Title
Effect of Radiologists' Diagnostic Work-up Volume on Interpretive Performance

Journal
Radiology, 273(2)

ISSN
0033-8419

Authors
Buist, Diana SM
Anderson, Melissa L
Smith, Robert A
et al.

Publication Date
2014-11-01

DOI
10.1148/radiol.14132806

Peer reviewed

Effect of Radiologists' Diagnostic Work-up Volume on Interpretive Performance[1]

Diana S. M. Buist, PhD, MPH
Melissa L. Anderson, MS
Robert A. Smith, PhD
Patricia A. Carney, PhD
Diana L. Miglioretti, PhD
Barbara S. Monsees, MD
Edward A. Sickles, MD
Stephen H. Taplin, MD, MPH
Berta M. Geller, EdD
Bonnie C. Yankaskas, PhD
Tracy L. Onega, PhD

Purpose: To examine radiologists' screening performance in relation to the number of diagnostic work-ups performed after abnormal findings are discovered at screening mammography by the same radiologist or by different radiologists.

Materials and Methods: In an institutional review board–approved HIPAA-compliant study, the authors linked 651 671 screening mammograms interpreted from 2002 to 2006 by 96 radiologists in the Breast Cancer Surveillance Consortium to cancer registries (standard of reference) to evaluate the performance of screening mammography (sensitivity, false-positive rate [FPR], and cancer detection rate [CDR]). Logistic regression was used to assess the association between the volume of recalled screening mammograms ("own" mammograms, where the radiologist who interpreted the diagnostic image was the same radiologist who had interpreted the screening image, and "any" mammograms, where the radiologist who interpreted the diagnostic image may or may not have been the radiologist who interpreted the screening image) and screening performance and whether the association between total annual volume and performance differed according to the volume of diagnostic work-up.

Results: Annually, 38% of radiologists performed the diagnostic work-up for 25 or fewer of their own recalled screening mammograms, 24% performed the work-up for 0–50, and 39% performed the work-up for more than 50. For the work-up of recalled screening mammograms from any radiologist, 24% of radiologists performed the work-up for 0–50 mammograms, 32% performed the work-up for 51–125, and 44% performed the work-up for more than 125. With increasing numbers of radiologist work-ups for their own recalled mammograms, the sensitivity (P = .039), FPR (P = .004), and CDR ($P < .001$) of screening mammography increased, yielding a stepped increase in women recalled per cancer detected from 17.4 for 25 or fewer mammograms to 24.6 for more than 50 mammograms. Increases in work-ups for any radiologist yielded significant increases in FPR (P = .011) and CDR (P = .001) and a nonsignificant increase in sensitivity (P = .15). Radiologists with a lower annual volume of any work-ups had consistently lower FPR, sensitivity, and CDR at all annual interpretive volumes.

Conclusion: These findings support the hypothesis that radiologists may improve their screening performance by performing the diagnostic work-up for their own recalled screening mammograms and directly receiving feedback afforded by means of the outcomes associated with their initial decision to recall. Arranging for radiologists to work up a minimum number of their own recalled cases could improve screening performance but would need systems to facilitate this workflow.

Online supplemental material is available for this article.

[1] From the Group Health Research Institute, Group Health Cooperative, 1730 Minor Ave, Suite 1600, Seattle, WA 98101 (D.S.M.B., M.L.A., D.L.M.); Cancer Control Science Department, American Cancer Society, Atlanta, Ga (R.A.S.); Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, Ore (P.A.C.); Department of Biostatistics, University of Washington School of Public Health, Seattle, Wash (D.L.M.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (B.S.M.); Department of Radiology, University of California, San Francisco, Calif (E.A.S.); Division of Cancer Control and Population Science, Behavioral Research Program, National Cancer Institute, Rockville, Md (S.H.T.); Department of Family Medicine, University of Vermont, College of Medicine, Burlington, Vt (B.M.G.); Department of Radiology, University of North Carolina, Chapel Hill, NC (B.C.Y.); and Department of Community and Family Medicine, Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH (T.L.O.). Received December 9, 2013; revision requested January 10, 2014; revision received March 24; accepted April 4; final version accepted April 18. This work was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274, SIRSG-07-275, SIRGS-06-281, SIRSG-09-270, SIRSG-09-271) and the Breast Cancer Stamp Fund. **Address correspondence to** D.S.M.B. (e-mail: *buist.d@ghc.org*).

A 2005 Institute of Medicine report (1) noted that the technical quality of mammography has improved since the 1992 Mammography Quality Standards Act but that optimal sensitivity and specificity have not been achieved—a conclusion reinforced by recent investigations (2). The Institute of Medicine report called for additional research on the relationship between interpretive volume and performance (1). Results on the association between mammography performance and volume, although inconsistent, generally suggest that higher-volume readers have lower false-positive rates (FPRs); findings on sensitivity are mixed (3). To address these gaps between observed and optimal screening accuracy (1), previous studies examined the relationship between interpretive volume and screening and diagnostic performance (4,5).

Contrary to the hypothesis suggested by the Institute of Medicine report that a higher interpretive volume would improve mammography performance, a study of a sample of U.S. radiologists found that volume did not explain much of the observed interradiologist variability in screening or diagnostic performance (4,5). The FPRs of radiologists with higher annual volumes were clinically and significantly lower than those of their lower-volume colleagues; however, the sensitivities were similar (4,5). Interpretive volume composition (ratio of screening volume relative to total volume) had the strongest influence on screening and diagnostic performance; a higher screening focus (ratio of screening to diagnostic mammograms) was associated with significantly lower screening sensitivity, cancer detection rate (CDR), and FPR (4,5), which suggests that having some element of diagnostic work-up could increase sensitivity and CDR. To our knowledge, only one study has examined whether radiologists' accuracy (defined as positive predictive value for biopsy recommendation) was influenced by monitoring a woman's images throughout the diagnostic process and found no significant influence (6). These findings indicate that interpretive volume alone is not the principal influence on performance; rather, volume might affect performance by allowing radiologists the opportunity to enhance their interpretative skills by performing work-up for diagnostic images that result from recalled screening mammograms interpreted by themselves or by other radiologists.

The purpose of this study was to to examine radiologists' screening performance in relation to the number of diagnostic examination work-ups after abnormal findings are discovered at screening mammography performed by the same radiologist or by different radiologists. In addition, we determined whether work-up of abnormal screening mammograms modified the association between annual interpretive volume and screening performance.

Advances in Knowledge

- Radiologists who interpreted a greater annual number of diagnostic mammograms that resulted from recall of screening mammograms they interpreted had consistently higher sensitivity (81.1% for 0–25 mammograms to 87.0% for >50 mammograms, $P = .039$) and cancer detection rates (CDRs) (3.1 per 1000 screening mammograms for 0–25 mammograms to 4.5 per 1000 screening mammograms for >50 mammograms, $P < .001$) than radiologists who interpreted fewer of these mammograms; however, the false-positive rate (FPR) was higher (6.7% for 0–25 mammograms to 10.3% for >50 mammograms, $P = .004$).
- These performance changes resulted in a stepped increase in the number of women recalled per cancer detected, ranging from 17.4 for radiologists who interpreted 25 or fewer of their recalled mammograms per year to 24.6 for radiologists who interpreted more than 50 of their recalled mammograms per year.
- Radiologists with a lower annual number of work-ups of recalled screening mammograms (0–50 mammograms vs >125 mammograms) had consistently lower FPRs (7.0% vs 10.3%, $P = .15$), sensitivity (80.8% vs 86.5%, $P = .011$), and CDRs (2.9 per 1000 vs 4.4 per 1000, $P = .001$) at all annual interpretive volumes.

Implication for Patient Care

- Arranging for radiologists to perform a minimum number of diagnostic work-ups that resulted from recall of screening mammograms they interpreted could improve screening mammography performance in the United States.

Materials and Methods

Subjects

The Breast Cancer Surveillance Consortium (BCSC) registries and Statistical

Published online before print
10.1148/radiol.14132806 **Content code:** BR

Radiology 2014; 273:351–364

Abbreviations:
BCSC = Breast Cancer Surveillance Consortium
BI-RADS = Breast Imaging Reporting and Data System
CDR = cancer detection rate
CI = confidence interval
DCIS = ductal carcinoma in situ
FPR = false-positive rate

Author contributions:
Guarantors of integrity of entire study, D.S.M.B., T.L.O.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, D.S.M.B., R.A.S., T.L.O.; clinical studies, D.S.M.B., D.L.M., B.C.Y., T.L.O.; experimental studies, D.S.M.B., B.S.M.; statistical analysis, D.S.M.B., M.L.A., R.A.S., T.L.O.; and manuscript editing, all authors

Funding:
This research was supported by the National Cancer Institute Breast Cancer Surveillance Consortium (U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01CA69976, U01CA63731, U01CA63736, U01CA70040, HHSN261201100031C). Portions of the data collection were supported by the National Cancer Institute and Agency for Healthcare Research and Quality (R01 CA107623) and the National Cancer Institute (K05 CA104699).

Conflicts of interest are listed at the end of this article.

Coordinating Center received institutional review board approval for active or passive consenting processes and a Federal Certificate of Confidentiality and other protections for participating women, physicians, and facilities. All procedures were compliant with the Health Insurance Portability and Accountability Act (7).

Data Collection

BCSC registries collect information about mammography performed at participating facilities in their defined catchment areas and link this information to state tumor registries or regional Surveillance Epidemiology and End Results programs to obtain population-based cancer data (8,9). Demographic and breast cancer risk factor data, including age, first-degree family history, and time since last mammographic examination, are collected with use of a self-reported questionnaire completed at each screening mammography examination. This study included data from six BCSC mammography registries (in California, North Carolina, New Hampshire, Vermont, Washington, and New Mexico). Because planned analysis required complete capture of all screening and diagnostic images for each radiologist, we limited our sample to radiologists who interpreted mammograms only in BCSC facilities (436 reader-years, 106 radiologists) (4,5). Eligible radiologists from the six registries who interpreted screening mammograms from 2005 to 2006 were invited to complete a self-administered mailed survey between January 1, 2006, and September 30, 2007 (10), and survey information was linked to BCSC data. We excluded 10 radiologists (43 reader-years) who interpreted mammograms at facilities with incomplete BCSC data on diagnostic mammograms during the study years. The radiologists and reader-years included in these analyses are a subset of those previously reported (5).

The two primary exposures of interest were "own" work-ups and "any" work-ups. The measurement of "own" started at a recalled screening examination and determined the number of those recalled screening mammograms with a diagnostic work-up (mammography with or without ultrasonography [US], counted as one examination) within 60 days (11) interpreted by the same radiologist who recalled the screening mammogram. The measurement of "any" started with the interpretation of any diagnostic mammogram (with or without US, counted as one examination), regardless of which radiologist recalled the screening mammogram. We followed the Breast Imaging Reporting and Data System (BI-RADS) lexicon and collected one overall assessment for diagnostic examinations with or without US (12,13). Because any work-ups included all diagnostic mammograms obtained for work-up of a positive screening examination, unlike the own work-ups, we did not require linkage to the recalled screening examination or that the diagnostic work-up be performed within 60 days of the screening examination. The two exposure measures overlap. For example, a work-up would be counted as both own and any if the same radiologist recalled the screening mammogram and interpreted the diagnostic follow-up mammogram within 60 days of the screening examination. Therefore, most diagnostic follow-up mammograms classified as own work-ups were also included as any work-ups (except when the only diagnostic follow-up was US).

Annual interpretive volume for 2001–2005 was collected and summed across all facilities for total, screening, and diagnostic volumes. Examination type was defined by using radiologists' indications for examinations (5). Diagnostic examinations included additional evaluation of a previous abnormal screening mammogram, short-interval follow-up, or evaluation of a breast symptom or mammographic abnormality with or without US.

Screening performance was based on the radiologist's initial assessment (positive or negative) of the screening mammogram linked to invasive carcinoma or ductal carcinoma in situ (DCIS) diagnoses collected from tumor registries and pathology databases and diagnosed within the follow-up period (1 year after the screening mammogram and before the next screening mammography examination) (13). Registry data were used to characterize the tumors with regard to histologic characteristic (DCIS vs invasive), stage (0–IV), tumor size, axillary lymph node involvement (negative or positive), grade (well differentiated to undifferentiated), and estrogen receptor status. We defined minimal detected and early stage cancers in three ways, as follows: *(a)* DCIS or invasive cancer 10 mm or smaller (12), *(b)* DCIS or invasive cancer smaller than 15 mm and node negative (4,5), or *(c)* DCIS or invasive cancer 10 mm or smaller and node negative (4,5).

Performance measures (sensitivity, FPR, and CDR) were derived from 651 671 screening mammograms (404 538 unique women, asymptomatic subjects with routine screening indication) interpreted from 2002 to 2006 by using standard BI-RADS and BCSC definitions (5). The mammograms and unique women reported herein are a subset of those previously reported (5). Sensitivity was defined as the proportion of screening mammograms interpreted as positive (defined as BI-RADS categories 0 [needs additional assessment], 4 [suspicious abnormality], 5 [highly suggestive of malignancy], or 3 [probably benign when associated with a recommendation for immediate follow-up, ie, more imaging, clinical examination, biopsy]) (12) diagnosed within the follow-up period. The FPR was defined as the proportion of positive screening examinations among all women without a breast cancer diagnosis within the follow-up period. The CDR was defined as the number of cancers detected within the follow-up period per 1000 screening mammograms interpreted.

Statistical Analysis

The two work-up volume measures (own and any) and annual total interpretive volume measures for each year were linked to screening performance in the following year (eg, 2005 volume was linked to 2006 performance). The Pearson correlation coefficient was used to estimate the strength of the linear relationship between the continuous measures of work-up volume.

Two breast imaging specialists (E.A.S. and B.S.M., with 38 and 34 years of experience, respectively) evaluated the data with all coauthors to assess feasibility for measurement and implementation and classified volumes as low (<25 mammograms), medium (26–50 mammograms), and high (>50 mammograms) and any work-up into low (<50 mammograms), medium (51–125 mammograms), and high (>125 mammograms). We calculated unadjusted screening performance by using these categoric volume diagnostic work-up measures. To assess the potential trade-off between sensitivity and FPR, we calculated the number of women recalled for each cancer detected. All *P* values are two sided.

Because a radiologist's case-mix distribution (average age and screening intervals) might have an effect on results, we computed adjusted performance measures by using internal standardization (14) to account for differences in radiologists' case-mix distributions (5). Internal standardization works by reweighting mammograms according to the relative difference between the radiologist's specific distribution of potential confounders (age and time since last mammography examination) and the corresponding distribution in the overall analytic sample. This process enables calculation of performance measures for radiologists as if their case mixes were the same as that in the overall population. To assess the relationship between the continuous work-up measures and adjusted performance, we stratified according to cancer status, fitting separate models for each performance measure by using the radiologist's initial mammographic assessment (positive or negative) as the binary outcome variable. Continuous diagnostic work-up measures were included in the regression models by using restricted cubic smoothing splines (15) to allow for nonlinearity and to permit a flexible shape for the relationship between the continuous volume measure and interpretive performance. We fit logistic models by using generalized estimating equations with robust standard errors to account for correlation between multiple observations from the same radiologist. Because the diagnostic work-up measures were heavily skewed with sparse data in high volumes, we restricted the range before fitting the models to ensure stable estimates of model parameters; therefore, model estimation excluded outliers (radiologists with >250 own and >600 any recalled mammograms). Model results are presented graphically with 95% confidence intervals (CIs), with the curves interpreted directly as the mean adjusted performance as a function of the exposure measure. *P* values for the estimated curves correspond to omnibus tests of whether there is any relationship between mean adjusted performance and work-up volume.

Similar methods were used to test the hypothesis that the relationship between total interpretive volume and screening performance is different for radiologists with low versus high volumes of diagnostic work-ups. Logistic regression models with cubic smoothing splines were used to estimate performance as a function of total interpretive volume. We restricted the range of total volume to 6000 or fewer mammograms because of sparse data in the tails. Interaction terms were included in the model to estimate separate curves for low and high levels of diagnostic work-up. *P* values correspond to omnibus tests of whether there is a difference in the shape (interaction term to assess effect modification) of the volume-performance relationship for radiologists with low versus high volumes of work-ups of recalled screening mammograms. Model results are presented graphically, with separate curves for low and high diagnostic work-up volume; the curves are the mean adjusted performance as a function of the total annual volume.

Statistical analyses were performed with software (SAS version 9.2, SAS Institute, Cary, NC [16], and Stata version 12.0, StataCorp, College Station, Tex [17]).

Results

The 96 radiologists in the study had a median age of 53 years (range, 37–72 years). Most radiologists worked full time (76%), had at least 20 years of experience (53%), and did not have fellowship training in breast imaging (95%) (Table 1). Time spent on breast imaging varied and was less than 20% for 24% of radiologists and 80%–100% for 32%. Thirty-eight percent of radiologists worked up 25 or fewer of their own recalled screening mammograms a year, 24% worked up 0–50, and 39% worked up more than 50. Twenty-four percent of radiologists worked up 0–50 of any recalled screening mammograms, 32% worked up 51–125, and 44% worked up more than 125.

Radiologists who performed work-up for a greater number of own or any recalled screening mammograms were more likely to have completed fellowship training, have greater annual interpretive volumes, and spend more than 40% of their time on breast imaging (Table 1). Associations between working up own and any recalled mammograms with volume (total and diagnostic) were similar, with higher-volume readers interpreting more own and any recalled screening mammograms. The work-up of own and any mammograms showed a positive correlation (Pearson correlation coefficient = 0.49, $P < .01$) (Fig E1 [online]).

The low-, medium-, and high-volume categories for the work-up of own recalled mammograms included 25%, 21%, and 53% of the screening mammograms used to calculate screening performance; the low-, medium-, and high-volume categories for the work-up of any recalled mammograms included 13%, 27%, and 60% of the screening mammograms used to calculate screening performance (Table 2). The characteristics of the women according to age, first-degree family history, or time since last mammographic examination did not differ according to low-, medium-, or high-volume category for either exposure measure (ie, own or any work-up). Most screening mammograms included in the performance outcome measures were obtained in women aged 40–59 years (60%), with 3% obtained in women younger than 40 years and 5% in women aged at

Table 1

Characteristics of Radiologists according to Average Annual Volume of Work-up of Own and Any Recalled Screening Mammograms

Characteristic	Total (*n* = 96)	Average Annual Work-up of Own Recalled Mammograms			Average Annual Work-up of Any Recalled Mammograms		
		0–25 Mammograms (*n* = 36)	26–50 Mammograms (*n* = 23)	>50 Mammograms (*n* = 37)	0–50 Mammograms (*n* = 23)	51–125 Mammograms (*n* = 31)	>125 Mammograms (*n* = 42)
Average annual work-up of own recalled mammograms							
0–25 mammograms	36 (38)	…	…	…	13 (57)	15 (48)	8 (19)
26–50 mammograms	23 (24)	…	…	…	5 (22)	5 (16)	13 (31)
>50 mammograms	37 (39)	…	…	…	5 (22)	11 (35)	21 (50)
Average annual work-up of any recalled mammograms							
0–50 mammograms	23 (24)	13 (36)	5 (22)	5 (14)	…	…	…
51–125 mammograms	31 (32)	15 (42)	5 (22)	11 (30)	…	…	…
>125 mammograms	42 (44)	8 (22)	13 (57)	21 (57)	…	…	…
Age at survey							
<45 y	24 (25)	10 (28)	5 (22)	9 (24)	7 (30)	9 (29)	8 (19)
45–54 y	29 (30)	10 (28)	7 (30)	12 (32)	4 (17)	6 (19)	19 (45)
≥55 y	43 (45)	16 (44)	11 (48)	16 (43)	12 (52)	16 (52)	15 (36)
Sex							
M	65 (68)	26 (72)	16 (70)	23 (62)	19 (83)	24 (77)	22 (52)
F	31 (32)	10 (28)	7 (30)	14 (38)	4 (17)	7 (23)	20 (48)
Works full time (≥40 h/wk)							
No	23 (24)	6 (17)	8 (36)	9 (25)	6 (26)	4 (13)	13 (32)
Yes	71 (76)	30 (83)	14 (64)	27 (75)	17 (74)	26 (87)	28 (68)
Primary affiliation with academic medical center							
No affiliation	69 (73)	27 (75)	16 (70)	26 (73)	16 (70)	23 (74)	30 (73)
Adjunct	7 (7)	2 (6)	2 (9)	3 (8)	1 (4)	2 (6)	4 (10)
Primary	19 (20)	7 (19)	5 (22)	7 (19)	6 (26)	6 (19)	7 (17)
Experience							
Years since graduated residency							
<10	15 (16)	8 (22)	3 (13)	4 (11)	3 (13)	8 (26)	4 (10)
10–19	30 (32)	11 (31)	8 (35)	11 (31)	5 (22)	9 (29)	16 (39)
≥20	50 (53)	17 (47)	12 (52)	21 (58)	15 (65)	14 (45)	21 (51)
Combined variable of fellowship training and years of experience in mammography interpretation							
No fellowship, <10 y	17 (18)	9 (25)	3 (13)	5 (14)	5 (22)	6 (19)	6 (14)
No fellowship, 10–19 y	32 (33)	12 (33)	9 (39)	11 (30)	2 (9)	14 (45)	16 (38)
No fellowship, ≥20 y	42 (44)	15 (42)	11 (48)	16 (43)	15 (65)	11 (36)	16 (38)
Fellowship, <10 y	1 (1)	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	1 (2)
Fellowship, ≥10 y	4 (4)	0 (0)	0 (0)	4 (11)	1 (4)	0 (0)	3 (7)
Time working in breast imaging							
<20%	22 (24)	11 (31)	3 (14)	8 (23)	7 (33)	10 (34)	5 (12)
20%–39%	25 (27)	10 (29)	10 (45)	5 (14)	3 (14)	12 (41)	10 (24)
40%–79%	16 (17)	3 (9)	4 (18)	9 (26)	3 (14)	3 (10)	10 (24)
80%–100%	29 (32)	11 (31)	5 (23)	13 (37)	8 (38)	4 (14)	17 (40)
Interpretive volume							
Average annual total volume							
480–999 mammograms	16 (17)	12 (33)	2 (9)	2 (5)	10 (43)	5 (16)	1 (2)

Table 1 (continues)

Table 1 (continued)

Characteristics of Radiologists according to Average Annual Volume of Work-up of Own and Any Recalled Screening Mammograms

Characteristic	Total (*n* = 96)	Average Annual Work-up of Own Recalled Mammograms			Average Annual Work-up of Any Recalled Mammograms		
		0–25 Mammograms (*n* = 36)	26–50 Mammograms (*n* = 23)	>50 Mammograms (*n* = 37)	0–50 Mammograms (*n* = 23)	51–125 Mammograms (*n* = 31)	>125 Mammograms (*n* = 42)
1000–1499 mammograms	20 (21)	9 (25)	4 (17)	7 (19)	7 (30)	9 (29)	4 (10)
1500–1999 mammograms	16 (17)	7 (19)	5 (22)	4 (11)	4 (17)	5 (16)	7 (17)
2000–2999 mammograms	23 (24)	6 (17)	9 (39)	8 (22)	0 (0)	9 (29)	14 (33)
3000–4999 mammograms	12 (13)	1 (3)	2 (9)	9 (24)	2 (9)	2 (6)	8 (19)
≥5000 mammograms	9 (9)	1 (3)	1 (4)	7 (19)	0 (0)	1 (3)	8 (19)
Average annual screening volume							
480–999 mammograms	24 (25)	16 (44)	3 (13)	5 (14)	13 (57)	7 (23)	4 (10)
1000–1499 mammograms	17 (18)	6 (17)	5 (22)	6 (16)	4 (17)	9 (29)	4 (10)
1500–1999 mammograms	26 (27)	9 (25)	10 (43)	7 (19)	4 (17)	9 (29)	13 (31)
2000–2999 mammograms	13 (14)	3 (8)	4 (17)	6 (16)	1 (4)	3 (10)	9 (21)
≥3000 mammograms	16 (17)	2 (6)	1 (4)	13 (35)	1 (4)	3 (10)	12 (29)
Average annual diagnostic volume							
<100 mammograms	11 (11)	9 (25)	2 (9)	0 (0)	11 (48)	0 (0)	0 (0)
100–199 mammograms	11 (11)	6 (17)	3 (13)	2 (5)	5 (22)	6 (19)	0 (0)
200–299 mammograms	24 (25)	11 (31)	3 (13)	10 (27)	6 (26)	13 (42)	5 (12)
300–499 mammograms	33 (34)	10 (28)	13 (57)	10 (27)	1 (4)	12 (39)	20 (48)
500–999 mammograms	8 (8)	0 (0)	1 (4)	7 (19)	0 (0)	0 (0)	8 (19)
≥1000 mammograms	9 (9)	0 (0)	1 (4)	8 (22)	0 (0)	0 (0)	9 (21)
Average annual percentage of examinations that are screening examinations							
<75%	6 (6)	2 (6)	1 (4)	3 (8)	0 (0)	1 (3)	5 (12)
75%–79%	13 (14)	3 (8)	4 (17)	6 (16)	0 (0)	2 (6)	11 (26)
80%–84%	35 (36)	13 (36)	9 (39)	13 (35)	1 (4)	14 (45)	20 (48)
85%–89%	21 (22)	6 (17)	5 (22)	10 (27)	8 (35)	7 (23)	6 (14)
≥90%	21 (22)	12 (33)	4 (17)	5 (14)	14 (61)	7 (23)	0 (0)
Average annual number of facilities where interpreting							
1	31 (32)	8 (22)	6 (26)	17 (46)	9 (39)	15 (48)	7 (17)
>1 to 2	39 (41)	15 (42)	11 (48)	13 (35)	11 (48)	10 (32)	18 (43)
>2 to 3	16 (17)	8 (22)	4 (17)	4 (11)	3 (13)	6 (19)	7 (17)
>3	10 (10)	5 (14)	2 (9)	3 (8)	0 (0)	0 (0)	10 (24)

Note.—Data are numbers of radiologists, with column percentages in parentheses.

least 80 years. The characteristics of women who had their mammograms interpreted at academic medical centers were not different from those of women whose mammograms were interpreted at nonacademic facilities (Table E1 [online]).

There were 3101 cancers in the study population; 2646 were detected with screening mammography. Among invasive cancers, stage distribution and median tumor size did not vary according to either exposure measure (own or any work-up) (Table 3). Of 455 interval-detected cancers, 89% were invasive cancers with a larger median size (19 mm) and a higher fraction (25%) were estrogen receptor–negative compared with screening-detected cancers (Table E2 [online]).

The unadjusted mean sensitivity was 85.3% (95% CI: 83.6%, 86.9%), the FPR was 9.1% (95% CI: 8.0%, 10.3%), and the CDR was 4.1 per 1000 screening mammograms (95% CI: 3.7, 4.5) (Table 4). As the number of own work-ups increased, the adjusted sensitivity, FPR, and CDR significantly increased (Fig 1), yielding a stepped increase in the number of women recalled per cancer detected from 17.4 for 25 or fewer mammograms to 24.6 for more than 50 mammograms (Table 4). Improved sensitivity and CDR were accompanied by an increase in the FPR with each category of own work-ups, which was consistent with the figures showing little improvement for volumes of more than 50 own work-ups or more

Table 2

Characteristics of Women Whose Mammograms Were Used to Calculate Screening Performance Measures according to Annual Volume of Own and Any Recalled Screening Mammograms

Parameter	Total (n = 651 671)	Average Annual Volume of Own Recalled Mammograms			Average Annual Volume of Any Recalled Mammograms		
		0–25 Mammograms (n = 164 834)	26–50 Mammograms (n = 139 244)	>50 Mammograms (n = 347 593)	0–50 Mammograms (n = 83 326)	51–125 Mammograms (n = 178 766)	>125 Mammograms (n = 389 579)
Age at screening mammography							
<40 y	19 734 (3)	5291 (3)	3566 (3)	10 877 (3)	2951 (4)	5584 (3)	11 199 (3)
40–49 y	183 967 (28)	46 148 (28)	38 422 (28)	99 397 (29)	24 258 (29)	48 034 (27)	111 675 (29)
50–59 y	205 916 (32)	51 849 (32)	45 142 (32)	108 925 (31)	25 531 (31)	55 102 (31)	125 283 (32)
60–69 y	128 089 (20)	33 003 (20)	27 304 (20)	67 782 (20)	15 998 (19)	37 665 (21)	74 426 (19)
70–79 y	83 566 (13)	21 250 (13)	18 137 (13)	44 179 (13)	10 854 (13)	24 410 (14)	48 302 (12)
≥80 y	30 399 (5)	7293 (4)	6673 (5)	16 433 (5)	3734 (4)	7971 (4)	18 694 (5)
First-degree family history							
No	487 338 (84)	129 070 (78)	94 037 (68)	264 231 (76)	57 378 (69)	143 974 (81)	285 986 (73)
Yes	95 606 (16)	23 315 (14)	20 792 (15)	51 499 (15)	11 047 (13)	24 447 (14)	60 112 (15)
Unknown	68 727 [11]	12 449 [8]	24 415 [17]	31 863 [9]	14 901 [18]	10 345 [6]	43 481 [11]
Time since last mammography							
No previous mammography	28 267 (5)	7379 (5)	6526 (5)	14 362 (4)	4045 (5)	8709 (5)	15 513 (4)
<2 y	529 614 (85)	114 178 (81)	114 178 (82)	281 802 (81)	67 988 (82)	142 019 (79)	319 607 (82)
3–4 y	39 880 (6)	8771 (6)	8771 (6)	20 615 (6)	5552 (7)	11 743 (7)	22 585 (6)
≥5 y	22 975 (4)	5032 (4)	5032 (4)	11 792 (3)	3086 (4)	6789 (4)	13 100 (3)
Unknown	30 935 [5]	4737 [4]	4737 [3]	19 022 [6]	2655 [3]	9506 [5]	18 774 [5]

Note.—Data are numbers of women, with column percentages in parentheses. Percentages in brackets (for unknown variables) were not included in total column percentages.

than 125 any work-ups (Fig 2). The one exception was CDR, where CDR was significantly (P = .039) reduced with increasing annual volume for radiologists who interpreted fewer than 50 of their own recalled mammograms.

Unadjusted sensitivity increased from 80.8% for radiologists with 50 or fewer any work-ups to 86.5% for those with more than 125 any work-ups (Table 4); however, the association between adjusted sensitivity and volume of any work-up was not significant (P = .15) (Fig 1d). An increase in the volume of any work-up yielded statistically significant increases in the FPR and CDR (Fig 1e, 1f).

Overall, 22.2 women out of 1000 were recalled for each cancer detected. The lowest number of women recalled per cancer detected was among radiologists who worked up the fewest numbers of own and any recalled mammograms; however, these radiologists also had the lowest sensitivity. Radiologists with the highest sensitivity and CDR and the lowest FPR had worked up more than 25 of own recalled mammograms or more than 50 of any recalled mammograms.

In general, the shape of the relationship between total interpretive volume and screening performance did not differ according to a low versus high volume of diagnostic follow-up (Fig 2). However, readers with fewer own or any work-ups had consistently lower sensitivity, FPR, and CDR at any given total volume. The stratified analysis also showed decreased FPRs with increasing total annual volume to a threshold of 2000.

Discussion

We found that radiologists with a higher annual volume of work-ups for recalled screening mammograms they initially interpreted had consistently higher screening sensitivities and CDRs; however, these performance improvements were accompanied by higher FPRs. We expected that a higher volume of diagnostic work-ups for a radiologist who interpreted the screening mammogram would be associated with better screening performance because of the radiologist's involvement throughout a case, possibly including interventional procedures (18,19). This constitutes direct feedback on the radiologist's clinical decisions. Performing analysis with continuous measures and accounting for potential confounders resulted in improved sensitivity for radiologists who annually work up diagnostic examinations resulting from at least 50 of their own recalled mammograms and a higher CDR for radiologists who

Table 3

Characteristics of Screening-detected Tumors according to Average Annual Volume of Own and Any Recalled Screening Mammograms

Characteristic	Total (*n* = 2646)	Average Annual Volume of Own Recalled Screening Mammograms			Average Annual Volume of Any Recalled Screening Mammograms		
		0–25 Mammograms (*n* = 467)	26–50 Mammograms (*n* = 611)	>50 Mammograms (*n* = 1568)	0–50 Mammograms (*n* = 251)	51–125 Mammograms (*n* = 653)	>125 Mammograms (*n* = 1742
Cancer histologic type							
DCIS	681 (26)	119 (25)	149 (24)	413 (26)	72 (29)	151 (23)	458 (26)
All invasive	1961 (74)	348 (75)	460 (76)	1153 (74)	178 (71)	501 (77)	1282 (74)
Unknown	4 [0]	0 [0]	2 [0]	2 [0]	1 [0]	1 [0]	2 [0]
Stage							
0	681 (27)	119 (27)	149 (25)	413 (28)	72 (30)	151 (25)	458 (27)
I	1151 (45)	208 (46)	290 (49)	653 (44)	102 (43)	272 (45)	777 (46)
II	558 (22)	97 (22)	127 (21)	334 (22)	41 (17)	136 (22)	381 (23)
III	136 (5)	20 (4)	30 (5)	86 (6)	22 (9)	42 (7)	72 (4)
IV	16 (1)	4 (1)	1 (0)	11 (1)	1 (0)	10 (2)	5 (0)
Unknown	104 (4)	19 (4)	14 (2)	71 (5)	13 (5)	42 (6)	49 (3)
Cancer size*							
≤5 mm	222 (12)	37 (11)	55 (12)	130 (12)	20 (12)	51 (11)	151 (12)
6–10 mm	463 (25)	83 (25)	119 (27)	261 (24)	33 (20)	101 (22)	329 (26)
11–15 mm	484 (26)	83 (25)	133 (30)	268 (25)	49 (30)	115 (25)	320 (26)
16–20 mm	274 (15)	66 (20)	49 (11)	159 (15)	25 (15)	90 (19)	159 (13)
>20 mm	430 (23)	65 (19)	89 (20)	276 (25)	36 (22)	111 (24)	283 (23)
Unknown	88 [4]	14 [4]	15 [3]	59 [5]	15 [8]	33 [7]	40 [3]
Median cancer size*	13	14	12	13	14	15	13
Minimal cancer†							
DCIS or invasive cancer ≤10 mm	1366 (53)	239 (53)	323 (54)	804 (53)	125 (53)	303 (49)	938 (55)
Invasive cancer >10 mm	1188 (47)	214 (47)	271 (46)	703 (47)	110 (47)	316 (51)	762 (45)
Unknown	92 [3]	14 [3]	17 [3]	61 [4]	16 [6]	34 [5]	42 [2]
Early stage at diagnosis (definition 1)							
DCIS or invasive cancer <15 mm + node negative	1554 (61)	263 (58)	380 (64)	911 (61)	151 (64)	344 (55)	1059 (63)
Other	991 (39)	187 (42)	216 (36)	588 (39)	84 (36)	278 (45)	629 (37)
Unknown	101 [4]	17 [4]	15 [2]	69 [4]	16 [6]	31 [4]	54 [3]
Early stage at diagnosis (definition 2)							
DCIS or invasive cancer ≤10 mm + node negative	1282 (50)	221 (49)	304 (51)	757 (50)	121 (51)	282 (45)	879 (52)
Other	1265 (50)	230 (51)	292 (49)	743 (50)	114 (49)	342 (55)	809 (48)
Unknown	99 [4]	16 [3]	15 [2]	68 [4]	16 [6]	29 [5]	54 [3]
Axillary lymph node status*							
Negative	1441 (76)	242 (72)	349 (77)	850 (76)	133 (80)	350 (72)	958 (77)
Positive	467 (24)	95 (28)	103 (23)	269 (24)	34 (20)	139 (28)	294 (23)
Unknown	53 [3]	11 [3]	8 [2]	34 [3]	11 [6]	12 [2]	30 [2]
Grade*							
1 (well differentiated)	464 (26)	89 (28)	112 (26)	263 (25)	41 (28)	106 (24)	317 (27)
2 (moderately differentiated)	802 (45)	142 (45)	199 (47)	461 (44)	63 (42)	201 (45)	538 (45)
3 (poorly differentiated)	508 (28)	82 (26)	115 (27)	311 (30)	45 (30)	138 (31)	325 (27)
4 (undifferentiated)	19 (1)	2 (1)	1 (0)	16 (2)	0 (0)	3 (1)	16 (1)
Unknown	168 [9]	33 [9]	33 [7]	102 [9]	29 [6]	53 [11]	86 [7]
Estrogen receptor status*							
Negative	267 (15)	45 (15)	55 (13)	167 (16)	24 (16)	63 (15)	180 (15)

Table 3 (continues)

Table 3 (continued)

Characteristics of Screening-detected Tumors according to Average Annual Volume of Own and Any Recalled Screening Mammogram

Characteristic	Total (*n* = 2646)	Average Annual Volume of Own Recalled Screening Mammograms			Average Annual Volume of Any Recalled Screening Mammograms		
		0–25 Mammograms (*n* = 467)	26–50 Mammograms (*n* = 611)	>50 Mammograms (*n* = 1568)	0–50 Mammograms (*n* = 251)	51–125 Mammograms (*n* = 653)	>125 Mammograms (*n* = 1742
Positive	1475 (85)	260 (85)	358 (87)	857 (84)	122 (84)	352 (85)	1001 (85)
Unknown	219 [11]	43 [12]	47 [10]	129 [11]	32 [18]	86 [17]	101 [8]
Progesterone receptor status*							
Negative	432 (25)	83 (27)	81 (20)	268 (26)	38 (27)	109 (26)	285 (24)
Positive	1297 (75)	220 (73)	332 (80)	745 (74)	104 (73)	306 (74)	887 (76)
Unknown	232 [12]	45 [13]	47 [10]	140 [12]	36 [20]	86 [17]	110 [9]

Note.—Numbers in parenthese are column percentages. Percentages in brackets (for unknown variables) were not included in total column percentages. There were 3101 cancers total (2646 screening-detected cancers and 455 interval cancers). Screening-detected cancers were based on the radiologist's initial assessment of the screening mammogram (positive or negative) linked to cancer diagnosis: invasive carcinoma or DCIS diagnosed within 1 year of screening mammography or before the subject's next screening mammography examination (13).

* Invasive cancers only.

† Defined as in reference 12.

Table 4

Unadjusted Sensitivity, FPR, CDR, and Number of Women Recalled per Cancer Detected according to Own and Any Recalled Screening Mammograms

Parameter	Sensitivity		FPR		Mean CDR†‡	No. of Women Recalled per Cancer Detected
	No. of Reader-Years*	Mean (%)†	No. of Reader-Years*	Mean (%)†		
Overall	380	85.3 (83.6, 86.9)	393	9.1 (8.0, 10.3)	4.1 (3.7, 4.5)	22.2
Annual volume of own recalled mammograms						
0–25 mammograms	130 (34.2)	81.1 (77.4, 84.3)	141 (35.9)	6.7 (5.6, 8.0)	3.1 (2.7, 3.5)	17.4
26–50 mammograms	85 (22.4)	84.6 (80.4, 88.0)	86 (21.9)	8.6 (7.2, 10.3)	4.0 (3.3, 4.8)	21.3
>50 mammograms	165 (43.4)	87.0 (84.9, 88.8)	166 (42.2)	10.3 (8.7, 12.1)	4.5 (4.1, 5.0)	24.6
Annual volume of any recalled mammograms						
0–50 mammograms	94 (24.7)	80.8 (76.4, 84.7)	105 (26.7)	7.0 (5.1, 9.4)	2.9 (2.5, 3.4)	18.1
51–125 mammograms	103 (27.1)	84.7 (81.3, 87.5)	104 (26.5)	7.6 (6.5, 8.8)	3.9 (3.3, 4.6)	18.8
>125 mammograms	183 (48.2)	86.5 (84.3, 88.3)	184 (46.8)	10.3 (8.8, 11.9)	4.4 (4.0, 4.9)	24.6

* Numbers in parentheses are column percentages. Thirteen reader-years were not associated with any cancers and did not contribute to the sensitivity estimate.

† Numbers in parentheses are 95% CIs.

‡ Number of cancers detected per 1000 screening mammograms.

annually work up more than 125 of any recalled mammograms. Despite variability in performance measures, on average, radiologists who worked up fewer recalled mammograms had consistently lower sensitivity, CDRs, and FPRs at any given total volume. Current U.S. Food and Drug Administration regulations require U.S. physicians to have interpreted 960 mammograms within the previous 24 months to meet continuing experience requirements. However, the regulations have no requirements about the indication for the examination (ie, they could all be screening examinations or they could all be diagnostic examinations).

We previously examined annual interpretive volume and screening (4) and diagnostic performance (5) and reported wide, unexplained variability in screening and diagnostic performance across radiologists within volume levels. We had expected to see the relationship between total interpretive volume (screening plus diagnostic images) or screening volume to be most strongly associated with screening performance and total volume or diagnostic volume to be most strongly

Figure 1

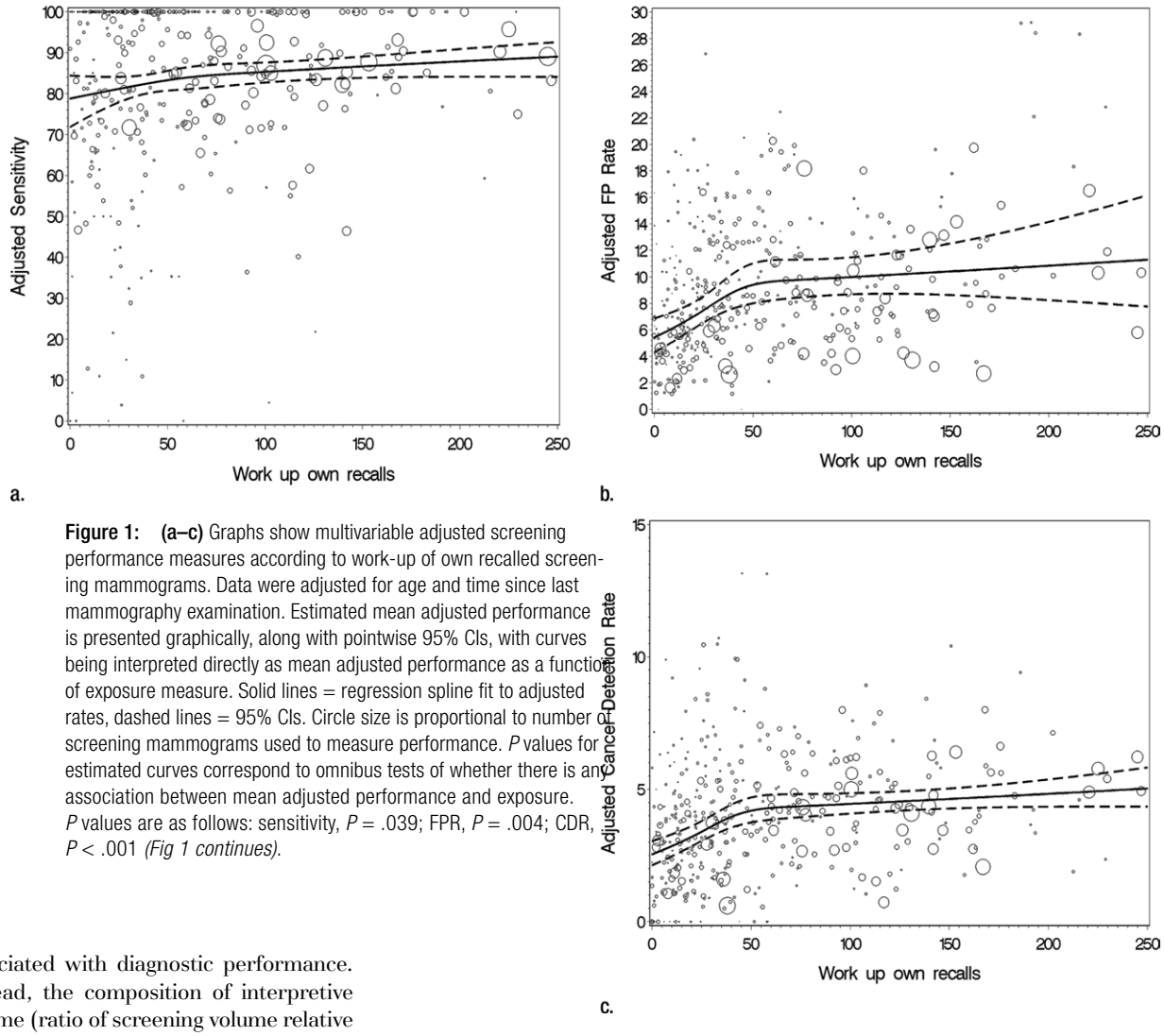


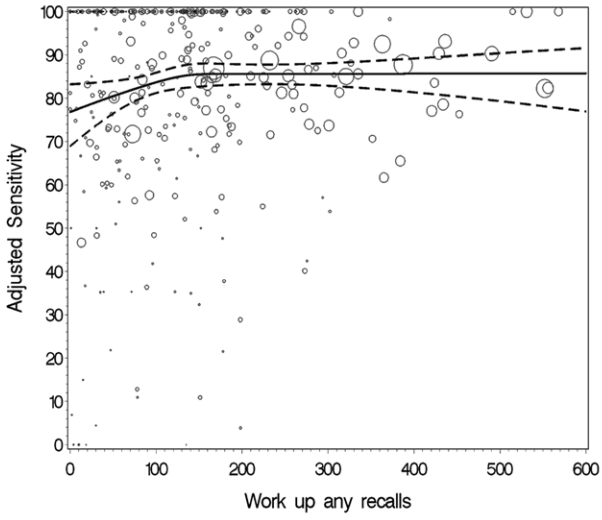
Figure 1: **(a–c)** Graphs show multivariable adjusted screening performance measures according to work-up of own recalled screening mammograms. Data were adjusted for age and time since last mammography examination. Estimated mean adjusted performance is presented graphically, along with pointwise 95% CIs, with curves being interpreted directly as mean adjusted performance as a function of exposure measure. Solid lines = regression spline fit to adjusted rates, dashed lines = 95% CIs. Circle size is proportional to number of screening mammograms used to measure performance. *P* values for estimated curves correspond to omnibus tests of whether there is an association between mean adjusted performance and exposure. *P* values are as follows: sensitivity, $P = .039$; FPR, $P = .004$; CDR, $P < .001$ *(Fig 1 continues).*

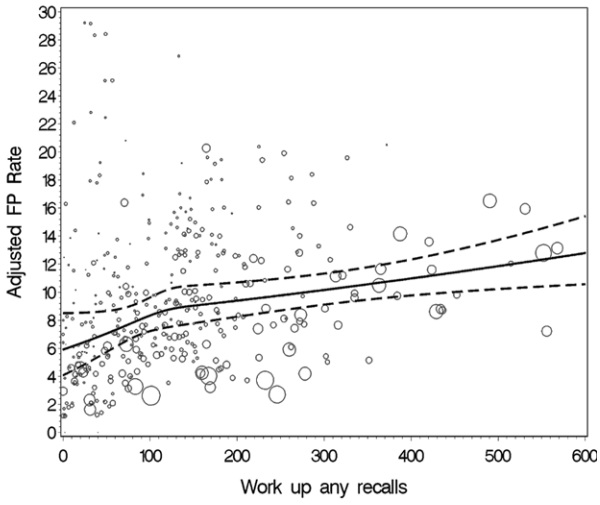
associated with diagnostic performance. Instead, the composition of interpretive volume (ratio of screening volume relative to total volume) was the greatest important factor influencing screening and diagnostic performance. Radiologists with higher annual volumes had clinically and significantly lower FPRs than their lower-volume colleagues but similar sensitivities (4,5). These earlier findings (4,5), combined with these current findings, suggest that increasing the current U.S. Mammography Quality Standards Act requirements for interpretive volume and requiring a minimum number of diagnostic work-ups for a radiologist's recalled screening mammograms could improve a radiologist's screening performance.

Despite previous findings suggesting that the proportion of screening or diagnostic examinations is most strongly associated with screening performance (4,5), we chose to investigate the total number of examinations rather than proportions because tracking examination numbers might be more practical for practices and radiologists. Tracking proportions requires more robust data collection, including total numbers of examinations according to type and proportion. Many mammography facilities cannot provide complete data volume according to interpretation type (screening vs diagnostic). Facilities also might not be able to link recalled mammograms with the radiologist who worked up the examinations. However, these findings suggest new approaches to organizing clinical work, and tracking screening and diagnostic volume may be worthwhile given the potential to improve interpretative performance.

Increasing the minimum number of interpretations might cause some radiologists with lower annual volumes to stop interpreting mammograms. Conversely, these findings may motivate those radiologists to increase their volumes. Workforce issues may also be less relevant today with the increasing use of digital mammography, which allows radiologists to interpret examinations

Figure 1 (continued)





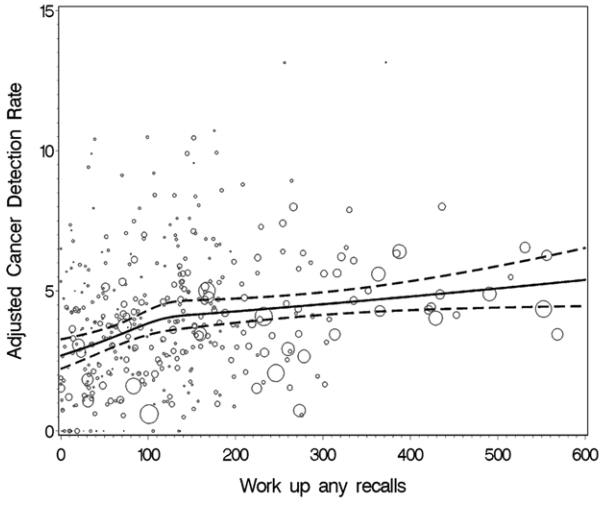


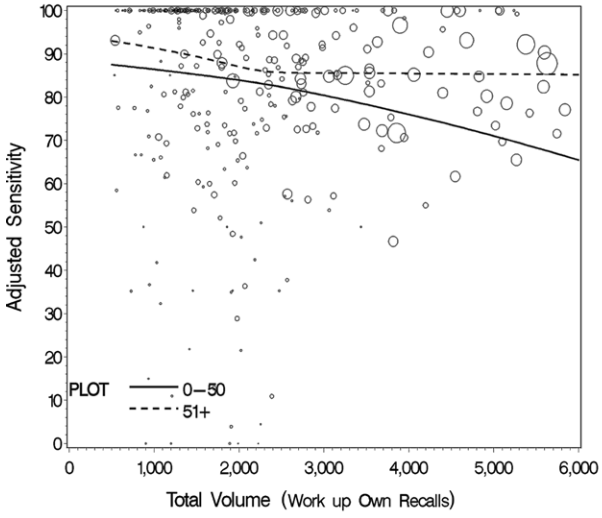
Figure 1: *(continued)* **(d–f)** Graphs show multivariable adjusted screening performance measures according to work-up of any recalled screening mammograms. Data were adjusted for age and time since last mammography examination. Estimated mean adjusted performance is presented graphically, along with pointwise 95% CIs, with curves being interpreted directly as mean adjusted performance as a function of exposure measure. Solid lines = regression spline fit to adjusted rates, dashed lines = 95% CIs. Circle size is proportional to number of screening mammograms used to measure performance. *P* values for estimated curves correspond to omnibus tests of whether there is any association between mean adjusted performance and exposure. *P* values are as follows: sensitivity, *P* = .15; FPR, *P* = .011; CDR, *P* = .001.

remotely. Our data support a minimum annual interpretive volume coupled with annual work-up of at least 50 of a radiologist's own recalled mammograms. This recommendation would require changes in how facilities capture and report current Mammography Quality Standards Act interpretation requirements and may require some facilities to reorganize their workflow. In addition, in the absence of a national reporting registry, tracking interpretive requirements across facilities, particularly if volume requirements span multiple years, would be challenging. An alternative would be to have radiologists review the work-up of a portion of their own recalled mammograms (category 0), even if they were not the radiologist that performed the work-up.

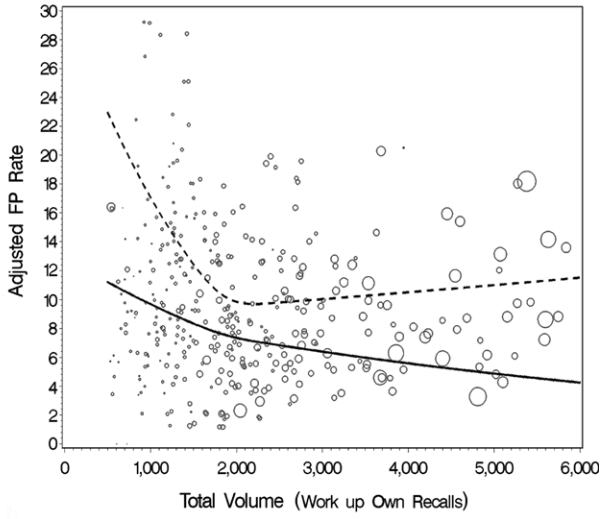
A common assumption is that improvements in sensitivity come at the expense of specificity, and vice versa, as reflected in traditional receiver operating curve analysis. However, this is not always the case. It is possible to improve both measures to the point where improvements in one measure reach a threshold beyond which the other is diminished (20). Thus, increases in FPRs associated with the improvement in sensitivity and CDRs potentially could be reduced with use of other strategies to improve interpretative performance, such as interventions for radiologists to improve interpretative performance (21–23), application of performance thresholds (20), providing additional audit feedback by reviewing the lesion that was sampled for biopsy, or providing additional feedback related to improving specificity (24–27). Some women who undergo screening mammography may regard the small increase in the FPR as an acceptable trade-off for improved sensitivity (28–31).

Disentangling the factors that influence interpretive performance (for mammography or any technology) requires in-depth longitudinal examinations on large populations that enable cause and effect to be established.

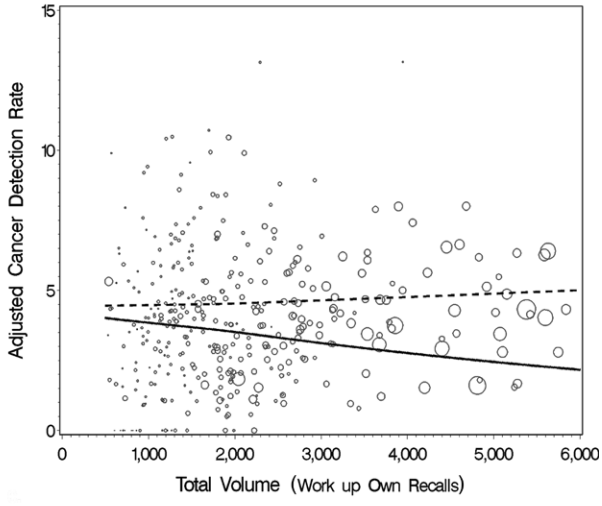
Figure 2



a.



b.



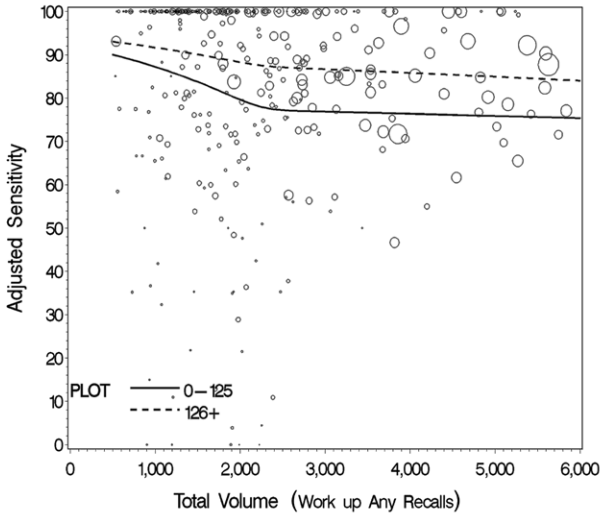
c.

Figure 2: **(a–c)** Graphs show multivariable adjusted screening performance measures according to total volume in terms of low (0–50 mammograms; solid line) and high (≥51 mammograms; dashed line) volume of own work-ups. Data are adjusted for age and time since last mammography examination. Plotted lines represent estimated mean adjusted performance as a function of total volume and according to radiologists with low (0–50 mammograms) and high (≥51 mammograms) volumes of own work-ups. Circle size is proportional to number of screening mammograms used to measure performance. *P* values correspond to omnibus tests of whether there is difference in shape (interaction term to assess effect modification) of volume-performance association according to low versus high volumes of work-up of recalled screening mammograms. *P* values are as follows: sensitivity, $P = .36$; FPR, $P = .09$; CDR, $P = .039$ *(Fig 2 continues).*

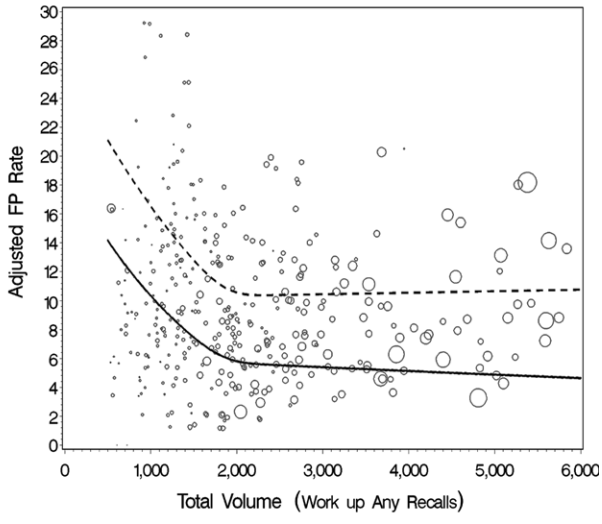
Current Mammography Quality Standards Act requirements had no supporting evidence when they were established to support the specifics of their requirements; they were a well-intentioned judgment call. Years later, we now have evidence demonstrating that the combination of higher volume and direct involvement in working up one's own recalled screening mammograms is associated with a higher sensitivity and CDR.

Our study had limitations. Mammography performance was derived from examinations performed between 2002 and 2006, when computer-assisted detection and digital mammography were not as ubiquitous as they are now; however, few studies have shown large clinical improvements in performance with these newer technologies (32–37). In addition, during the study period computer-assisted detection was not commonly used in the BCSC (only 29% of screening mammograms). This was not a trial where we manipulated work-up volumes to test whether changing the composition would improve an individual radiologist's performance. Our analysis did not include double-reading for screening or diagnostic mammograms; however, only 3% of radiologists reported double-reading for 20% or more of their screening mammograms. Instead of prespecifying the categories for exposure distribution, our cutpoints were determined after examining the data. We took this approach to address feasibility for measurement, implementation, and policy. For example, with recalls in the range of 10% and minimum interpretive volume being 980 mammograms during a 2-year period, it would be reasonable for a radiologist to be eligible to review 50 recalled screening mammograms. We also picked cut-off values that we thought would be feasible for implementation rather than basing them only off the variable distribution–in other words, at standard intervals (eg, 50 vs 53 mammograms).

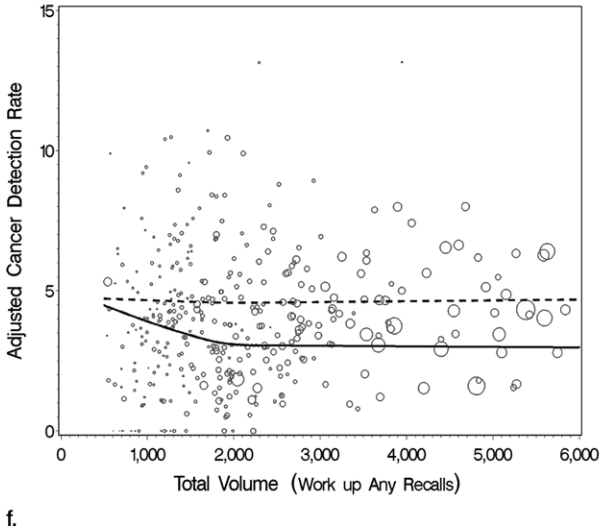
Figure 2 (continued)



d.



e.



f.

Figure 2: *(continued)* **(d–f)** Graphs show multivariable adjusted screening performance measures according to total volume in terms of low (0–125 mammograms; solid line) and high (≥126 mammograms; dashed line) volume of any work-ups. Data are adjusted for age and time since last mammography examination. Plotted lines represent estimated mean adjusted performance as a function of total volume and according to radiologists with low (0–125 mammograms) and high (≥126 mammograms) volumes of any work-ups. Circle size is proportional to number of screening mammograms used to measure performance. *P* values correspond to omnibus tests of whether there is difference in shape (interaction term to assess effect modification) of volume-performance association according to low versus high volumes of work-up of recalled screening mammograms. *P* values are as follows: sensitivity, $P = .92$; FPR, $P = .63$; CDR, $P = .47$.

Finally, given the small number of breast imaging specialists, we could not examine the influence of own recalled mammograms and specialty training. We also had no information on how radiologists collaborate during the work-up process and therefore could not examine how these interactions affected performance metrics.

Results of our analyses suggest that radiologists' screening performance could improve with work-up of more than 50 of their own recalled screening mammograms. Our findings support the BI-RADS strong recommendations to track all recalled screening mammograms, for separate auditing for screening and diagnostic examinations, and for more extenstive auditing. This study, combined with previous investigations (4,5), supports an increase in annual volume requirements and a minimum diagnostic volume of recalled screening cases for U.S. radiologists who interpret mammograms.

Acknowledgments: The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the United States. For a full description of these sources, please see *http://www.breastscreening.cancer.gov/work/acknowledgement.html*. We thank the BCSC investigators, participating women, mammography facilities, and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: *http://breastscreening.cancer.gov/*. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

Disclosures of Conflicts of Interest: D.S.M.B. No relevant conflicts of interest to disclose. **M.L.A.** No relevant conflicts of interest to disclose. **R.A.S.** No relevant conflicts of interest to disclose. **P.A.C.** No relevant conflicts of interest to disclose. **D.L.M.** No relevant conflicts of interest to disclose. **B.S.M.** No relevant conflicts of interest to disclose. **E.A.S.** No relevant conflicts of interest to disclose. **S.H.T.** No relevant conflicts of interest to disclose. **B.M.G.** No relevant conflicts of interest to disclose. **B.C.Y.** No relevant conflicts of interest to disclose. **T.L.O.** No relevant conflicts of interest to disclose.

References

1. Institute of Medicine. Improving breast imaging quality standards. Washington, DC: National Academies Press, 2005.
2. Ichikawa LE, Barlow WE, Anderson ML, et al. Time trends in radiologists' interpretive performance at screening mammography from the community-based Breast Cancer Surveillance Consortium, 1996–2004. Radiology 2010;256(1):74–82.
3. Hébert-Croteau N, Roberge D, Brisson J. Provider's volume and quality of breast cancer detection and treatment. Breast Cancer Res Treat 2007;105(2):117–132.
4. Haneuse S, Buist DS, Miglioretti DL, et al. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. Radiology 2012;262(1):69–79.
5. Buist DS, Anderson ML, Haneuse SJ, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 2011;259(1):72–84.
6. Halladay JR, Yankaskas BC, Bowling JM, Alexander C. Positive predictive value of mammography: comparison of interpretations of screening and diagnostic images by the same radiologist and by different radiologists. AJR Am J Roentgenol 2010;195(3):782–785.
7. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138(3):168–175.
8. 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(4):1001–1008.
9. National Cancer Institute; Breast Cancer Surveillance Consortium Web site. http://breastscreening.cancer.gov/. Updated February 7, 2014. Accessed June 3, 2014.
10. Elmore JG, Jackson SL, Abraham L, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 2009;253(3):641–651.
11. Rosenberg RD, Haneuse SJ, Geller BM, et al; Breast Cancer Surveillance Consortium. Timeliness of follow-up after abnormal screening mammogram: variability of facilities. Radiology 2011;261(2):404–413.
12. American College of Radiology. ACR BI-RADS — mammography. In: ACR Breast Imaging and Reporting and Data System, breast imaging atlas. 4th ed. Reston, Va: American College of Radiology, 2003.
13. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006;241(1):55–66.
14. Greenland S. Introduction to regression modeling. In: Rothman KJ, Greenland S, eds. Modern epidemiology. 2nd ed. Philadelphia, Pa: Lippincott-Raven, 1998; 401–434.
15. Hastie T, Tibshirani R, Friedman J. Basis expansions and regularization. 5 5. In: The elements of statistical learning: data mining, inference and prediction. New York, NY: Springer-Verlag, 2001; 127–133.
16. SAS Institute. SAS/GRAPH 9.2 reference. 2nd ed. Cary, NC: SAS Institute, 2010.
17. StataCorp. Stata statistical software: release 12. College Station, Tex: StataCorp, 2011.
18. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95(4):282–290.
19. Miglioretti DL, Haneuse SJ, Anderson ML. Statistical approaches for modeling radiologists' interpretive performance. Acad Radiol 2009;16(2):227–238.
20. Carney PA, Parikh J, Sickles EA, et al. Diagnostic mammography: identifying minimally acceptable interpretive performance criteria. Radiology 2013;267(2):359–367.
21. Carney PA, Bogart A, Sickles EA, et al. Feasibility and acceptability of conducting a randomized clinical trial designed to improve interpretation of screening mammography. Acad Radiol 2013;20(11):1389–1398.
22. Geller B, Bogart TA, Carney PA, et al. Educational interventions to improve screening mammography interpretation: a randomized, controlled trial. AJR Am J Roentgenol 2014;202(6)W586–W596.
23. Carney PA, Bowles EJ, Sickles EA, et al. Using a tailored web-based intervention to set goals to reduce unnecessary recall. Acad Radiol 2011;18(4):495–503.
24. Adcock K. Initiative to improve mammogram interpretation. Perm J 2004;8(2):12–18.
25. Geller BM, Ichikawa L, Miglioretti DL, Eastman D. Web-based mammography audit feedback. AJR Am J Roentgenol 2012;198(6):W562–W567.
26. Aiello Bowles EJ, Geller BM. Best ways to provide feedback to radiologists on mammography performance. AJR Am J Roentgenol 2009;193(1):157–164.
27. Elmore JG, Aiello Bowles EJ, Geller B, et al. Radiologists' attitudes and use of mammography audit reports. Acad Radiol 2010;17(6):752–760.
28. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. U.S. women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. BMJ 2000;320(7250):1635–1640.
29. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Ann Intern Med 2007;146(7):502–510.
30. Defrank JT, Brewer N. A model of the influence of false-positive mammography screening results on subsequent screening. Health Psychol Rev 2010;4(2):112–127.
31. DeFrank JT, Rimer BK, Bowling JM, Earp JA, Breslau ES, Brewer NT. Influence of false-positive mammography results on subsequent screening: do physician recommendations buffer negative effects? J Med Screen 2012;19(1):35–41.
32. Cole EB, Zhang Z, Marques HS, et al. Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the Digital Mammographic Imaging Screening Trial. AJR Am J Roentgenol 2012;199(3):W392–W401.
33. Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. J Natl Cancer Inst 2011;103(15):1152–1161.
34. Fenton JJ, Xing G, Elmore JG, et al. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. Ann Intern Med 2013;158(8):580–587.
35. Hubbard RA, Zhu W, Onega TL, et al. Effects of digital mammography uptake on downstream breast-related care among older women. Med Care 2012;50(12):1053–1059.
36. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study. Ann Intern Med 2011;155(8):493–502.
37. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005;353(17):1773–1783.