Influence of Annual Interpretive Volume on Screening Mammography Performance in the United States[1]

Diana S. M. Buist, PhD, MPH
Melissa L. Anderson, MS
Sebastien J. P. A. Haneuse, PhD[2]
Edward A. Sickles, MD
Robert A. Smith, PhD
Patricia A. Carney, PhD
Stephen H. Taplin, MD, MPH
Robert D. Rosenberg, MD
Berta M. Geller, EdD
Tracy L. Onega, PhD
Barbara S. Monsees, MD
Lawrence W. Bassett, MD
Bonnie C. Yankaskas, PhD
Joann G. Elmore, MD, MPH
Karla Kerlikowske, MD
Diana L. Miglioretti, PhD

Purpose: To examine whether U.S. radiologists' interpretive volume affects their screening mammography performance.

Materials and Methods: Annual interpretive volume measures (total, screening, diagnostic, and screening focus [ratio of screening to diagnostic mammograms]) were collected for 120 radiologists in the Breast Cancer Surveillance Consortium (BCSC) who interpreted 783 965 screening mammograms from 2002 to 2006. Volume measures in 1 year were examined by using multivariate logistic regression relative to screening sensitivity, false-positive rates, and cancer detection rate the next year. BCSC registries and the Statistical 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 terms of the Health Insurance Portability and Accountability Act.

Results: Mean sensitivity was 85.2% (95% confidence interval [CI]: 83.7%, 86.6%) and was significantly lower for radiologists with a greater screening focus ($P = .023$) but did not significantly differ by total ($P = .47$), screening ($P = .33$), or diagnostic ($P = .23$) volume. The mean false-positive rate was 9.1% (95% CI: 8.1%, 10.1%), with rates significantly higher for radiologists who had the lowest total ($P = .008$) and screening ($P = .015$) volumes. Radiologists with low diagnostic volume ($P = .004$ and $P = .008$) and a greater screening focus ($P = .003$ and $P = .002$) had significantly lower false-positive and cancer detection rates, respectively. Median invasive tumor size and proportion of cancers detected at early stages did not vary by volume.

Conclusion: Increasing minimum interpretive volume requirements in the United States while adding a minimal requirement for diagnostic interpretation could reduce the number of false-positive work-ups without hindering cancer detection. These results provide detailed associations between mammography volumes and performance for policymakers to consider along with workforce, practice organization, and access issues and radiologist experience when reevaluating requirements.

Supplemental material: *http://radiology.rsna.org/lookup/suppl/doi:10.1148/radiol.10101698/-/DC1*

[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., S.J.P.A.H., D.L.M.); Department of Biostatistics, University of Washington School of Public Health, Seattle, Wash (S.J.P.A.H., D.L.M.); Department of Radiology, University of California, San Francisco, Calif (E.A.S.); 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.); Division of Cancer Control and Population Science, Applied Research Program, National Cancer Institute, Bethesda, Md (S.H.T.); Department of Radiology, University of New Mexico, Health Sciences Center, Albuquerque, NM (R.D.R.); Health Promotion Research, University of Vermont, College of Medicine, Burlington, Vt (B.M.G.); Department of Community and Family Medicine, Dartmouth Medical School, Norris Cotton Cancer Center, Lebanon, NH (T.L.O.); Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Mo (B.S.M.); Department of Radiology, David Geffen School of Medicine at UCLA, Los Angeles, Calif (L.W.B.); Department of Radiology, University of North Carolina, Chapel Hill, NC (B.C.Y.); Department of Medicine, University of Washington School of Medicine, Seattle, Wash (J.G.E.); and Department of Epidemiology and Biostatistics and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, Calif (K.K.). Received August 24, 2010; revision requested October 19; revision received October 29; accepted November 10; final version accepted November 22. Supported in part by the American Cancer Society and made possible by a generous donation from the Longaberger Company's Horizon of Hope Campaign (grants SIRGS-07-271-01, SIRGS-07-272-01, SIRGS-07-274-01, SIRGS-07-275-01, SIRGS-06-281-01, and ACS A1-07-362), the Breast Cancer Stamp Fund, the Agency for Healthcare Research and Quality and National Cancer Institute (grant CA107623), and the National Cancer Institute Breast Cancer Surveillance Consortium (grants U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, and U01CA70040). **Address correspondence to** D.S.M.B. (e-mail: (*buist.d@ghc.org*).
[2] **Current address:** Department of Biostatistics, Harvard School of Public Health, Boston, Mass.

Mammography is the only screening test that has been demonstrated in trials to reduce breast cancer mortality (1). An Institute of Medicine report (2) noted that while the technical quality of mammography has improved since implementation of the Mammography Quality Standards Act of 1992 (MQSA), optimal sensitivity and specificity have not yet been achieved (3). The report called for additional research on the relationship between interpretive volume and performance (2).

Compared with the United States, other countries with established screening mammography programs have lower false-positive rates but comparable cancer detection rates (CDRs) (4–6). Hypothesized reasons include the shorter screening intervals and the lower interpretive volume requirements in the United States (960 mammograms every 2 years—five- to 10-fold lower than in other countries) (7–9).

Mammography performance and volume findings, though inconsistent (10), generally suggest that higher-volume readers have lower false-positive rates with no sensitivity difference. Prior studies had limitations that could account for the inconsistencies, including different approaches for measuring volume (self report vs observational data), study designs, and settings (test set vs clinical practice), performance measures, covariates, and modeling methods. In a small Canadian study (11), abnormal interpretations were lower and CDRs were higher for radiologists performing between 2000 and 3999 interpretations annually compared with those for radiologists performing fewer than 2000 interpretations annually. Two conflicting studies used overlapping populations from the Breast Cancer Surveillance Consortium (BCSC) (12,13). Barlow et al (14) found that higher self-reported volume was associated with greater sensitivity and higher false-positive rates. Smith-Bindman et al (15) defined volume by measured number of BCSC examinations and found that higher volume was associated with lower false-positive rates, with no impact on sensitivity. Barlow et al included no validation of self-reported volume, and Smith-Bindman et al did not capture volume from mammograms interpreted at non-BCSC facilities. Inconsistent findings with overlapping study populations highlight how different conclusions may be drawn by using different analytic methods and measures.

Our purpose was to examine whether U.S. radiologists' interpretive volume affects their screening mammography performance.

Advances in Knowledge

- We found no clear association between interpretive volume and sensitivity.
- Performance across radiologists within volume levels had wide, unexplained variability, reinforcing the ideas that the volume-performance relationship is complex and several factors may influence it.
- Screening performance is unlikely to be affected by volume alone, but rather by a balance in the interpreted examination composition; radiologists with a greater screening focus had significantly lower sensitivity ($P = .023$), cancer detection ($P = .002$), and false-positive rates ($P = .003$).
- Radiologists with higher annual volumes had clinically and statistically significantly lower false-positive rates with similar sensitivities as their colleagues with lower annual volumes.

Materials and Methods

BCSC registries and the Statistical Coordinating Center have 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 are compliant with the terms of the Health Insurance Portability and Accountability Act (16).

Sample Group

This study included six BCSC mammography registries (in California, North Carolina, New Hampshire, Vermont, Washington, and New Mexico) that have previously been described (12,13). BCSC registries collect information on mammography examinations 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. Demographic and breast cancer risk factor data are collected by using a self-reported questionnaire completed at each mammography examination.

Implication for Patient Care

- Our results provide detailed associations between mammography volumes and performance for policymakers to consider along with workforce, practice organization, and access issues and radiologist experience when reevaluating requirements.

Published online before print
10.1148/radiol.10101698

Radiology 2011; 259:72–84

Abbreviations:
BCSC = Breast Cancer Surveillance Consortium
CDR = cancer detection rate
CI = confidence interval
DCIS = ductal carcinoma in situ
MQSA = Mammography Quality Standards Act of 1992

Author contributions:
Guarantors of integrity of entire study, D.S.M.B., D.L.M.; 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., P.A.C., S.H.T., B.M.G., T.L.O., K.K., D.L.M.; clinical studies, D.S.M.B., S.H.T., T.L.O., B.C.Y., K.K., D.L.M.; statistical analysis, M.L.A., S.J.P.A.H., R.A.S., D.L.M.; and manuscript editing, all authors

Funding:
This research was supported by the Agency for Healthcare Research and Quality and National Cancer Institute (grant CA107623) and the National Cancer Institute Breast Cancer Surveillance Consortium (grants U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, and U01CA70040).

Potential conflicts of interest are listed at the end of this article.

BCSC radiologists who interpreted screening mammograms between 2005 and 2006 were invited to complete a self-administered mailed survey, as previously described (17); 214 radiologists completed it. The survey asked radiologists to indicate all facilities (outside the BCSC) at which they interpreted mammograms between 2001 and 2005. Registry staff members contacted non-BCSC facilities and collected complete volume information for all radiologists. Our final study sample included 120 radiologists with a mean of 4.0 years of volume measures, resulting in 481 total reader-years of data for the analytic sample; 91% of radiologists (109 of 120) interpreted mammograms at only BCSC facilities. The demographic characteristics, time spent in breast imaging, and experience of the radiologists included in the analytic sample were similar to those of the radiologists included in the original sample (Table E1 [online]).

Interpretive volume, collected for each radiologist by facility and examination year, was summed across all facilities to obtain the annual volume for each radiologist. We collected total, screening, and diagnostic volumes for each year. We included mammograms in the volume estimate only if the radiologist was the primary reader. The radiologist's indication for the examination was used to categorize screening and diagnostic volumes. Diagnostic examinations included additional evaluation of a prior mammogram, short-interval follow-up, or evaluation of a breast symptom or mammographic abnormality. Total volume included all screening and diagnostic examinations, with screening and diagnostic mammograms interpreted by the same radiologist on the same day counted as one examination. The latter definition differs from that used in the Breast Imaging Reporting and Data System audits (18), where mammograms performed on the same day contribute independently to volume. These two approaches yielded little difference in calculating total volume. For the 2.5% of mammograms with missing indications, we attributed their indication to screening or diagnostic volume on the basis of the proportion of screening mammograms where indications were observed for that reader.

For each year, we also computed the "screening focus" for each reader, defined as the percentage of total mammograms that were screening examinations.

All volume measures for each year were linked to screening performance in the following year (eg, volume in 2005 was linked to performance in 2006). Performance data were based solely on screening mammograms interpreted within BCSC facilities, because although volume data were available from non-BCSC facilities, linkage to cancer follow-up data was not possible for mammograms interpreted at non-BCSC facilities. Only 4% (19 of 481) of our study's reader-years included radiologists who interpreted at non-BCSC facilities; among these readers, a mean of 54% of their total volume comprised non-BCSC mammograms.

Women given a diagnosis of invasive carcinoma or ductal carcinoma in situ (DCIS) within 1 year of the screening mammogram and before their next screening mammogram were included as representing breast cancer cases (19). Tumor characteristics were collected from tumor registries and pathology databases. We defined early stage cancers in three ways (19): *(a)* DCIS or invasive cancer that was 10 mm or smaller, *(b)* node-negative invasive cancer or DCIS that was 10 mm or smaller, and *(c)* node-negative invasive cancer or DCIS that was smaller than 15 mm.

Performance measures included sensitivity, false-positive rate, and CDR. Sensitivity was defined as the proportion of screening examinations interpreted as positive (defined as Breast Imaging Reporting and Data System category 0 [needs additional assessment], category 4 [suspicious abnormality], or category 5 [highly suggestive of malignancy]) among all women who were given a diagnosis of breast cancer within the 1-year follow-up period. False-positive rate was defined as the proportion of positive screening examinations among all women without a breast cancer diagnosis within the follow-up period. CDR was defined as the number of cancers detected per 1000 screening mammograms interpreted. Performance measures were derived from 783965 screening examinations (in 476079 unique women) interpreted between 2002 and 2006; these included mammograms in asymptomatic women with a routine screening indication.

Analysis

We calculated crude performance measures according to categoric versions of the four continuous volume measures. The association between continuous volume measures and screening performance was modeled by using restricted cubic smoothing splines (20), which permit a flexible shape while avoiding arbitrary cutpoints. We computed the smoothing splines with three knots placed at the 33rd, 50th, and 67th percentiles of the volume distribution. Volume distributions were heavily skewed with sparse information in the tails; therefore, we restricted each volume range before fitting the logistic regression models. Estimations were limited to a total volume of 6000 or fewer mammograms, a screening volume of 5000 or fewer mammograms, a diagnostic volume of 2000 or fewer mammograms, and a screening focus of 65% or greater to ensure that we had adequate information for stable estimates for model parameters.

To measure the potential tradeoff between sensitivity and false-positive rates, we calculated the number of women recalled for each cancer detected. To adjust for differing case-mix distributions across radiologists, we computed adjusted performance measures by using internal standardization (21). Internal standardization works by reweighting each mammogram according to the relative difference between the radiologist-specific distribution of potential confounders (age and time since last mammogram) and the corresponding distribution in the overall analytic sample. Intuitively, this process results in calculated performance measures for radiologists had their case mixes been the same as that in the overall population. Standardizing removes differences in the potential confounders across radiologists and therefore removes the potential

Table 1

Characteristics of Radiologists according to Total Interpretive Volume

Characteristic	Total*	Mean Annual Total Volume (No. of Mammograms)					
		480–999	1000–1499	1500–1999	2000–2999	3000–4999	≥5000
No. of radiologists*	120	20 (17)	25 (21)	19 (16)	29 (24)	16 (13)	11 (9)
Age at survey (y)							
<45	31 (26)	30	28	42	17	25	9
45–54	33 (28)	20	24	32	28	44	18
≥55	56 (47)	50	48	26	55	31	73
Sex							
Male	80 (67)	65	80	58	72	50	64
Female	40 (33)	35	20	42	28	50	36
Work full time (≥40 h/wk)							
No	30 (25)	30	24	22	25	25	27
Yes	88 (75)	70	76	78	75	75	73
Primary affiliation with academic medical center							
No affiliation	90 (76)	65	88	79	86	69	40
Adjunct	7 (6)	0	8	5	3	0	30
Primary	22 (18)	35	4	16	10	31	30
Experience							
Time since graduation from residency (y)							
<10	19 (16)	25	20	21	10	13	0
10–19	37 (31)	20	32	53	28	33	18
≥20	63 (53)	55	48	26	62	53	82
Combined variable of fellowship training and duration of mammography interpretation (y)							
No fellowship, <10	19 (16)	35	20	32	0	6	0
No fellowship, 10–19	38 (32)	15	36	42	45	19	18
No fellowship, ≥20	53 (44)	50	44	16	45	50	73
Fellowship, <10	3 (2)	0	0	0	3	13	0
Fellowship, ≥10	7 (6)	0	0	11	7	13	9
Time working in breast imaging (%)							
<20	30 (26)	37	48	25	17	6	9
20–39	29 (25)	21	28	44	28	19	0
40–79	19 (16)	11	0	25	21	31	18
80–100	38 (33)	32	24	6	34	44	73
Interpretive volume							
Mean annual screening volume (no. of mammograms)							
480–999	31 (26)	100	44	0	0	0	0
1000–1499	19 (16)	0	56	26	0	0	0
1500–1999	31 (26)	0	0	74	59	0	0
2000–2999	18 (15)	0	0	0	41	31	9
≥3000	21 (18)	0	0	0	0	69	91
Mean annual diagnostic volume (no. of mammograms)							
<100	15 (12)	50	12	5	0	6	0
100–199	13 (11)	35	16	5	0	6	0
200–299	30 (25)	15	48	42	24	0	0
300–499	39 (32)	0	24	47	66	25	9
500–999	12 (10)	0	0	0	10	44	18
≥1000	11 (9)	0	0	0	0	19	73
Screening focus: mean annual percentage of all examinations that are screening							
<75	10 (8)	0	12	0	3	13	36
75–79	14 (12)	10	8	21	7	6	27
80–84	41 (34)	25	36	37	41	38	18

Table 1 (continues)

Table 1 (continued)

Characteristics of Radiologists according to Total Interpretive Volume

Characteristic	Total*	Mean Annual Total Volume (No. of Mammograms)					
		480–999	1000–1499	1500–1999	2000–2999	3000–4999	≥5000
85–89	33 (28)	35	20	21	41	25	9
≥90	22 (18)	30	24	21	7	19	9
Mean annual no. of facilities where interpreting							
1	44 (37)	30	44	42	34	31	36
>1 To 2	48 (40)	55	24	42	34	56	36
>2 To 3	16 (13)	15	24	11	14	6	0
>3	12 (10)	0	8	5	17	6	27
No. of years of volume and performance data							
1	10 (8)	25	4	11	3	6	0
2	4 (3)	0	4	0	3	13	0
3	21 (18)	10	24	21	7	25	27
4	25 (21)	35	8	16	21	25	27
5	60 (50)	30	60	53	66	31	45

Note.—Unless otherwise specified, data are column percentages. There were missing values for the work full time question ($n = 2$), the primary affiliation with an academic medical center question ($n = 1$), the time since graduation from residency question ($n = 1$), and the time working in breast imaging question ($n = 4$).

* Data are numbers of radiologists, with percentages in parentheses.

impact of confounding. We then used generalized estimating equations to model the marginal association between the continuous adjusted performance measures and volume (20). Given the flexibility provided by the restricted cubic spline framework, the estimated mean adjusted performance is presented graphically, along with pointwise 95% confidence intervals (CIs), with the curves being interpreted directly as the mean adjusted performance as a function of the volume measure. *P* values for the estimated curves correspond to omnibus tests of whether there is any association between mean adjusted performance and volume. Under the null hypothesis, there would be no association, and the estimated relationship would be a flat line. We stratified by cancer status, fitting separate models for each performance measure by using a binary outcome based on the radiologist's initial mammogram assessment of positive or negative. Robust sandwich standard error estimates were calculated to ensure appropriate accounting of correlation by the same radiologist (22).

Sensitivity analyses to examine the robustness of our results to various data and modeling assumptions included varying the number and location of the smoothing splines knots and individually excluding each registry to ensure that none overly influenced results.

All analyses were performed by using software (SAS, version 9.2; SAS Institute, Cary, NC), including the restricted cubic splines programming (23). Two-sided *P* values less than .05 were considered to indicate a statistically significant difference.

We simulated the effect of increasing minimum interpretive volume in the United States on the basis of an estimated 34 million women aged 40–79 years undergoing screening each year (24–26) by using software (Stata, version 11; Stata, College Station, Tex). Cancer status was assigned on the basis of a cancer rate of 5.0 cancers per 1000 mammograms. Each woman was randomly assigned to one of the study radiologists, with the selection probability proportional to the reader's observed volume. We used our primary multivariate model results to obtain the estimated probability of recall, on the basis of the woman's cancer status and the reader's observed volume, then randomly generated a mammogram result on the basis of the estimated probability of a recall. To simulate increasing the volume requirements, we generated a second mammogram result for each woman. If the woman's originally assigned radiologist had a volume greater than the threshold (ie, screening volume > 1500 mammograms), then the estimated probability of recall remained unchanged. If the original reader had a volume below the threshold, then the woman was randomly assigned to a reader with a volume above the threshold. We compared the simulated test results against cancer status to estimate the number of cancers that were correctly recalled and the number of false-positive tests before and after reader reassignment. We based cost estimates on the mean Medicare reimbursement (27) of $107 per screening examination. We were unable to provide any estimates of cost savings for fewer missed cancers, because these costs are not well documented.

Results

We studied 120 radiologists with a median age of 54 years (range, 37–74 years); most worked full time (75%), had 20 or more years of experience (53%), and had no fellowship training in breast imaging (92%) (Table 1). Time spent in breast imaging varied, with 26% of radiologists working less than 20% and 33% working 80%–100% of their time in breast imaging. Most

Table 2

Characteristics of Women with Screening Mammograms between 2002 and 2006 in Relation to their Radiologists' Mean Annual Total Interpretive Volume

Parameter	Total*	Mean Annual Total Volume (No. of Mammograms)					
		480–999	1000–1499	1500–1999	2000–2999	3000–4999	≥5000
No. of mammograms*	783965	39072 (5.0)	89857 (11.5)	91901 (11.7)	217693 (27.8)	154385 (19.7)	191057 (24.4)
Patient age (y)							
<40	23370	3.3	2.7	2.6	2.5	2.8	3.8
40–49	220779	29.1	26.4	26.2	26.1	28.7	31.7
50–59	249495	32.4	32.1	32.6	31.7	32.1	31.2
60–69	155608	20.0	20.2	20.2	21.1	19.4	18.5
70–79	99370	11.4	13.7	13.5	13.6	12.2	11.4
≥80	35343	3.9	4.8	5.0	5.0	4.8	3.5
First-degree family history of breast cancer							
No	586788	85.0	83.4	83.0	84.0	83.1	84.4
Yes	113559	15.0	16.6	17.0	16.0	16.9	15.6
Unknown	83618	9.5	10.0	8.2	9.9	19.1	6.5
Time since last mammogram (y)							
No previous mammogram	35617	4.8	5.2	4.7	4.6	5.0	4.6
<2	629544	82.5	83.2	84.2	84.1	84.1	85.4
3–4	52392	8.0	7.6	7.1	7.3	6.8	6.3
≥5	29783	4.7	4.0	4.0	4.0	4.1	3.6
Unknown	36629	5.2	2.0	3.2	4.1	6.1	6.0

Note.—Unless otherwise specified, data are column percentages. Unknown percentages are not included in column percentages.

* Data are numbers of mammograms, and numbers in parentheses are percentages.

(61%) interpreted 1000–2999 mammograms annually, with 9% interpreting 5000 or more mammograms. The highest-volume readers had a lower screening focus and were more likely to have been in clinical practice for 20 or more years and to have completed a breast imaging fellowship.

Mean annual screening volume ranged from 474 to 6255 mammograms (median, 1640 mammograms), and mean annual diagnostic volume ranged from 41 to 3315 mammograms (median, 305 mammograms). Mean annual screening volume was distributed as follows: 26% (31 of 120) of radiologists interpreted fewer than 1000 screening mammograms, 16% (19 of 120) of radiologists interpreted 1000–1499 screening mammograms, 26% (31 of 120) of radiologists interpreted 1500–1999 screening mammograms, 15% (18 of 120) of radiologists interpreted 2000–2999 screening mammograms, and 18% (21 of 120) of radiologists interpreted 3000 or more screening mammograms. Approximately 10% of radiologists interpreted fewer than 100 (15 of 120) or 1000 or more (11 of 120) diagnostic mammograms annually. Approximately 20% (24 of 120) of radiologists had a lower screening focus (<80% screening), and 18% (22 of 120) had a greater screening focus (≥90% screening). Radiologists with a lower screening focus had a higher total volume and were more commonly women, older, and more likely to have completed a breast imaging fellowship (Table E2 [online]).

Most of the screening mammograms included in the performance outcome measures had been obtained in women aged 40–59 years (60%), with fewer than 3% obtained in women younger than 40 years and fewer than 5% obtained in women 80 years of age or older (Table 2). Sixteen percent of women had a first-degree family history. Fewer than 5% of the mammograms were first mammograms, and comparison films were available for 86% of examinations. Characteristics of women with screening mammograms did not differ by radiologist total volume (Table 2). Characteristics of women with screening mammograms are detailed according to their radiologists' screening volume, diagnostic volume, and screening focus in Tables E3–E5 (online), respectively.

A total of 3321 cancers were detected (Table 3). Radiologists detected a median of 25 cancers (interquartile range, 11–46) over the study period. Among invasive cancers, median tumor size did not vary by radiologists' volume; overall median tumor size was 13 mm (range, <1 to 130 mm). The proportion of cancers detected at early stages did not vary across radiologist volume. Of 575 missed cancers, 507 (88%) were invasive, with a median size of 19 mm.

Unadjusted mean sensitivity was 85.2% (95% CI: 83.7%, 86.6%), false-positive rate was 9.1% (95% CI: 8.1%, 10.1%), and CDR was 4.2 cancers per 1000 mammograms (95% CI: 3.9, 4.6). Unadjusted screening sensitivity showed no consistent trends with any volume measure (Table 4), except that sensitivity decreased with higher screening percentage.

Table 3

Characteristics of Detected Tumors according to Radiologist Total Interpretive Volume

Parameter	Total*	Mean Annual Total Volume (No. of Mammograms)					
		480–999	1000–1499	1500–1999	2000–2999	3000–4999	≥5000
No. of cancers detected*	3321	143 (4)	405 (12)	452 (14)	915 (28)	753 (23)	653 (20)
Cancer histologic type							
DCIS	847 (26)	24	23	25	23	29	29
Invasive	2470 (74)	76	77	75	77	71	71
Stage							
0	847 (26)	25	24	25	24	29	29
I	1452 (45)	43	46	46	47	43	45
II	722 (23)	20	21	25	23	22	22
III	161 (5)	12	8	2	5	5	4
IV	24 (1)	0	1	2	1	0	0
Unknown	115 (3)	5	6	2	7	1	2
Cancer size (mm)†							
≤5	275 (12)	17	12	9	12	12	12
6–10	592 (25)	20	26	26	26	22	26
11–15	597 (25)	24	27	27	24	27	21
16–20	347 (15)	14	17	13	15	14	15
>20	561 (24)	26	19	25	23	25	25
Unknown	98 (4)	7	7	5	3	2	3
Median cancer size (mm)†	13	13	13	14	13	13	13
Minimal cancer‡							
DCIS or invasive cancer ≤ 10 mm	1714 (53)	53	52	52	53	53	56
Invasive cancer > 10 mm	1505 (47)	47	48	48	47	47	44
Unknown	102 (3)	6	6	4	3	2	2
Early stage at diagnosis with definition 1							
DCIS or node-negative invasive cancer ≤ 10 mm	1608 (50)	49	49	48	49	51	53
Other	1607 (50)	51	51	52	51	49	47
Unknown	106 (3)	6	5	4	4	1	2
Early stage at diagnosis with definition 2							
DCIS or node-negative invasive cancer < 15 mm	1945 (61)	62	60	59	59	62	62
Other	1268 (39)	38	40	41	41	38	38
Unknown	108 (3)	6	5	4	4	1	2
Axillary lymph node status†							
Negative	1823 (75)	73	74	75	73	76	79
Positive	593 (25)	27	26	25	27	24	21
Unknown	54 (2)	5	4	2	3	0	1
Grade†							
1: Well differentiated	597 (26)	21	27	26	24	29	27
2: Moderately differentiated	1024 (45)	41	48	47	47	42	44
3: Poorly differentiated	623 (27)	38	25	27	28	27	28
4: Undifferentiated	22 (1)	0	0	0	1	2	1
Unknown	204 (8)	16	13	8	9	6	4
Estrogen receptor status†							
Negative	329 (15)	24	11	14	13	17	16
Positive	1898 (85)	76	89	86	87	83	84
Unknown	243 (10)	18	16	6	10	7	10

Note.—Unless otherwise specified, data are column percentages. Unknown percentages are not included in column percentages.

* Data are numbers of cancers, with percentages in parentheses.

† Invasive cancers only.

‡ Defined as in the report by Rosenberg et al (19).

Table 4

Screening Performance Measures of False-Positive Rate, Sensitivity, and Number of Women Recalled per Cancer Detected according to Radiologist Interpretive Volume

	Sensitivity		False-Positive Rate			
Parameter	No. of Reader-Years*	Mean Value (%)†	No. of Reader-Years*	Mean Value (%)†	Mean CDR‡	No. of Women Recalled per Cancer Detected
Overall	464	85.2 (83.7, 86.6)	481	9.1 (8.1, 10.1)	4.2 (3.9, 4.6)	22.3
Annual total volume						
<480	11 (2.4)	83.9 (72.5, 91.1)	15 (3.1)	7.7 (4.8, 12.1)	3.4 (2.4, 4.7)	19.3
480–999	56 (12.1)	84.0 (79.2, 87.9)	63 (13.1)	11.0 (9.0, 13.4)	4.3 (3.6, 5.1)	27.0
1000–1499	99 (21.3)	89.1 (85.7, 91.8)	104 (21.6)	11.2 (9.4, 13.3)	4.6 (4.2, 5.1)	25.9
1500–1999	78 (16.8)	84.0 (79.7, 87.5)	79 (16.4)	8.3 (7.0, 9.9)	4.2 (3.5, 5.0)	20.8
2000–2999	112 (24.1)	84.3 (81.5, 86.8)	112 (23.3)	8.3 (7.1, 9.6)	4.4 (3.9, 4.9)	20.5
3000–4999	61 (13.2)	86.6 (83.1, 89.4)	61 (12.7)	8.4 (7.2, 9.7)	4.7 (3.9, 5.5)	20.2
≥5000	47 (10.1)	84.0 (81.7, 86.1)	47 (9.8)	9.5 (7.0, 12.7)	3.6 (3.1, 4.2)	23.5
Annual screening volume						
<480	14 (3.0)	88.7 (79.2, 94.1)	18 (3.7)	9.9 (6.4, 15.0)	4.2 (3.2, 5.5)	23.2
480–999	91 (19.6)	84.2 (80.2, 87.5)	98 (20.4)	11.2 (9.6, 13.0)	4.3 (3.7, 4.9)	27.5
1000–1499	98 (21.2)	88.8 (85.1, 91.7)	103 (21.4)	10.6 (9.1, 12.3)	4.9 (4.4, 5.5)	24.8
1500–1999	85 (18.3)	84.0 (80.2, 87.2)	86 (17.9)	7.7 (6.4, 9.2)	4.1 (3.5, 4.7)	19.2
2000–2999	96 (20.7)	84.5 (81.9, 86.8)	96 (20.0)	8.3 (7.4, 9.4)	4.4 (3.9, 5.0)	20.6
≥3000	80 (17.2)	85.0 (82.4, 87.2)	80 (16.6)	9.1 (7.2, 11.4)	4.0 (3.4, 4.6)	22.2
Annual diagnostic volume						
<100	48 (10.3)	82.5 (75.0, 88.1)	58 (12.1)	6.7 (5.4, 8.4)	3.3 (2.6, 4.1)	17.2
100–199	72 (15.5)	82.9 (78.8, 86.4)	76 (15.8)	6.8 (5.6, 8.3)	3.7 (3.2, 4.2)	17.3
200–299	99 (21.3)	84.4 (81.7, 86.9)	102 (21.2)	8.4 (7.3, 9.8)	4.3 (3.7, 4.9)	20.9
300–499	139 (30.0)	86.2 (83.0, 88.8)	139 (28.9)	9.5 (8.0, 11.1)	4.6 (4.0, 5.3)	22.8
500–999	54 (11.6)	85.9 (83.6, 88.0)	54 (11.2)	10.5 (9.0, 12.2)	4.6 (3.9, 5.3)	25.3
≥1000	52 (11.2)	85.7 (83.0, 88.1)	52 (10.8)	9.8 (7.4, 12.8)	4.1 (3.6, 4.7)	23.7
Screening focus (%)						
<75	61 (13.2)	86.3 (83.5, 88.7)	62 (12.9)	9.9 (7.4, 13.2)	4.5 (3.8, 5.4)	23.8
75–79	64 (13.8)	88.8 (85.6, 91.4)	64 (13.3)	11.6 (10.0, 13.4)	5.1 (4.5, 5.8)	27.0
80–84	119 (25.7)	85.3 (82.5, 87.7)	122 (25.4)	9.7 (8.4, 11.2)	4.2 (3.8, 4.7)	23.6
85–89	126 (27.2)	84.2 (81.2, 86.7)	133 (27.7)	9.1 (7.8, 10.7)	4.2 (3.7, 4.8)	22.6
≥90	94 (20.3)	81.8 (78.1, 85.0)	100 (20.8)	5.6 (4.4, 7.0)	3.4 (2.7, 4.2)	14.5

Note.—Volume represents number of mammograms.

* Data in parentheses are percentages. Seventeen reader-years were not associated with any cancers and therefore did not contribute to the sensitivity estimate.

† Data in parentheses are 95% CIs.

‡ Number of cancers detected per 1000 screening mammograms.

Radiologists with lower screening volumes had higher false-positive rates, except radiologists who interpreted fewer than 480 mammograms annually, whose false-positive rates were lower but had wide 95% CIs. Interpreters with the highest diagnostic volume had higher false-positive rates. The lowest false-positive rates were among radiologists with a screening focus of 90% or greater (5.6% [95% CI: 4.4%, 7.0%]), and this same group had the lowest CDRs (3.4 [95% CI: 2.7, 4.2]). The highest false-positive rates and CDRs were among radiologists with a screening focus of less than 80% (10.7% [95% CI: 8.9%, 12.7%] and 4.8 [95% CI: 4.2, 5.4], respectively).

Overall, 22.3 women were recalled for each cancer detected—slightly fewer for radiologists with lower diagnostic volume and higher total and screening volumes. Radiologists with a screening focus of 90% or greater recalled a mean of 14.5 women for each cancer detected but had lower sensitivity than radiologists with lower screening focus percentages. Radiologists with a screening focus of less than 80% had higher sensitivity but recalled 23.8–27.0 women per cancer detected.

Figures 1–3 show the adjusted screening performance measures according to volume. Radiologists with lower volume and a greater screening focus had greater variability in all performance outcomes. Sensitivity estimates varied little across volume, with the exception of a lower sensitivity for radiologists with a greater screening focus (P = .023). False-positive rates were significantly lower for radiologists at neither

Figure 1

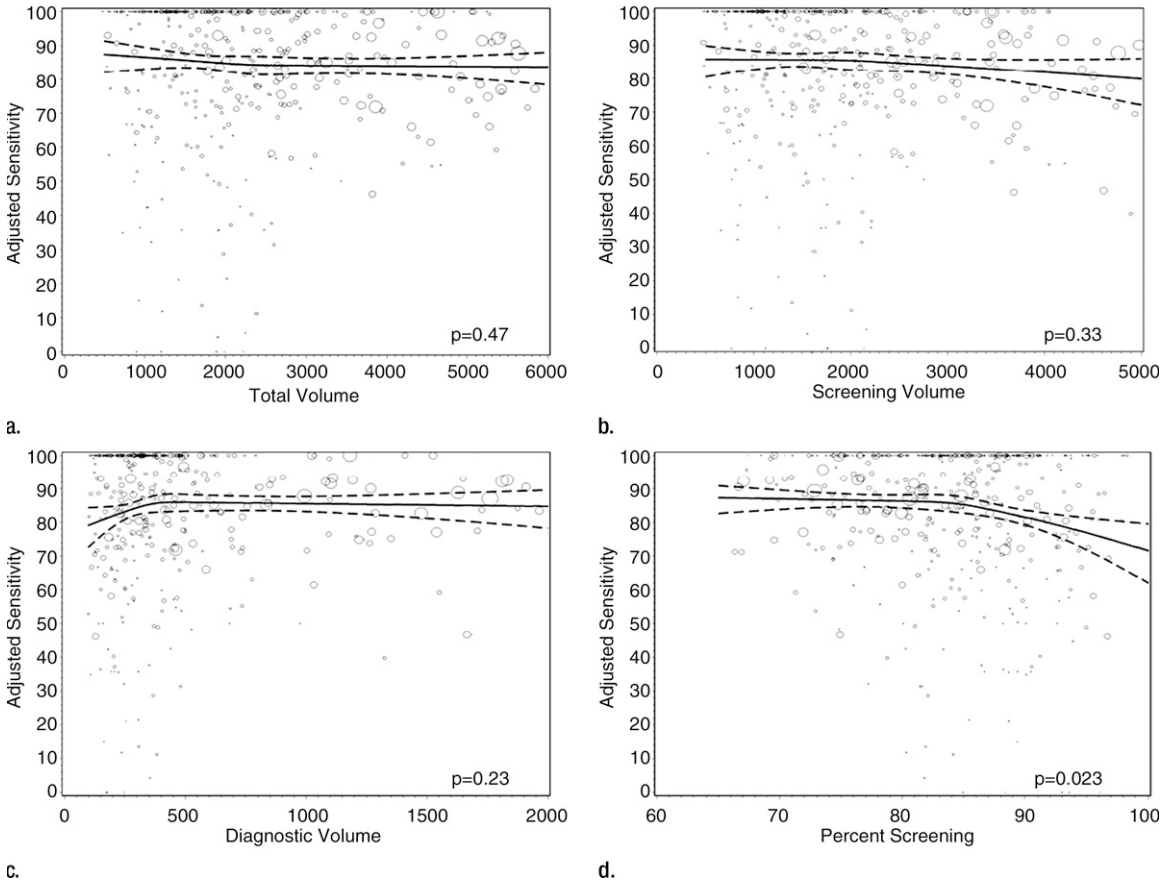

Figure 1: Graphs show adjusted sensitivity according to interpretive volume, in terms of **(a)** total volume, **(b)** screening volume, **(c)** diagnostic volume, and **(d)** percentage of total mammograms that represented screening examinations. Sensitivity was adjusted for age and time since last mammogram. Lines = regression spline fit to adjusted rates; dashed lines = 95% CIs; and ○ = adjusted sensitivity, with size proportional to the number of total cancers. Smoothing splines had three knots placed at the 33rd, 50th, and 67th percentiles of the volume distribution; estimations were limited to total volume of 6000 or fewer mammograms, screening volume of 5000 or fewer mammograms, diagnostic volume of 2000 or fewer mammograms, and a screening focus of 65% or greater. Estimated mean adjusted performance is presented graphically, along with pointwise 95% CIs, with the curves being interpreted directly as the mean adjusted performance as a function of the volume measure. *P* values for the estimated curves correspond to omnibus tests of whether there is any association between mean adjusted performance and volume.

the high nor the low extreme (mean, approximately 1500–4000 mammograms per year) for total ($P = .008$) and screening ($P = .015$) volumes. Radiologists with the lowest diagnostic volume also had lower false-positive rates ($P = .004$). Screening CDR was lower for low-volume diagnostic interpreters ($P = .008$) and for radiologists with greater screening focus ($P = .002$), but did not differ across total ($P = .30$) or screening volumes ($P = .56$).

In our simulation, on the basis of an estimated 34 million women aged 40–79 years undergoing screening each year (24–26), we found that increasing volume requirements could reduce the number of work-ups with a very small reduction in cancer detection. We estimate that increasing the annual minimum total volume requirements to at least 1000 mammograms would result in 43 629 fewer women being recalled, at the expense of missing 40 cancers while detecting 143 215 cancers. Shifting annual total volume requirements to at least 1500 mammograms would result in 92 838 fewer women being recalled, at the expense of missing 761 cancers while detecting 142 494 cancers. Basing volume requirements on annual screening volume and changing the minimum to at least 1000 mammograms would result in 71 110 fewer women being recalled, at the expense of missing 415 cancers while detecting 141 413 cancers; and changing the minimum to at least 1500 would result in 117 187 fewer women being recalled, at the expense of missing 361 cancers while detecting 141 467 cancers.

The direction and clinical interpretation of results did not change after we completed the sensitivity analyses described above.

Figure 2

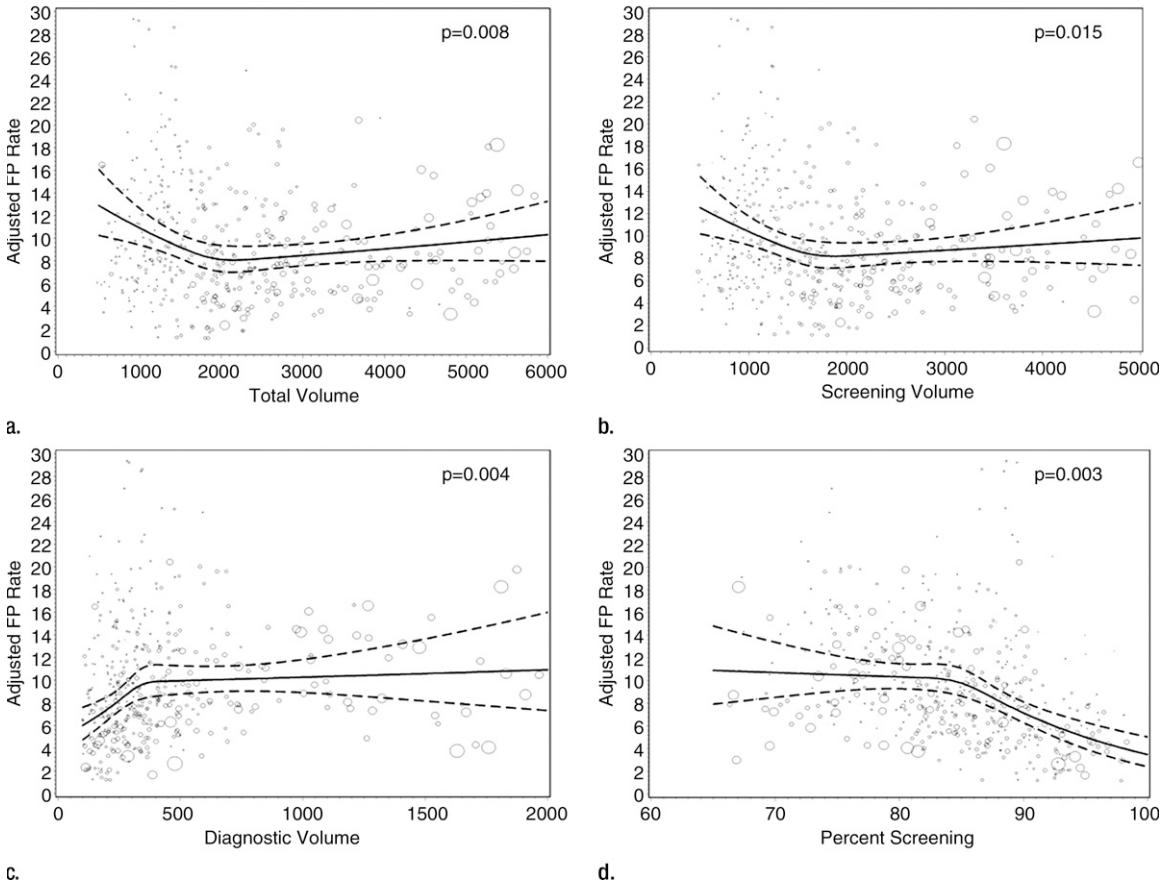

Figure 2: Graphs show adjusted false-positive rates according to interpretive volume, in terms of **(a)** total volume, **(b)** screening volume, **(c)** diagnostic volume, and **(d)** percentage of total mammograms that represented screening examinations. False-positive rates were adjusted for age and time since last mammogram. Lines = regression spline fit to adjusted rates; dashed lines = 95% CIs; and ○ = adjusted false-positive rate, with size proportional to the number of screening mammograms used to measure performance. Smoothing splines had three knots placed at the 33rd, 50th, and 67th percentiles of the volume distribution; estimations were limited to total volume of 6000 or fewer mammograms, screening volume of 5000 or fewer mammograms, diagnostic volume of 2000 or fewer mammograms, and a screening focus of 65% or greater. Estimated mean adjusted performance is presented graphically, along with pointwise 95% CIs, with the curves being interpreted directly as the mean adjusted performance as a function of the volume measure. *P* values for the estimated curves correspond to omnibus tests of whether there is any association between mean adjusted performance and volume.

Discussion

Current Food and Drug Administration regulations state that U.S. physicians interpreting mammograms must interpret 960 mammograms within the previous 24 months to meet continuing experience requirements. There has been interest in increasing radiologists' continuing experience requirements on the assumption that higher volume requirements would improve overall interpretive performance, particularly sensitivity. However, the Institute of Medicine evaluated approaches to improving the quality of mammographic interpretation and concluded that data were insufficient to justify regulatory changes to the MQSA volume requirement and called for new studies (2).

Our study was designed to examine various measures of mammography interpretive volume in relation to screening performance outcomes. Contrary to our expectations, we observed no clear association between volume and sensitivity. We found that higher interpretive volume was associated with clinically and statistically important lower rates of false-positive results and numbers of women recalled per cancer detected—without a corresponding decrease in sensitivity or CDR. We also observed lower CDRs in radiologists with low diagnostic volumes. Performance across radiologists within volume levels had wide, unexplained variability, reinforcing the ideas that the volume-performance relationship is complex and several factors may influence it.

Screening performance is unlikely to be affected by volume alone, but rather

Figure 3

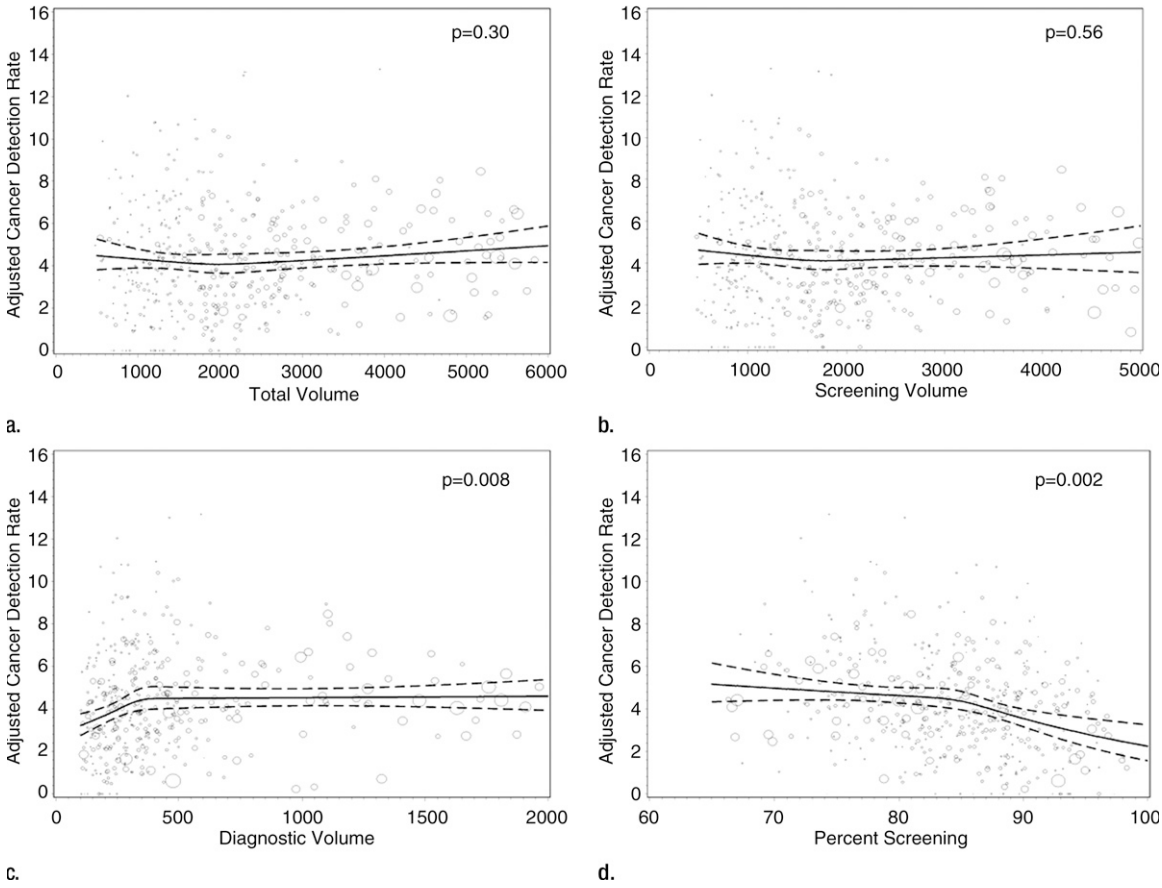

Figure 3: Graphs show adjusted CDRs according to interpretive volume, in terms of **(a)** total volume, **(b)** screening volume, **(c)** diagnostic volume, and **(d)** percentage of total mammograms that represented screening examinations. CDRs were adjusted for age and time since last mammogram. Lines = regression spline fit to adjusted rates; dashed lines = 95% CIs; and ○ = adjusted CDR, with size proportional to the number of screening mammograms used to measure performance. Smoothing splines had three knots placed at the 33rd, 50th, and 67th percentiles of the volume distribution; estimations were limited to total volume of 6000 or fewer mammograms, screening volume of 5000 or fewer mammograms, diagnostic volume of 2000 or fewer mammograms, and a screening focus of 65% or greater. Estimated mean adjusted performance is presented graphically, along with pointwise 95% CIs, with the curves being interpreted directly as the mean adjusted performance as a function of the volume measure. *P* values for the estimated curves correspond to omnibus tests of whether there is any association between mean adjusted performance and volume.

by a balance in the interpreted examination composition. Radiologists with greater screening focus had significantly lower sensitivities and CDRs and significantly lower false-positive rates. We expected that radiologists performing more diagnostic work-ups would have better performance, because of their involvement in seeing a case from screening through work-up, including possible involvement with interventional procedures (22,28). Many large groups decentralize their screening: General partners interpret screening mammograms in office practices where they also interpret other types of imaging studies, while designated breast imagers focus on screening and diagnostic breast imaging examinations. Although our results suggest clinically important differences for radiologists with greater diagnostic volumes, we cannot establish cause and effect and did not evaluate the additional potential influence of performing interventional procedures.

The techniques and skills required for interpreting different images differ and will also generate different performance measures—for example, recognizing normal and benign variants is important for false-positive rates, whereas detecting subtle cancer findings is important for sensitivity. Screening is performed to accurately identify individuals who need additional work-up, whereas diagnostic imaging is performed to accurately evaluate areas of suspicion. A previous study (15) found that radiologists with a greater diagnostic focus have higher screening false-positive rates, perhaps because they are more accustomed to higher cancer prevalence.

Interpretive volume collection and reporting would be required to change if volume requirements included minimal diagnostic interpretations.

The complexity of so many factors (eg, years of experience, number of cancers interpreted, screening vs diagnostic volume, training, and the innate skills of the interpreter) will continue to challenge researchers and policymakers. Does experience at high interpretive volume improve performance, or do radiologists who interpret more accurately choose to interpret high volumes? Radiologists who take 4 years to interpret 5000 screening mammograms may vary importantly in performance compared with radiologists who attain this volume in 1 year (22). Subspecialty training may also influence the experience-and-volume interplay. Our highest-volume radiologists included a mix of some with many years of experience and newer graduates with fellowship training, and within this group were varying volumes of screening and diagnostic interpretations.

Our results and conclusions are specific to screening performance, which comprises 80% of U.S. mammography. Whereas most prior studies examined only screening, we examined different volume measures and collected volume across all facilities where radiologists interpreted over 5 years. We modeled the association between volume interpreted in the prior year with performance in the following year, as opposed to including future volume measures in association with past performance. We based performance measures on actual practice, not test-set performance.

Statistical variability issues complicate measuring volume-performance outcomes. Cancer is rare in screening settings. In our study, radiologists detected a mean of 4.2 cancers per 1000 mammograms with a sensitivity of 85.2%. Because false-negative cases are rare (one per 1000 mammograms) and some are visible only in retrospect (29–32), it could take many years for a low-volume reader to miss a finding that an expert might identify. This is a smaller problem for high-volume readers. Low-volume radiologists see only a few cancers each year, which makes it difficult to measure sensitivity accurately in this group (33). Additionally, many believe that regular feedback improves performance, and radiologists who interpret any screening examination without the opportunity to see the results of their abnormal interpretations could not build on that experience. We could not explore the influence of feedback.

Breast cancer screening costs $3.6 billion annually in the United States (27). This does not include the costs of false-positive examination work-ups, which amount to approximately $1.6 billion per year, or avoid time, trouble, and anxiety for women. Our simulation estimated that the costs of false-positive findings would be reduced by $21.8 and $46.4 million (34,35) if the Food and Drug Administration required annual total volume requirements of greater than 1000 or fewer than 1500 mammograms, respectively. Basing volume requirements on annual screening volume and changing the minimum to greater than 1000 or fewer than 1500 mammograms would lower false-positive work-up costs by $35.6 million and $58.6 million, respectively.

There is no single "best" performance metric that can be used to help set policy. Our simulation results demonstrate that changing MQSA volume requirements or adding minimum numbers of screening and diagnostic examinations could result in modest improvements in some screening outcomes at a cost to others. An estimated 20% of radiologists interpret fewer than 1000 mammograms per year, but these account for only 6% of all U.S. mammograms (2). In our study, 17% of radiologists interpreted fewer than 1000 mammograms annually, and 38% interpreted fewer than 1500 mammograms annually. Workforce issues are crucial when considering interpretation requirements, because raising the minimum number of interpretations might cause some lower-volume radiologists to stop interpreting mammograms.

In conclusion, radiologists with higher annual volumes had clinically and statistically significantly lower false-positive rates with similar sensitivities as their colleagues with lower volumes. Radiologists with a greater screening focus had significantly lower sensitivities and CDRs and significantly lower false-positive rates. Recommending any increase in U.S. volume requirements will entail crucial decisions about the relative importance of cancer detection versus false-positive examinations and workforce issues, because changes could curtail workforce supply and women's access to mammography. To achieve higher sensitivity while lowering false-positive rates, further studies need to elucidate the interrelationships between training, experience, volume, and performance measures. Several requirements may need to be considered simultaneously, such as minimum volume in addition to combined minimum performance requirements (36).

Acknowledgments: The collection of cancer incidence data this study used 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://breastscreening.cancer.gov/work/acknowledgement.html*. We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study. We also thank Melissa Rabelhofer, BA, and Rebecca Hughes, BA, for their assistance in manuscript preparation and editing. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at *http://breastscreening.cancer.gov/*.

Disclosures of Potential Conflicts of Interest: **D.S.M.B.** No potential conflicts of interest to disclose. **M.L.A.** No potential conflicts of interest to disclose. **S.J.P.A.H.** No potential conflicts of interest to disclose. **E.A.S.** No potential conflicts of interest to disclose. **R.A.S.** No potential conflicts of interest to disclose. **P.A.C.** No potential conflicts of interest to disclose. **S.H.T.** No potential conflicts of interest to disclose. **R.D.R.** No potential conflicts of interest to disclose. **B.M.G.** No potential conflicts of interest to disclose. **T.L.O.** No potential conflicts of interest to disclose. **B.S.M.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: is a paid consultant on the Hologic Medical Advisory Board; institution received a grant from the National Institutes of Health for photoacoustic breast imaging; and received an honorarium from the University of Alabama at Birmingham as payment for lectures. Other relationships: none to disclose. **L.W.B.** No potential conflicts of interest to disclose. **B.C.Y.** No potential conflicts of interest to disclose. **J.G.E.** No potential conflicts of interest to disclose. **K.K.** No potential conflicts of interest to disclose. **D.L.M.** No potential conflicts of interest to disclose.

References

1. IARC Working Group on the Evaluation of Cancer-Preventive Strategies. Breast cancer screening. Lyon, France: IARC Press, 2002.
2. Institute of Medicine. Improving breast imaging quality standards. Washington, DC: National Academies Press, 2005.
3. 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.
4. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United kingdom. JAMA 2003;290(16):2129–2137.
5. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. J Natl Cancer Inst 2003;95(18):1384–1393.
6. Hofvind S, Vacek PM, Skelly J, Weaver DL, Geller BM. Comparing screening mammography for early breast cancer detection in Vermont and Norway. J Natl Cancer Inst 2008;100(15):1082–1091.
7. Department of Health UK. Breast Screening Programme, England: 1999-2000. http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/StatisticalWorkAreas/Statistical healthcare/DH_4015755. Updated February 8, 2007. Accessed August 25, 2009.
8. Warren Burhenne L. Screening Mammography Program of British Columbia standardized test for screening radiologists. Semin Breast Dis 2003;6(3):140–147.
9. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. Ann Oncol 2008;19(4):614–622.
10. 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.
11. Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000;215(2):563–567.
12. 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.
13. National Cancer Institute. Breast Cancer Surveillance Consortium Homepage. http://breastscreening.cancer.gov/. Updated November 16, 2009. Accessed March 5, 2010.
14. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96(24):1840–1850.
15. Smith-Bindman R, Chu P, Miglioretti DL, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst 2005;97(5):358–367.
16. 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.
17. 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.
18. American College of Radiology. 1998 MQSA (Mammography Quality Standards Act) final rule released. Radiol Manage 1998;20(4):51–55.
19. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006;241(1):55–66.
20. Hastie T, Tibshirani R, Friedman J. Basis expansions and regularization. In: The elements of statistical learning: data mining, inference and prediction. New York, NY: Springer-Verlag, 2001; 127–133.
21. Greenland S. Introduction to regression modeling. In: Rothman KJ, Greenland S, eds. Modern epidemiology. 2nd ed. Philadelphia, Pa: Lippincott-Raven, 1998; 401–434.
22. Miglioretti DL, Haneuse SJ, Anderson ML. Statistical approaches for modeling radiologists' interpretive performance. Acad Radiol 2009;16(2):227–238.
23. Harrell FE. General aspects of fitting regression models. In: Regression modeling strategies. New York, NY: Springer-Verlag, 2001; 18–24.
24. US Census Bureau. Table 1. Total Population by Age and Sex for the United States: 2000. http://www.census.gov/population/cen2000/phc-t08/phc-t-08.xls. Updated February 25, 2002. Accessed March 23, 2010.
25. Carney PA, Dietrich AJ, Freeman DH Jr, Mott LA. The periodic health examination provided to asymptomatic older women: an assessment using standardized patients. Ann Intern Med 1993;119(2):129–135.
26. Centers for Medicare & Medicaid Services. In the Medicare population in the year 2005, 11.4 million non-HMO women between the ages of 50 to 79 years of age received a reimbursed screening mammogram. http://www.cms.hhs.gov/PrevntionGenInfo/Downloads/mammography_age_2005.pdf. Accessed June 18, 2008.
27. GE Healthcare. Medicare reimbursement for mammography services. http://www.gehealthcare.com/usen/community/reimbursement/docs/MammographyOverview.pdf. Updated December 12, 2006. Accessed June 13, 2010.
28. 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.
29. Moberg K, Grundström H, Lundquist H, Svane G, Havervall E, Muren C. Radiological review of incidence breast cancers. J Med Screen 2000;7(4):177–183.
30. Britton PD, McCann J, O'Driscoll D, Hunnam G, Warren RM. Interval cancer peer review in East Anglia: implications for monitoring doctors as well as the NHS breast screening programme. Clin Radiol 2001;56(1):44–49.
31. Saarenmaa I, Salminen T, Geiger U, et al. The visibility of cancer on previous mammograms in retrospective review. Clin Radiol 2001;56(1):40–43.
32. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001;219(1):192–202.
33. Drye EE, Chen J. Evaluating quality in small-volume hospitals. Arch Intern Med 2008;168(12):1249–1251.
34. Burnside E, Belkora J, Esserman L. The impact of alternative practices on the cost and quality of mammographic screening in the United States. Clin Breast Cancer 2001;2(2):145–152.
35. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002;94(5):369–375.
36. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology 2010;255(2):354–361.