rijken H, Holland R. Breast cancer screening in the Netherlands: audit and training of radiologists. Semin Breast Disease 2003;6:114–122.

7. Mammography Quality Standards Act of 1992. Mammography facilities requirement for accrediting bodies, and quality standards and certifying requirements: interim rules (21 CFR 900) December 21;1993 58:57558–57572.

8. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the mammography quality standards act (MQSA). AJR 1995;165:19–25. [PubMed: 7785586]

9. Linver M, Newman J. MQSA: the final rule. Radiol Technol 1999;70:338–353. [PubMed: 10101751] quiz 354–336

10. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR 1997;169:1001–1008. [PubMed: 9308451]

11. Bradley EH, Holmboe ES, Mattera JA, Roumanis SA, Radford MJ, Krumholz HM. Data feedback efforts in quality improvement: lessons learned from US hospitals. Qual Saf Health Care 2004;13:26–31. [PubMed: 14757796]

12. Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. Does telling people what they have been doing change what they do? A systematic review of the effects of audit and feedback. Qual Saf Health Care 2006;15:433–436. [PubMed: 17142594]

13. Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2006:CD000259. [PubMed: 16625533]

14. Rowan MS, Hogg W, Martin C, Vilis E. Family physicians' reactions to performance assessment feedback. Can Fam Physician 2006;52:1570–1571. [PubMed: 17279238]

15. Adcock KA. Initiative to improve mammogram interpretation. Permanente Journal 2004;8:12–18.

16. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA 2001;285:2871–2879. [PubMed: 11401608]

17. Nass, S.; Balls, J. Improving breast imaging quality standards. Washington, DC: National Academy of Sciences; 2005.

18. American College of Radiology (ACR). ACR Breast Imaging Reporting and Data System: breast imaging atlas. Reston, VA: American College of Radiology; 2003. ACR BI-RADS–mammography.

19. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR 2001;177:543–549. [PubMed: 11517044]

20. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME guide no. 7. Med Teach 2006;28:117–128. [PubMed: 16707292]

21. Sickles EA. Auditing your breast imaging practice: an evidence-based approach. Semin Roentgenol 2007;42:211–217. [PubMed: 17919523]

22. Hadjianastassiou VG, Karadaglis D, Gavalas M. A comparison between different formats of educational feedback to junior doctors: a prospective pilot intervention study. J R Coll Surg Edinb 2001;46:354–357. [PubMed: 11768577]
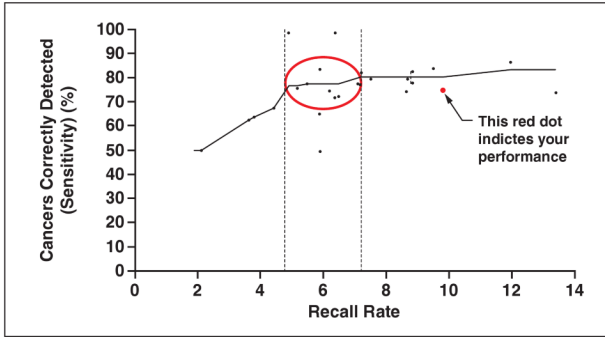
A
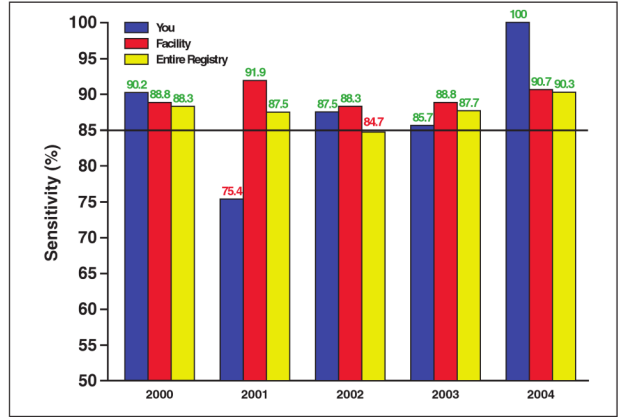
B

**Run Date: 04/09/2002**

**Annual Radiologist Report**

**Mammogram Dates: 01/01/1994 to 12/31/2000**

**Radiologist:**

| | Screening Mammograms | | | | All Mammograms | | | |
|---|---|---|---|---|---|---|---|---|
| | Radiologist | | State | | Radiologist | | State | |
| | # | % | # | % | # | % | # | % |
| Total mammograms | 4715 | 83.69 | 357007 | 85.90 | 5634 | 100.00 | 415593 | 100.00 |
| Negative | 3315 | 70.31 | 277248 | 77.66 | 3628 | 64.39 | 301926 | 72.65 |
| Negative w/benign findings | 577 | 12.24 | 42037 | 11.77 | 761 | 13.51 | 56709 | 13.65 |
| Probably benign | 291 | 5.96 | 13114 | 3.67 | 453 | 8.04 | 25233 | 6.07 |
| Suspicious | 222 | 4.71 | 3932 | 1.10 | 398 | 7.06 | 7782 | 1.87 |
| Malignant | 34 | 0.72 | 449 | 0.13 | 67 | 1.19 | 1153 | 0.28 |
| Incomplete assessment | 286 | 6.07 | 20227 | 5.67 | 327 | 5.80 | 22790 | 5.48 |
| Recall rate (starting 5/1/96) | | 10.99 | | 5.25 | | | | |
| Abnormal Interpretation rate (starting 5/1/96) | | 13.36 | | 5.79 | | | | |
| Total mammograms | 4758 | 88.67 | 353417 | 88.55 | 5366 | 100.00 | 399137 | 100.00 |
| True-negative | 4393 | 92.33 | 344007 | 97.34 | 4866 | 90.68 | 385261 | 96.52 |
| True-positive | 86 | 1.01 | 1512 | 0.43 | 127 | 2.37 | 2644 | 0.66 |
| False-negative | 11 | 0.23 | 619 | 0.18 | 17 | 0.32 | 933 | 0.23 |
| False-positive | 268 | 5.63 | 7279 | 2.06 | 356 | 6.63 | 10299 | 2.58 |
| Accuracy (TP + TN) / ALL | | 94.14 | | 97.77 | | 93.05 | | 97.19 |
| Sensitivity TP / (TP + FN) | | 88.66 | | 70.95 | | 88.19 | | 73.92 |
| Age: under 50 yrs | | 72.73 | | 64.11 | | 76.19 | | 70.81 |
| Age: 50 yrs and older | | 93.33 | | 72.75 | | 93.14 | | 74.97 |
| Specificity TN / (TN + FP) | | 94.25 | | 97.93 | | 93.18 | | 97.40 |
| Age: under 50 yrs | | 95.06 | | 97.65 | | 93.42 | | 96.82 |
| Age: 50 yrs and older | | 93.72 | | 98.07 | | 93.01 | | 97.71 |
| PPV (interpretation) | | 24.29 | | 17.20 | | 26.29 | | 20.43 |
| PPV (biopsy/cancer yield) | | 32.09 | | 30.95 | | 34.79 | | 34.96 |
| Biopsy rate | | 7.52 | | 2.55 | | | | |
| Cancer detection rate | | 18.07 | | 4.22 | | | | |
| Stage 0 + 1 cancers | 53 | 71.62 | 819 | 73.85 | 71 | 68.93 | 1166 | 62.19 |
| Minimal cancers | 44 | 54.32 | 711 | 52.78 | 57 | 49.14 | 956 | 42.32 |
| Node positive cancers | 14 | 25.45 | 198 | 23.19 | 22 | 27.50 | 506 | 32.65 |

C

| Assessment | Biopsy Results | |
|---|---|---|
| | Cancer (biospy comfirmed malignancy within 1 year) | Cancer (benign biospy or no cancer within 1 year) |
| Mammograms positive (BI-RADS 0, 4, 5) | TP = 8 (True-Positive) | FP = 68 (False-Positive) |
| Mammograms negative (BI-RADS 1, 2, 3) | FN = 2 (False-Negative) | TN = 992 (True-Negative) |

Sensitivity = TP / TP + FN = 8 / 10 = 80%
Specificity = TN / TN + FP = 992 / 1,060 = 93%
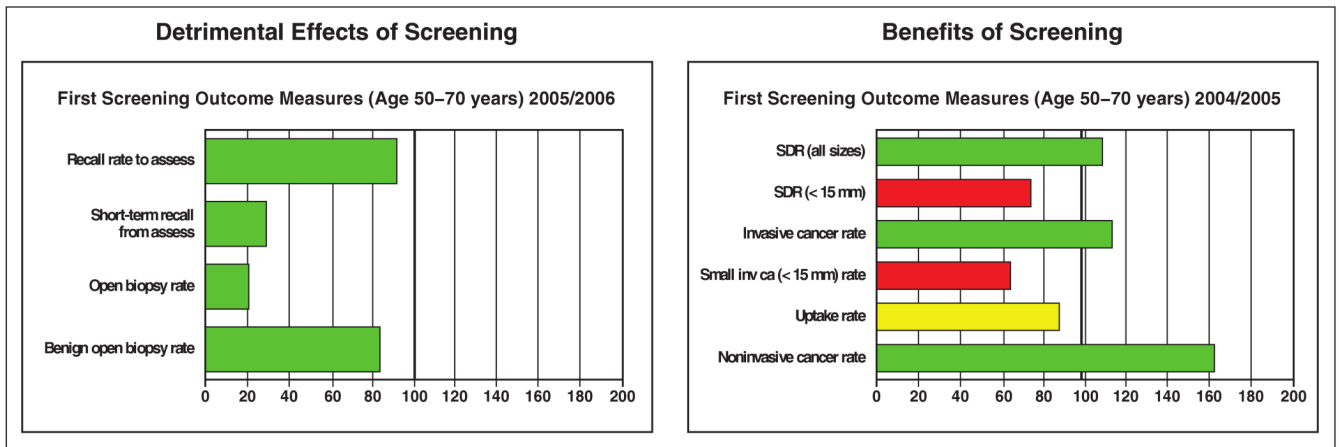PPV1 = TP / TP + FP = 8 / 100 = 11%

D

E

**Detrimental Effects of Screening**

First Screening Outcome Measures (Age 50–70 years) 2005/2006

**Benefits of Screening**

First Screening Outcome Measures (Age 50–70 years) 2004/2005

F

**Fig. 1. Examples of audit reports presented during focus groups and interviews**

**A**, Recall by sensitivity scattergram.

**B**, Vertical bar graph with benchmark (85%) shows sensitivity among screening examinations. A fourth vertical bar representing sensitivity data for Breast Cancer Surveillance Consortium was added to this figure in subsequent focus groups.

**C**, Single-page table. TP = true-positive, TN = true-negative, FP = false-positive, FN = false-negative, PPV = positive predictive value.

**D**, 2 × 2 table. PPV1 = percentage of all positive screening examinations that result in a cancer diagnosis within 1 year.

**E**, Color-coded chart.

**F**, Color-coded bar graph. Green or red indicates radiologist achieved or failed to achieve both minimum standard and target. Yellow indicates achieved minium standard but failed target. SDR = standardized detection ratio, inv ca = invasive cancer.

**Table 1**
**Performance Measures and Definitions Included in the Basic Clinical Relevant Mammography Audit of the American College of Radiology**

| Performance Measure | Definition |
| --- | --- |
| Recall rate | Percentage of screening examinations recommended for additional imaging (BI-RADS category 0) |
| Abnormal interpretation rate | Percentage of screening examinations interpreted as positive (BI-RADS categories 0, 4, and 5) |
| Recommendation for biopsy or surgical consultation | Number of recommendations for biopsy or surgical consultation (BI-RADS categories 4 and 5) |
| Known false-negatives | Diagnosis of cancer within 1 year of a negative mammogram (BI-RADS categories 1 or 2 for screening examination; BI-RADS categories 1, 2, or 3 for diagnostic examination) |
| Sensitivity | Probability of detecting a cancer when a cancer exists (true-positive examinations / all examinations with cancer detected within 1 year) |
| Specificity | Probability of interpreting an examination as negative when a cancer does not exist (true-negative examinations / all examinations without cancer detected in 1 year) |
| PPV1 | Percentage of all positive screening examinations that result in a cancer diagnosis within 1 year |
| PPV2 | Percentage of all screening or diagnostic examinations recommended for biopsy or surgical consultation that result in a cancer diagnosis within 1 year |
| PPV3 | Percentage of all biopsies done after a positive screening or diagnostic examination that result in a cancer diagnosis within 1 year |
| Cancer detection rate | Number of cancers correctly detected per 1,000 screening mammograms |
| Minimal cancers | Percentage of all cancers that are invasive and $\leq 1$ cm, or ductal carcinoma in situ |
| Node-negative cancers | Percentage of all invasive cancers that are node negative |

Note—PPV = positive predictive value.

**Table 2**
**Radiologists' Likes and Dislikes for Various Formats of Audit Reports**

| Format | Description | Likes | Dislikes |
|---|---|---|---|
| Recall by sensitivity scattergram | • Shows how sensitivity (on *y*-axis) increases with recall rate (on *x*-axis)<br><br>• Includes individual radiologists as single points and a regression line summarizing the relation between sensitivity and recall<br><br>• Highlights range of recall rates for which radiologists had the best sensitivity | • Shows how one performance measure relates to another<br><br>• Easy to compare performance with everyone else<br><br>• Visually easy to see individual performance<br><br>• Highlighted range gives a goal | • Graph was not intuitive and required explanation<br><br>• Could be confusing if it included a large number of radiologists |
| Vertical bar graph with benchmark | • Shows sensitivity (on *y*-axis) over time (past 5 years on *x*-axis) with different vertical bars for individual radiologists, practice facility, entire state, and BCSC[a]<br><br>• Includes single line that indicates 85% sensitivity as benchmark | • Easy to compare individual performance with facility or region<br><br>• Benchmark gives a goal, radiologists can easily see if they were over or under line<br><br>• Seeing data over time shows trends | • Multiple graphs take up space and cannot all be presented on a single page |
| Single-page table | • Includes all statistics outlined in BI-RADS audit<br><br>• Shows cumulative numbers and percentages for screening mammograms done by individual radiologists compared with entire state from time they started practicing mammography through past year<br><br>• Provides same statistics for all mammograms (screening and diagnostic) | • Cumulative numbers provide more stable estimates compared with single years<br><br>• Provides all measures on a single page<br><br>• Compares with state data | • A lot of numbers on one page<br><br>• Wanted to see screening and diagnostic examinations separately<br><br>• No comparisons to colleagues at one's own facility or over time |
| 2 × 2 table | • Shows positive and negative mammogram assessments on left and cancer diagnoses across top for a single year<br><br>• Includes numbers for true-positives, false-positives, false-negatives, and true-negatives<br><br>• Includes equations and calculations for sensitivity, specificity, and PPV | • Straightforward and easy to understand<br><br>• Like having sensitivity, specificity, and PPV calculated | • Radiologists would not calculate performance measures on their own<br><br>• No comparison with colleagues or over time<br><br>• No benchmarks |
| Color-coded chart | • Summarizes performance over 1 year across facilities on a single page | • Can be useful on a facility level or for a director | • Did not provide actual numbers |

| Format | Description | Likes | Dislikes |
|---|---|---|---|
| | • Performance is indicated with colored boxes instead of numbers (green = achieved target, yellow = achieved minimum standard but not target, red = failed minimum standard) | • Colors are easy to see | • Radiologists had no idea where they were within a group |
| Color-coded horizontal bar graph[b] | • Summarizes outcomes associated with detriments (recall and biopsy rates) and benefits (cancer detection rates) of screening mammography<br><br>• Color-coded bars show whether radiologists achieved target (green), achieved minimum standard but not target (yellow), or failed to reach minimum standard (red)<br><br>• Targets are displayed as over or under 100% | • Colors are easy to see | • 100% target was not meaningful |

Note—BCSC = Breast Cancer Surveillance Consortium, PPV = positive predictive value.

[a]Added in subsequent focus groups.

[b]Not shown to last two groups.