Berta M. Geller, EdD
William E. Barlow, PhD
Rachel Ballard-Barbash, MD, MPH
Virginia L. Ernster, PhD
Bonnie C. Yankaskas, PhD
Edward A. Sickles, MD
Patricia A. Carney, PhD
Mark B. Dignan, PhD
Robert D. Rosenberg, MD
Nicole Urban, ScD
Yingye Zheng, MS
Stephen H. Taplin, MD, MPH

Index terms:
Breast, ACR Reporting and Data System
Breast neoplasms, diagnosis, 00.30
Breast radiography, quality assurance, 00.11
Cancer screening, 00.11, 00.30

Published online before print
10.1148/radiol.2222010620
Radiology 2002; 222:536–542

Abbreviation:
BI-RADS = Breast Imaging Reporting and Data System

[1] From the Dept of Health Promotion Research, Univ of Vermont, 1 S Prospect St, Burlington, VT 05401-3444 (B.M.G.). Affiliations for all other authors and author contributions are listed at the end of this article. Received March 15, 2001; revision requested April 16; revision received June 18; accepted July 25. Supported by grants U01CA86076 (W.E.B.), U01CA86082 (P.A.C.), U01CA70013 (B.M.G.), U01CA63740 (E.A.S., V.L.E.), U01CA63731 (S.H.T.), U01CA70040 (B.C.Y.), and R01CA63146 (N.U.) from the National Cancer Institute. **Address correspondence to** B.M.G. (e-mail: *bgeller@zoo.uvm.edu*).

All opinions and findings are the sole responsibility of the authors. The views expressed do not necessarily represent those of the U.S. Government.

See also the article by Taplin et al (pp 529–535) in this issue.

Use of the American College of Radiology BI-RADS to Report on the Mammographic Evaluation of Women with Signs and Symptoms of Breast Disease[1]

PURPOSE: To examine whether mammographic assessments and recommendations are linked as expected, based on the Breast Imaging Reporting and Data System (BI-RADS), for the evaluation of women with signs and symptoms of breast disease.

MATERIALS AND METHODS: Eight mammography registries from the Breast Cancer Surveillance Consortium contributed mammographic data from 1996 through 1997 for women 25 years of age or older, with signs or symptoms of breast cancer. The association of assessments and recommendations and the relationship of self-reported symptoms to assessments are described.

RESULTS: A total of 51,673 diagnostic mammograms were included in the analyses and the expected management recommendation was provided 85%–90% of the time for mammograms classified as assessment categories 1, 2, 4, or 5. Category 3 ("probably benign finding") had the most variability in associated management recommendations, with only 40% (2,998 of 7,423) of cases associated with the recommendation for short interval follow-up. Of the 1,648 category 0 mammograms ("needs additional imaging") that did not have a final assessment, 64% were recommended for additional imaging, while another 20% of the cases were recommended for either a consultation or biopsy. The number of women who reported a lump as a symptom decreased with age but was associated with higher BI-RADS assessments.

CONCLUSION: BI-RADS assessment categories were generally used as intended for all categories but 0 and 3. Additional education about the use of these categories may be warranted. The inconsistencies between assessment category and management recommendations may present difficulties in conducting outcome audits.

Mammography is generally the first breast imaging procedure used in the assessment of women who present with breast signs or symptoms that may be indicative of cancer. The results of a mammogram need to be communicated to the referring physician in a consistent understandable format that includes the imaging findings, the probability of cancer, and the recommendation for the course of action. To assist with the communication of mammographic interpretation, the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) in 1992 to standardize reporting (1). We have combined registry data from seven geographic regions (eight registries) across the United States, which together form the Breast Cancer Surveillance Consortium (2), to describe the use of the BI-RADS for reporting the assessments and recommendations for a large number of women for whom the indication for mammography, as classified by the radiology facility, was "evaluation of a breast problem." To our knowledge, the effectiveness of using the BI-RADS for this subset (women with signs and symptoms of breast

disease) of diagnostic mammographic examinations has not been studied. Ultimately, evaluating and enhancing the use of the reporting system will assist in determining how breast imaging results contribute to an efficient and accurate diagnosis of breast cancer. The purpose of our study was to examine whether mammographic assessments and recommendations are linked as expected, based on the BI-RADS, for the evaluation of women with signs and symptoms of breast disease.

MATERIALS AND METHODS

Study Population

Data on mammograms performed from January 1996 through December 1997 from the eight mammography registries of the Breast Cancer Surveillance Consortium were pooled to conduct these analyses. The registries are located in Colorado, New Hampshire, New Mexico, North Carolina, San Francisco, Vermont, and western Washington state (two sites). Details of the registries and the Breast Cancer Surveillance Consortium can be found elsewhere (2). Data confidentiality procedures are described by Carney et al (3). In accordance with federal regulations protecting human subjects (4), registries received a waiver or alteration of informed consent from their institutional review board when they were approved to collect data from multiple participating mammography facilities within their geographic region.

We restricted the study to mammographic data in women aged 25 years or older for whom the indication for mammography, as classified by the radiology facility, was "evaluation of a breast problem." We excluded diagnostic mammograms that were classified by the radiology facility as either "further evaluation of an abnormal screening mammogram" or "short interval follow-up." Mammograms of women who self-reported previous breast cancer or breast implants also were excluded from this analysis. If a woman underwent more than one diagnostic mammogram in the 2-year study period, only the first mammogram was included in the analysis.

Data

Information was prospectively collected from patients, mammographic technologists, and radiologists who interpret mammograms by using standardized regional data collection forms and radiologists' reports. Data collected included patient demographic characteristics, relevant clinical history, indication for mammographic examination, and the radiologist's assessment and recommendations based on the mammographic examination. Five of the eight sites consistently collected patients' self-reported symptoms during the study period. For those sites, a general question was asked to elicit whether the woman had current symptoms, and more specific questions were asked about whether she had a lump or nipple discharge. Information on pain as a self-reported symptom was not collected routinely by all of the mammography registries. We were not able to distinguish between whether the lump was detected by means of a clinical breast examination (a sign) or a breast self examination (a symptom), so they were grouped together as symptoms. After encrypting women's, radiologists', and facilities identifiers, the data were transferred to a central statistical coordinating center for analysis.

Radiologic Assessment

The BI-RADS includes six assessment categories. Four of these categories include specific recommendations for further diagnostic evaluation according to the following: category 0, "needs additional imaging evaluation"; category 3, "short interval follow-up suggested"; category 4, "biopsy should be considered"; and category 5, "appropriate action should be taken." The remaining categories 1 and 2 imply a recommendation for repeating mammography at the routine screening interval. The Breast Cancer Surveillance Consortium used the standard BI-RADS reporting categories, but the assessment and recommendation were collected as individual measures (ie, they were not linked on the reporting form).

Radiologists used the following American College of Radiology BI-RADS six assessment categories: 0, "need additional imaging"; 1, "negative"; 2, "benign finding"; 3, "probably benign finding"; 4, "suspicious abnormality"; and 5, "highly suggestive of malignancy" (1). If a BI-RADS assessment category of 1–5 was assigned, we called it a "final" assessment. Our final assessment included mammograms that consisted of additional mammographic views or ultrasonography (US) on the same day of diagnostic mammogram. If the first diagnostic mammogram had a BI-RADS assessment category 0 ("need additional imaging"), we searched the database for the next mammogram or US examination performed within 90 days of the initial category 0 mammogram that had a BI-RADS assessment category of 1–5 and used that final assessment category. If no breast imaging subsequent to a category 0 mammogram was found, the mammogram was considered unresolved but was retained in the data.

We also collected follow-up data on whether additional mammographic or US imaging or biopsy was performed following diagnostic mammography. Data collected on these subsequent examinations or procedures include the type of procedure(s) performed but do not include a BI-RADS assessment category for the subsequent radiologic examinations. In this study, we searched these files to determine whether any additional imaging or biopsy was performed following an unresolved category 0 mammogram.

Where available, assessments are recorded according to breast (breast level), otherwise they are recorded according to woman (woman level). For women with a breast-level assessment, we computed a woman-level assessment by using the higher (more suspicious) BI-RADS assessment category of the two breast-level assessments. For the woman-level assessment, we used the following hierarchy of assessment categories: 1, 2, 3, 0, 4, or 5. We considered category 0 more suspicious than categories 1, 2, and 3 because it calls for an immediate diagnostic test.

As noted earlier, recommendations are recorded independently of assessments and categorized into one of the following five hierarchical groups: *(a)* normal interval follow-up, *(b)* short interval follow-up, *(c)* needs additional imaging (includes additional mammographic views, US, or magnetic resonance imaging) and other work-up not otherwise specified, *(d)* clinical examination or surgical consultation, or *(e)* biopsy or fine-needle aspiration. The highest level recommendation was used if there was more than one recommendation per mammogram.

Some breast imaging facilities in the mammography registries automatically link assessments to recommendations. This may be done according to center policy or may be performed by the software being used. Because these linkages did not allow us to examine the association between the assessment and recommendations, we excluded the data from radiology facilities with automatic linkage.

Final assessments and associated management recommendations recorded by interpreting radiologists were reported. We could not collect more detailed data

TABLE 1
Final Assessment of Diagnostic Mammograms by Age

Final Assessment Category	Age Category					Total
	25–39 Years	40–49 Years	50–59 Years	60–69 Years	70+ Years	
0 (incomplete [unresolved])	325 (3.2)	640 (3.6)	350 (3.3)	185 (2.8)	148 (2.4)	1,648 (3.2)
1 (negative)	6,519 (63.3)	9,414 (53.5)	5,453 (50.6)	3,474 (51.8)	3,212 (51.0)	28,072 (54.3)
2 (benign finding)	1,508 (14.7)	3,725 (21.2)	2,543 (23.6)	1,497 (22.3)	1,316 (20.9)	10,589 (20.5)
3 (probably benign finding)	1,505 (14.6)	2,809 (16.0)	1,725 (16.0)	1,009 (15.0)	868 (13.8)	7,916 (15.3)
4 (suspicious abnormality)	375 (3.6)	835 (4.7)	537 (5.0)	396 (5.9)	466 (7.4)	2,609 (5.1)
5 (highly suggestive of malignancy)	64 (0.6)	182 (1.0)	160 (1.5)	149 (2.2)	284 (4.5)	839 (1.6)
Total (denominator)	10,296 (19.9)	17,605 (34.1)	10,768 (20.8)	6,710 (13.0)	6,294 (12.2)	51,673 (100)

Note.—Data in parentheses are percentages.

TABLE 2
Association of Final Recommendation and Final Assessment of Diagnostic Mammograms

Final Recommendation	Category 1 (negative)	Category 2 (benign findings)	Category 3 (probably benign findings)	Category 4 (suspicious abnormality)	Category 5 (highly suggestive of malignancy)	All Categories
Normal interval	22,189 (85.5)*	8,751 (83.5)*	935 (12.6)	15 (0.6)	0 (0)	31,890 (67.6)
Short-term interval	464 (1.8)	573 (5.5)	2,998 (40.4)*	21 (0.8)	2 (0.2)	4,058 (8.6)
Additional imaging or other	859 (3.3)	441 (4.2)	1,993 (26.8)	340 (13.6)	82 (9.9)	3,715 (7.9)
Clinical examination or surgical consult	2,247 (8.7)*	604 (5.8)*	832 (11.2)	365 (14.6)*	202 (24.3)*	4,250 (9.0)
Biopsy or fine-needle aspiration	199 (0.8)	112 (1.1)	665 (9.0)	1,767 (70.5)*	545 (65.6)*	3,288 (7.0)
Total (denominator)	25,958 (55.0)	10,481 (22.2)	7,423 (15.7)	2,508 (5.3)	831 (1.8)	47,201 (100)

Note.—Data in parentheses are percentages. To be included in the percentage total, examinations had to have both a BI-RADS category of 1–5 and a known final recommendation. A total of 4,472 (8.65%) of the original 51,673 mammograms were excluded for missing either a final BI-RADS category (n = 1,454), or a final recommendation (n = 2,824), or both (n = 194). Included in the table are 72 cases where there was an inconsistency in laterality between the recommendation and the BI-RADS category (ie, the breast with the lower BI-RADS category had a more aggressive recommendation).

* Data represent the BI-RADS category suggested or logical associations of assessment with recommendation.

on cases of apparent inconsistency between assessment and management recommendation because we did not have access to the full text of all the interpretations rendered for such cases. We also provided a demographic description of the population including age, race, and education for the purposes of comparison with the general population undergoing mammography (two geographic sites and several facilities at other sites do not routinely collect educational data).

Analysis

Standard χ^2 tests were performed on the data in each table. Because of the large sample, even small differences across categories were statistically significant at $P < .001$. Therefore, we have not presented tests of statistical significance. As is often the case, these statistically significant differences in percentages may not be clinically important.

We report the proportion of women assigned to BI-RADS assessment categories according to age, recommendation, and reported symptoms.

RESULTS

A total of 59,321 diagnostic mammograms met the initial inclusion criteria. Of these, 1,338 (2.3%) were excluded because they were missing an assessment; an additional 6,310 (11%) were excluded because they were from facilities that automatically linked the assessments and recommendations. Therefore, 51,673 diagnostic mammograms were included in the analysis. The following racial information was reported for 40,065 (78%) women: 88%, white (n = 35,179); 8%, black (n = 3,240); 1.5%, Asian (n = 599); 0.8%, Native American (n = 323); and the remainder percentage was "other" or mixed race (n = 724). For the 73% of women who responded to a question about Hispanic origin, 4.7% (n = 1,375) said they were Hispanic. Of the 55% (n = 28,666) of women that reported educational status, 64% (n = 18,346) reported some post high school education.

Fifty-four percent of the women in this study were younger than 50 years old (Table 1). The percentage of women with mammograms assigned assessment categories 4 and 5 increased with age. After age 40, the use of category 0 decreased with each decade of age, and a decrease in the use of category 3 was noted at age 50. The use of a normal assessment (categories 1 and 2 combined) declined minimally with age; however, the use of the "negative" category declined, while use of the "benign finding" category increased with age.

Table 2 shows the association between the final assessment category and the management recommendation. The expected management recommendation was given 85%–90% of the time for mammograms classified as assessment categories 1, 2, 4, or 5. Mammograms assigned to categories 1 and 2 were recommended for normal interval follow-up 85.5% and 83.5% of the time, respectively. Ten per-

TABLE 3
Percentage of Women Reporting Symptoms and Total Number in Each Age and BI-RADS Assessment Classification

Patient Age (y)	1 (negative)	2 (benign findings)	3 (probably benign findings)	4 (suspicious abnormality)	5 (highly suggestive of malignancy)	All BI-RADS Categories
	Any Self-reported Symptoms before Mammogram					
25–44	4,900/6,117 (80.1)	1,508/1,923 (78.4)	1,028/1,226 (83.9)	334/379 (88.1)	74/80 (92.5)	7,844/9,725 (80.7)
45–54	2,560/3,606 (71.0)	1,402/2,049 (68.4)	660/890 (74.2)	230/294 (78.2)	109/113 (96.5)	4,961/6,952 (71.4)
55–64	1,271/1,990 (63.9)	582/993 (58.6)	263/386 (68.1)	101/136 (74.3)	74/84 (88.1)	2,291/3,589 (63.8)
65 or older	1,611/2,572 (62.6)	653/1,120 (58.3)	315/502 (62.8)	222/288 (77.1)	161/193 (83.4)	2,962/4,675 (63.4)
All ages	10,342/14,285 (72.4)	4,145/6,085 (68.1)	2,266/3,004 (75.4)	887/1,097 (80.9)	418/470 (88.9)	18,058/24,941 (72.4)
	Self-reported Breast Lump					
25–44	3,509/5,900 (59.5)	1,278/1,888 (67.7)	835/1,202 (69.5)	296/369 (80.2)	71/79 (89.9)	5,989/9,438 (63.5)
45–54	1,629/3,464 (47.0)	1,161/2,003 (58.0)	468/862 (54.3)	187/280 (66.8)	102/109 (93.6)	3,547/6,718 (52.8)
55–64	713/1,909 (37.4)	424/967 (43.9)	153/378 (40.5)	75/132 (56.8)	63/80 (78.8)	1,428/3,466 (41.2)
65 or older	769/2,453 (31.4)	379/1,077 (35.2)	170/488 (34.8)	174/275 (63.3)	141/181 (77.9)	1,633/4,474 (36.5)
All ages	6,620/13,726 (48.2)	3,242/5,935 (54.6)	1,626/2,930 (55.5)	732/1,056 (69.3)	377/449 (84.0)	12,597/24,096 (52.3)

Note.—Data represent the number of women who self-reported symptoms or a breast lump (numerator) out of the total number in that category and age group (denominator). Data in parentheses are percentages. There were 26,378 diagnostic mammograms at the five sites that reported symptoms. There were 1,437 women with unknown overall symptoms and 2,282 women for whom a response to the question about a breast lump was missing.

cent of the "negative" assessments and 7% of the "benign finding" assessments were recommended for either a clinical consultation or biopsy. For categories 4 and 5, 85% and 90%, respectively, had a recommendation for either a clinical consultation or biopsy. Most of the remaining women in categories 4 and 5 were recommended for additional imaging (13.6% and 9.9%, respectively). Fourteen percent of the mammograms had more than one management recommendation (not shown). Only 3,018 (5.8%) of 51,673 mammograms were excluded from this analysis because they were missing a recommendation.

Category 3 mammograms had the most variability in associated management recommendations. Forty percent of the category 3 mammograms had a short interval follow-up recommended, while 27% had additional imaging recommended. Category 0 also had inconsistent associated management recommendations. Of the 1,648 category 0 mammograms that did not have a final assessment, 64% (n = 1,055) were recommended for additional imaging, while another 20% (n = 330) were recommended for either a consultation or biopsy (data not shown). Within 90 days of the initial examination, 61% (n = 1,006) of these did have additional imaging or a biopsy or both, although no final BI-RADS assessment was provided (data not shown). Of the 61%, 191 (19%) were US examinations.

For the five mammography registries that collected symptom data, symptom information was available for 24,941 (94.6%) of the 26,378 diagnostic mammograms that had a final assessment. For mammograms for which the information was available, current breast symptoms were reported by 72.4% (18,058 of 24,941) of the women. Table 3 shows the percentage of women in each age and BI-RADS category that reported any current symptoms. For example, among women aged 25–44 years, 80.1% (4,900 of 6,117) of those in BI-RADS category 1 reported any symptoms versus 92.5% (74 of 80) in BI-RADS category 5. We also show the percentage according to age for all BI-RADS categories together. The percentage reporting any symptoms decreased with age from 80.7% (7,844 of 9,725) in women aged 25–44 years to 63.4% (2,962 of 4,675) in women aged 65 years or older. Current symptom reporting increased with increasing BI-RADS codes except "benign findings," which had a slightly smaller percentage with symptoms compared with "negative."

Almost all women (24,096 of 26,378) were asked about the presence of a breast lump. A breast lump was the most frequent symptom, reported by 52.3% (12,597 of 24,096) of the women (Table 3). This percentage also decreased with age from 63.5% (5,989 of 9,438) in women aged 25–44 years to 36.5% (1,633 of 4,474) in women aged 65 years or greater. Generally, the percentage of women reporting a lump increased with the final BI-RADS assessment, although Category 2 had a higher percentage than Category 3 for every age group except 25–44 years. Nipple discharge was reported by 7.6% (1,692 of 22,352) of women, although there did not appear to be a consistent relationship with the BI-RADS codes (data not shown). Overall, reporting of symptoms and, particularly, breast lumps, were associated with higher-numbered BI-RADS codes and younger age.

DISCUSSION

We believe these 1996–1997 data provide an important and sentinel baseline measurement of performance in community and academic practice of the use of the American College of Radiology BI-RADS for standardized reporting prior to the final regulations of the Mammography Quality Standards Act (MQSA) (5). We expect that current implementation of the final Federal regulations requiring the use of BI-RADS assessment categories in all mammography reports in the United States, as well as more widespread use of computer programs designed specifically to capture the BI-RADS assessment codes, will result in greater standardization in mammography reporting. These data can be compared in the future to data obtained after the MQSA final regulations went into effect in April 1999. The other major change that has occurred since 1997 was the publication of a third edition of the BI-RADS manual, which for the first time includes illustrations of mammographic findings and a detailed chapter on how to conduct an outcomes audit (1). If these improvements in explaining how to use BI-RADS result in more successful implementation

of the reporting system, they may also decrease the frequency of inconsistencies in mammography reports.

The Breast Cancer Surveillance Consortium completed a similar study on the use of BI-RADS in screening (asymptomatic) mammography (6). There were similarities and differences in the relationship between the assessments and resulting management recommendations between the current study and the study (6) with screening mammograms. Normal interval follow-up was recommended most of the time for mammograms assigned an assessment category of "negative" or "benign finding" in both studies. However, women in the screening study who were found to have a category 4 "suspicious abnormality" were much less likely to have a recommendation of clinical consultation or biopsy than were women who underwent evaluation of a breast problem in this analysis (64% and 85%, respectively). In both screening and diagnostic mammography, category 3 (probably benign finding) mammograms showed similar variation in the association to management recommendation. Our study also had findings similar to those of a study by the National Breast and Cervical Cancer Early Detection Program (NBCCEDP), which reported a slight decrease in the use of category 3 for patients aged 40 years and older (7). Our study showed a decline starting at age 50. The proportion of category 3 mammograms in the NBCCEDP study was much lower than in our study but was almost exactly the same when the NBCCEDP data were limited to women who had abnormal clinical breast examinations (15.8% in NBCCEDP, 15.3% in Breast Cancer Surveillance Consortium). Unlike our study, the NBCCEDP study showed a decrease in the use of combined categories 4 and 5 with increasing age. The two studies differ in that the NBCCEDP underwent mostly screening mammography, and our study was limited to women who presented with signs and symptoms of breast disease.

As shown in Table 2, 40.4% (2,998 of 7,423) of category 3 assessments were associated with the recommendations suggested in the American College of Radiology BI-RADS guidelines. However, unlike other assessment categories, the remaining 59.6% (4,425 of 7,423) were distributed across the other four management recommendation categories. By itself, category 3 represents an equivocal clinical assessment. In such circumstances, it is possible that radiologists may prefer using a more complete statement to describe the follow-up management they would recommend. The present data collection systems for this research effort did not have the capability of capturing text fields from the complete mammography report, largely because the necessary computer systems are not in use in all clinical practices. Therefore, we were unable to use a method of text field searching to examine whether recommendations were more consistent than we could observe with the standard recommendation categories. Berg et al (8), in a study of interobserver reliability, found that the highest disagreement between assessment and recommendation was for categories 3 and 4, and concluded that this was likely due to the variation in the radiologist's intervention threshold. This may help to explain the variation in management recommendation in our study. Sickles (9) suggests that category 3 should be assigned only following a complete imaging work-up. It seems likely that this category was frequently used before evaluation was complete, as 26.8% (1,993 of 7,423) of the women with category 3 mammograms were recommended to return for immediate additional imaging.

Category 0 appears to be used inconsistently. Some radiologists use category 0 even when imaging is complete and they are recommending biopsy. In the current study, 20% of the category 0 mammograms were associated with a recommendation for either a clinical consultation or a biopsy. This practice likely involves a clinical situation in which there are sufficiently abnormal mammographic findings to recommend biopsy, but additional imaging is useful for diagnostic and treatment evaluation. The BI-RADS guidelines suggest the use of the assessment category 0 mainly for screening examinations that require immediate additional imaging to resolve the inconclusive assessment of category 0. However, these data indicate that category 0 is also being used for diagnostic examinations, as does a recent article by Poplack et al (10). It may be helpful for the BI-RADS documentation to be more explicit about this area of clinical ambiguity. Perhaps, for this circumstance, it would be more consistent with the clinical process to suggest assigning assessment category 4 or 5 and recommending both biopsy and additional imaging for further diagnostic precision or positioning prior to biopsy.

If the likelihood of cancer is high when category 0 is used, then it might be important to include this assessment code in audits of outcomes of mammography. Audit data from a large-scale mammography practice report a 10.7% (121 of 1,132) rate of breast cancer among women assigned this assessment code for screening mammography (11). However, current federal regulations allow the use of category 0 with the expectation that all these cases will eventually be reassigned to a final assessment category (5). Therefore, the regulations indicate that assessment category 0 examinations need not be tracked for biopsy results. As a result, exclusion of category 0 cases that have no further radiologic follow-up will preclude identification of a fraction of cancer cases and subsequently bias the reporting of outcomes from a mammography practice. Therefore, radiologists who follow the practice of assigning category 0 for diagnostic mammographic examinations should be aware of the potential problem and the consequence that a detailed outcomes audit (one that identifies cancer cases by linkage to a regional tumor or pathology registry) may demonstrate a higher than expected percentage of false-negative interpretations.

Three percent of the mammograms in the current study were not assigned a final BI-RADS assessment. We were able to use our follow-up data to find that 1,006 (61%) of 1,648 of women whose mammography had no final assessment underwent either additional imaging or biopsy. If an US or another breast imaging procedure are the final imaging procedure following an initial mammography, BI-RADS does not provide a method for completing the assessment, hence, it is frequently left as an unresolved category 0 examination. In addition, some women choose to go directly undergo a biopsy instead of returning for the recommended additional imaging. Mammography in these women remains unresolved, as well.

Not all women classified by the radiology facility as presenting for evaluation of a breast problem reported current symptoms, such as pain or a breast lump. Women may not report symptoms because they are not aware of them or are not concerned about them. It is possible that physicians sometimes referred women for diagnostic mammography because physical findings were detected at clinical breast examination but were not discussed with the women. Almost 89% of the women who underwent mammographic assessment of category 5 findings, "highly suggestive of malignancy," reported current symptoms compared with 72.4% of the women whose mam-

mographic findings were categorized as "negative." Reporting current symptoms or a lump may be a good predictor for an abnormal mammogram. It is questionable whether there is value in asking about nipple discharge because this symptom is so rare and did not vary substantially across categories.

There are several limitations to our study. As discussed earlier, we did not have the ability to review the complete text of mammography reports to clarify the purpose and intent of an individualized diagnostic evaluation plan. This would have had great value in the interpretation of data that appeared inconsistent with standard recommendations for diagnostic evaluation of breast lesions. These data were collected during a period when the revolution in computer and information technology use was transforming communication and reporting. Although medical practice lags in its use of this technology, it is an opportune time for it to start taking advantage of this revolution.

Two other specific data limitations should be noted. Our follow-up data were incomplete when women went outside our catchment area for further care. It was possible that more than 61% of the incomplete assessments (1% of all the mammograms in this study) underwent follow-up, but our data systems did not capture it. Second, until 1998, women with Medicare received coverage for screening mammography every other year. Whether or not this influenced examination coding as "evaluation of a breast problem" as a means of obtaining Medicare reimbursement to pay for diagnostic mammography, which was covered without the every-other-year restriction, is unclear. If this occurred to a large degree we would expect to see this effect mostly among the negative and benign assessments (categories 1 and 2, respectively). However, our data suggest that an increase of self-reported symptoms is associated with an increase in the severity of the assessment for all women, including women aged 65 years and older. Medicare-eligible women appear to report current symptoms less frequently than do younger women, and there was a clear decline in reporting symptoms as women got older.

BI-RADS does not associate a management recommendation with assessment categories 1 and 2. Although it is reasonable to presume that the consistent recommendation in most cases should be repeat mammography at a normal interval, clinical consultation may be recommended for category 1 or 2 cases in the presence of palpable abnormalities not visible at mammography (12). It would be helpful if, in the future, the outcome audit system suggested in the BI-RADS manual would address whether it is preferable, for purposes of tracking positive cases, to consider such interpretations to be positive (because of the management recommendation for clinical consultation) or negative (because of the assessment as category 1 or 2).

The BI-RADS assessment categories were generally used as intended for all but the two categories 0 and 3. Both of these categories are associated with uncertainty. Additional education may be helpful to encourage that category 0 be used as intended, as a temporary incomplete assessment mainly for screening examinations, to be resolved with additional imaging. It would be useful to develop a consistent method to resolve category 0 assessments that would take into account the women who do not undergo additional mammographic views but rather directly undergo US or biopsy. At a minimum, there should be a uniform recommendation for how to treat incomplete assessments in outcome audits (ie, how to assign a final assessment by using other imaging results). Perhaps by using the free text search technology to review the inconsistent reports we will be able to identify patterns that will lead to algorithms to finalize and resolve these inconsistencies.

BI-RADS assessment category 3 (probably benign finding) is intended for findings that should be monitored at a shorter than normal interval to establish stability, with a very low expectation of finding cancer. Our findings suggest that radiologists may be using category 3 for purposes other than those originally designed. At least part of the problem may be that there is controversy among radiologists, due to limited observational data, on the proper timing of follow-up examination for surveillance of probably benign lesions (13).

In the section of BI-RADS devoted to "Follow-up and Outcome Monitoring," a positive diagnostic mammogram is defined as "one that requires a tissue diagnosis (BI-RADS category 4 and 5)," while a negative diagnostic mammogram is defined as "one that is negative, has a benign or probably benign finding (BI-RADS category 1, 2, and 3)," that is, one that does not require tissue diagnosis. However, as long as assigned assessment categories are not associated with the management recommendations, outcome audits will not be uniform. In this study, 11% of recommendations for biopsy and 26% of recommendations for clinical consultation were associated with one of the negative assessment categories. Whether one chooses the assessment category or the management recommendation to define a positive and negative mammogram will substantially affect outcome audits. The function of auditing is commonly conceptualized as requiring a dichotomous assessment categorization (ie, normal or abnormal). Medical diagnosis is rarely dichotomous. Uncertainty often continues for some time in the course of diagnostic evaluation. The apparently inconsistent nontrivial use of categories 0 and 3 by some radiologists reflects this medical reality. Outcome audits may need to be revised to incorporate the reality of uncertainty in medical diagnosis.

The American College of Radiology has published three editions of BI-RADS since 1993, with improvements in each edition. The inconsistent association between assessment category and recommendation suggest that it would be helpful in a future edition of BI-RADS to have more detailed descriptions of the intended use of assessment categories 0 and 3 and their associated management recommendations. Also, the addition of a consistent procedure to resolve category 0 diagnostic mammographic examinations after US or biopsy would produce more uniform outcome audits, thereby facilitating comparisons across facilities and regions. In the meantime, BI-RADS provides a uniform structure that will continue to aid in the evaluation of the accuracy of mammography.

Author affiliations: Departments of Biostatistics (W.E.B., Y.Z.) and Family Medicine (S.H.T.), University of Washington, Seattle; Applied Research Branch, CSRP, DCCPS, National Cancer Institute, Bethesda, Md (R.B.B.); Department of Epidemiology and Biostatistics, University of California, San Francisco (V.L.E.); Department of Radiology, University of North Carolina, Chapel Hill (B.C.Y.); Department of Radiology, University of California at San Francisco School of Medicine (E.A.S.); Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH (P.A.C.); AMC Cancer Research Center, Denver, Colo (M.B.D.); University of New Mexico Medical Center, New Mexico Tumor Registry, Albuquerque (R.D.R.); Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, Seattle, Wash (N.U.); and Center for Health Studies, Group Health Cooperative of Puget Sound, Seattle, Wash (S.H.T., W.E.B.).

Author contributions: Guarantors of integrity of entire study, B.M.G., W.E.B.; study concepts, all authors; study design, B.M.G., W.E.B., S.H.T., R.B.B.; literature research, B.M.G.; data acquisi-

tion, B.M.G., W.E.B., R.B.B., V.L.E., B.C.Y., P.A.C., M.B.D., R.D.R., N.U., S.H.T.; data analysis/interpretation, all authors; statistical analysis, W.E.B., Y.Z.; manuscript preparation, B.M.G.; manuscript definition of intellectual content, B.M.G., W.E.B., R.B.B., E.A.S., S.H.T.; manuscript editing, B.M.G., W.E.B., R.B.B., V.L.E.; manuscript revision/review, B.M.G., W.E.B., R.B.B., V.L.E., E.A.S., B.C.Y., P.A.C., M.B.D., R.D.R., N.U., S.H.T.; manuscript final version approval, all authors.

References

1. D'Orsi CJ, Bassett LW, Feig SA, et al. Illustrated Breast Imaging Reporting and Data System: illustrated BI-RADS. 3rd ed. Reston, Va: American College of Radiology, 1998.
2. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR Am J Roentgenol 1997; 169:1001–1008.
3. Carney PA, Geller BM, Moffett H, et al. Medico-legal issues and protective policies and procedures for data integrity and confidentiality in a large multi-center research program: the National Cancer Institute's Breast Cancer Surveillance Consortium. Am J Epidemiol 2000; 152:371–378.
4. Federal Policy for the Protection of Human Subjects (Source: 56 FR 28003, June 18, 1991).
5. Food and Drug Administration. Quality Mammography Standards. Final Rules-21 CFR Parts 16 and 900 [docket No. 95N-0192] RIN 0910-AA24 edition. Washington, DC: Department of Health and Human Services, 1997; 55925–55926.
6. Taplin SH, Ichikawa LE, Kerlikowske K, et al. Concordance of Breast Imaging Reporting and Data System Assessments and Management Recommendations in Screening Mammography. Radiology 2002; 222:529–535.
7. Caplan LS, Blackman D, Nadel M, Monticciolo DL. Coding mammograms using the classification "probably benign finding: short interval follow-up suggested." AJR Am J Roentgenol 1999; 172:339–342.
8. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000; 174:1769–1777.
9. Sickles E. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology 1994; 192: 439–442.
10. Poplack SP, Tosteson AN, Grove MR, Wells WA, Carney PA. Mammography in 53,803 women from the New Hampshire Mammography Network. Radiology 2000; 217:832–840.
11. Rosenberg R, Yankaskas B, Hunt W, et al. Effect of variations in operational definitions on performance estimates for screening mammography. Acad Radiol 2000; 7:1058–1068.
12. Patel MR, Gary JW. Negative mammograms in symptomatic patients with breast cancer. Acad Radiol 1998; 5:26–33.
13. Sickles EA, Rubin E. Probably benign breast lesions: when should follow-up be recommended and what is the optimal follow-up protocol? (editorial). Radiology 1999; 213:11–21.