Title

National Performance Benchmarks for Screening Digital Breast Tomosynthesis: Update from the Breast Cancer Surveillance Consortium.

Journal

Radiology, 307(4)

ISSN

0033-8419

Authors

Lee, Christoph I

Abraham, Linn

Miglioretti, Diana L

et al.

Publication Date

2023-05-01

DOI

10.1148/radiol.222499

Peer reviewed

Radiology

ORIGINAL RESEARCH • BREAST IMAGING

National Performance Benchmarks for Screening Digital Breast Tomosynthesis: Update from the Breast Cancer Surveillance Consortium

Christoph I. Lee, MD • Linn Abraham, MS • Diana L. Miglioretti, PhD • Tracy Onega, PhD • Karla Kerlikowske, MD • Janie M. Lee, MD • Brian L. Sprague, PhD • Anna N. A. Tosteson, ScD • Garth H. Rauscher, PhD • Erin J. A. Bowles, MPH • Roberta M. diFlorio-Alexander, MD • Louise M. Henderson, PhD • for the Breast Cancer Surveillance Consortium

From the Department of Radiology, University of Washington School of Medicine, Hutchinson Institute for Cancer Outcomes Research, Fred Hutchinson Cancer Center, 825 Eastlake Ave E, LG-200, Seattle, WA 98109 (C.I.L., J.M.L.); Department of Health Systems & Population Health, University of Washington School of Public Health, Seattle, Wash (C.I.L.); Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, Wash (C.I.L., L.A., D.L.M., J.M.L., E.J.A.B.); Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, Davis, Calif (D.L.M.); Department of Population Health Sciences, and the Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah (T.O.); Department of Medicine, Department of Epidemiology and Biostatistics, and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, San Francisco, Calif (K.K.); Department of Surgery, Office of Health Promotion Research, Larner College of Medicine at the University of Vermont and University of Vermont Cancer Center, Burlington, Vt (B.L.S.); The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth and Norris Cotton Cancer Center, Lebanon, NH (A.N.A.T.); Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, Ill (G.H.R.); Department of Radiology, Geisel School of Medicine at Dartmouth, Lebanon, NH (R.M.d.A.); and Department of Radiology, University of North Carolina, Chapel Hill, NC (L.M.H.). Received September 28, 2022; revision requested November 16; revision received February 3, 2023; accepted February 20. **Address correspondence to** C.I.L. (email: *stophlee@uw.edu*).

Supported by the Breast Cancer Surveillance Consortium with funding from the National Cancer Institute (P01CA154292, R01CA262023, R50CA211115). Data collection was supported by the Patient-Centered Outcomes Research Institute Program Award (PCS-1504-30370), the National Cancer Institute (U54CA163303), and the Agency for Healthcare Research and Quality (R01 HS018366-01A1). Collection of Vermont Breast Cancer Surveillance System data was supported by the University of Vermont Cancer Center with funds generously awarded by the Lake Champlain Cancer Research Organization (032800). Cancer and vital status data collection was supported by several state public health departments and cancer registries *(https://www.bcsc-research.org/work/acknowledgement.html)*.

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

See also the editorial by Lee and Moy in this issue.

Radiology 2023; 307(4):e222499 • https://doi.org/10.1148/radiol.222499 • Content code: BR

Background: It is important to establish screening mammography performance benchmarks for quality improvement efforts.

Purpose: To establish performance benchmarks for digital breast tomosynthesis (DBT) screening and evaluate performance trends over time in U.S. community practice.

Materials and Methods: In this retrospective study, DBT screening examinations were collected from five Breast Cancer Surveillance Consortium (BCSC) registries between 2011 and 2018. Performance measures included abnormal interpretation rate (AIR), cancer detection rate (CDR), sensitivity, specificity, and false-negative rate (FNR) and were calculated based on the American College of Radiology Breast Imaging Reporting and Data System, fifth edition, and compared with concurrent BCSC DM screening examinations, previously published BCSC and National Mammography Database benchmarks, and expert opinion acceptable performance ranges. Benchmarks were derived from the distribution of performance measures across radiologists (n = 84 or n = 73 depending on metric) and were presented as percentiles.

Results: A total of 896 101 women undergoing 2 301 766 screening examinations (458 175 DBT examinations [median age, 58 years; age range, 18–111 years] and 1 843 591 DM examinations [median age, 58 years; age range, 18–109 years]) were included in this study. DBT screening performance measures were as follows: AIR, 8.3% (95% CI: 7.5, 9.3); CDR per 1000 screens, 5.8 (95% CI: 5.4, 6.1); sensitivity, 87.4% (95% CI: 85.2, 89.4); specificity, 92.2% (95% CI: 91.3, 93.0); and FNR per 1000 screens, 0.8 (95% CI: 0.7, 1.0). When compared with BCSC DM screening examinations from the same time period and previously published BCSC and National Mammography Database performance benchmarks, all performance measures were higher for DBT except sensitivity and FNR, which were similar to concurrent and prior DM performance measures. The following proportions of radiologists achieved acceptable performance ranges with DBT: 97.6% for CDR, 91.8% for sensitivity, 75.0% for AIR, and 74.0% for specificity.

Conclusion: In U.S. community practice, large proportions of radiologists met acceptable performance ranges for screening performance metrics with DBT.

Supplemental material is available for this article.

Mammography is the primary breast cancer screening test, and it has been shown to reduce breast cancer mortality (1). All major policy bodies, including the United States Preventive Services Task Force and American Cancer Society, recommend mammography for routine screening of women at average risk for breast cancer (2,3). In the time since most screening mammography randomized controlled trials were conducted, mammographic imaging technology transitioned from screen-film mammography to digital mammography (DM) (4). After gaining Food and Drug Administration approval in 2011, digital breast tomosynthesis (DBT) was rapidly adopted in the United States and is now the most common breast cancer screening modality (5–7). As of September

Abbreviations

AIR = abnormal interpretation rate, BCSC = Breast Cancer Surveillance Consortium, BI-RADS = Breast Imaging Reporting and Data System, CDR = cancer detection rate, DBT = digital breast tomosynthesis, DM = digital mammography, FNR = false-negative rate, PPV = positive predictive value, PPV_1 = abnormal interpretation, PPV_2 = recommended for tissue diagnosis, PPV_3 = PPV of biopsy performed

Summary

Digital breast tomosynthesis in U.S. community practice has improved breast cancer screening performance, with improvements or stability in all measures compared with previously published mammography performance benchmarks.

Key Results

- In this retrospective study involving 896 101 women who underwent 2 301 766 screening examinations, the digital breast tomosynthesis (DBT) mean abnormal interpretation rate (AIR) was 8.3% (95% CI: 7.5, 9.3), cancer detection rate (CDR) was 5.8 per 1000 examinations (95% CI: 5.4, 6.1), sensitivity was 87.4% (95% CI: 85.2, 89.4), and specificity was 92.2% (95% CI: 91.3, 93.0).
- With DBT, 97.6%, 91.8%, 75.0%, and 74.0% of radiologists assessed achieved the recommended acceptable performance ranges for CDR, sensitivity, AIR, and specificity, respectively.

2022, 84% of all mammography screening facilities in the United States had DBT units (8).

Breast cancer mortality reduction from routine screening is contingent on radiologists' interpretive performance. Since the Mammography Quality Standards Act was enacted in 1992, screening facilities and interpreting radiologists have been required to meet minimum quality standards (9). The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) is the practice standard in the United States and regularly updates published screening performance benchmarks that aid in quality improvement efforts (10).

In 2006, the Breast Cancer Surveillance Consortium (BCSC) began publishing screening performance benchmarks that have, in part, formed the basis of the performance benchmarks for the BI-RADS atlas (11). The BCSC is uniquely positioned to assess trends in screening mammography performance in U.S. community settings given its large diverse population and linkage to state and regional tumor registries that guarantee complete capture of cancer outcomes. With changes in technology over time, the BCSC has periodically updated their U.S. community screening performance benchmarks, most recently for modern DM (12) and breast MRI (13). The purpose of this study was to establish performance benchmarks for DBT screening and assess mammography-based performance trends over time in U.S. community practice.

Materials and Methods

Study Patients

We included data for all adult women aged at least 18 years who underwent DBT or DM screening at a BCSC facility between 2011 and 2018. We selected 2011 as the study start date since DBT obtained Food and Drug Administration approval (5) and some BCSC facilities began offering DBT that year. Our study end date was 2018 because some state and regional tumor registries have a lag of up to 3 years for complete cancer outcomes reporting (cancer capture through 2021). We excluded mammograms obtained in women younger than 18 years, those obtained for a nonscreening indication, those obtained within 9 months of a prior mammogram, those with a BI-RADS 6 (known malignancy) assessment, those that had a known issue with data quality, and those obtained in patients who had less than 1 year of complete cancer data available or who had a missing initial result or cancer status (Fig 1). To describe changes in screening performance measures over time, we report previously published BCSC screening performance benchmarks from 2007 to 2013 (11,12) for descriptive comparison purposes only.

Data Collection and Definitions

Screening examination data were obtained from five BCSC registries (Carolina Mammography Registry, Metropolitan Chicago Breast Cancer Registry, New Hampshire Mammography Network, San Francisco Mammography Registry, and Vermont Breast Cancer Surveillance System). The BCSC registries are collectively representative of the U.S. general population in terms of age, race, and ethnicity (12,14,15). Each registry links their mammography data with state or regional tumor registries and pathology databases for complete cancer capture. Data are pooled at a central BCSC Statistical Coordinating Center. Each BCSC registry and Statistical Coordinating Center received institutional review board approval for active or passive consenting processes or a waiver of consent to enroll women, link and pool data, and perform analyses. All procedures were compliant with the Health Insurance Portability and Accountability Act, and a federal certificate of confidentiality provides further protections to participants.

All BCSC registries systematically collect detailed breast cancer risk factor data for each woman at the time of screening, including age, self-reported race, self-reported ethnicity, family history of breast cancer, personal history of breast cancer, date of last mammography, menopausal status, and breast biopsy history. The BCSC 5-year risk score is calculated for each woman aged 35–74 years without a personal history of breast cancer based on age, race, ethnicity, family history, breast density, and breast biopsy history (16). All registry data include the BI-RADS assessments, recommendations, and breast density determination made by interpreting radiologists for each examination.

This study follows the BI-RADS fifth edition definitions for all screening performance metrics (17). For measures other than positive predictive value (PPV) (examinations with abnormal interpretation [PPV_1], recommendations for tissue diagnosis [PPV_2], and PPV of biopsy performed [PPV_3]), a positive screening examination was defined as one with an initial assessment of BI-RADS 0, 3, 4, or 5. For PPV_2 and PPV_3, a positive screening examination was defined as having a final assessment (after diagnostic evaluation) of BI-RADS 4 or 5. For screening examinations with an initial BI-RADS 0 assessment, the final assessment was determined from diagnostic imaging records up to 90 days after the screening examination. In accordance with

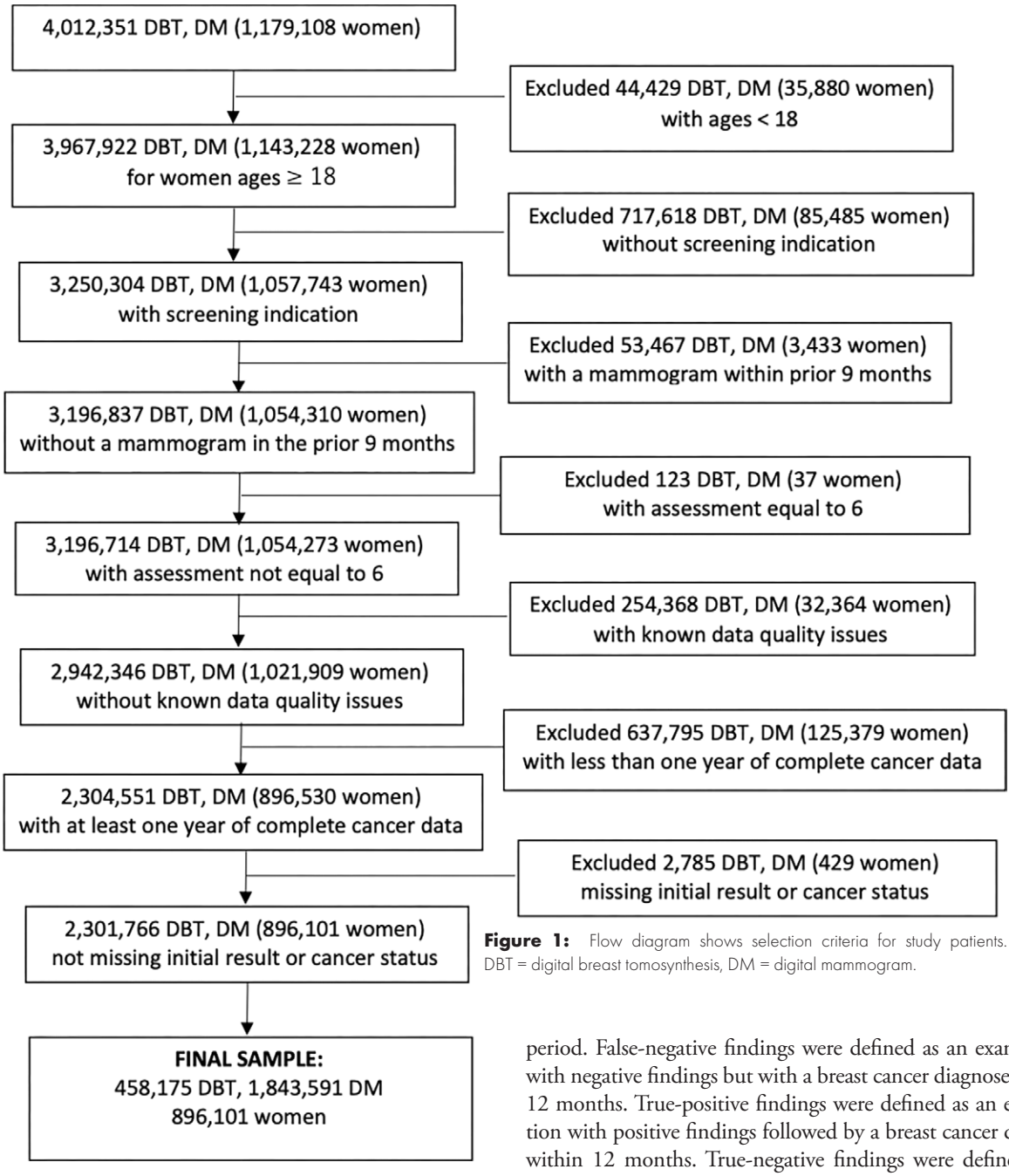

Figure 1: Flow diagram shows selection criteria for study patients. DBT = digital breast tomosynthesis, DM = digital mammogram.

BI-RADS auditing guidance, all examinations with a BI-RADS 6 assessment (biopsy-proven malignancies) were excluded from our analyses. Women were considered to have breast cancer if there was a recorded diagnosis of invasive breast cancer or ductal carcinoma in situ within 12 months after screening mammography and before the next screening examination. Only mammograms with at least 12 months of complete cancer follow-up data were included in our analysis.

Outcomes

Per the BI-RADS fifth edition definitions, false-positive findings were defined as an examination with positive findings but with no breast cancer diagnosed during the 12-month follow-up period. False-negative findings were defined as an examination with negative findings but with a breast cancer diagnosed within 12 months. True-positive findings were defined as an examination with positive findings followed by a breast cancer diagnosis within 12 months. True-negative findings were defined as an examination with negative findings with no breast cancer diagnosis within 12 months. Abnormal interpretation rate (AIR, recall rate) was calculated by dividing the number of positive (recalled) screening examinations (excluding initial BI-RADS 0 assessments for technical repeat or comparison with prior examinations) by the total number of screens. Cancer detection rate (CDR) was calculated for all cancers and separately for invasive cancers by dividing the number of examinations with true-positive findings by the total number of screening examinations. False-negative rate (FNR) was calculated by dividing the number of examinations with false-negative findings by the total number of screen examinations. Sensitivity was calculated by dividing the number of examinations with true-positive findings by the total number of examinations associated with cancer (ie, those with true-positive or false-negative findings). Specificity was

calculated by dividing the number of examinations with true-negative findings by the total number of examinations without cancer (ie, those with true-negative or false-positive findings). Characteristics of cancers detected included American Joint Committee on Cancer (AJCC) anatomic stage, AJCC prognostic disease stage, axillary nodal status, minimal cancer (ductal carcinoma in situ or ≤10 mm invasive cancer), advanced cancer (prognostic stage IIb or higher; if missing, then anatomic stage IIb or higher), and invasive cancer size.

Statistical Analyses

We report the same clinically relevant descriptive measures as in a prior performance benchmark article to facilitate comparison over time (11,12). Overall rates were based on the entire sample. We computed 95% CIs using generalized estimated equations with a working independence correlation structure to account for clustering within women, radiologists, and facilities (18,19). To reduce variability from low-volume interpreters, we included only radiologists who, during the study period, interpreted at least 1000 DBT studies for DBT performance benchmarks and at least 1000 DM studies for DM performance measures associated with cancer yield (CDR, FNR, and PPVs). We also required at least 10 cancer events detected with DBT and DM for sensitivity, specificity, and characteristics of tumors detected with each modality. We assessed the distribution of radiologists for each performance benchmark and calculated the median and IQR, as well as the 10th and 90th percentiles. The proportion of radiologists within acceptable ranges previously established by expert opinion is also reported (20). All analyses were performed by a statistician (L.A.) using SAS software (version 9.4; SAS Institute). Graphic presentations were produced using STATA 12 software (StataCorp).

Results

Patient Characteristics

After applying exclusion criteria (35 880 women younger than 18 years, 85 485 with a nonscreening indication, 3433 with a mammogram obtained in the prior 9 months, 37 with BI-RADS 6 assessment, 32 364 with a known image quality issue, 125 379 with less than 1 year complete cancer capture, 429 missing initial results or cancer status), our final study cohort included data from 896 101 women who underwent a total of 2 301 766 screening examinations (458 175 DBT and 1 843 591 DM) between 2011 and 2018 (Fig 1). A total of 525 unique interpreting radiologists from 89 unique imaging facilities were included in our analysis.

The median age of patients for all examinations was 58.0 years for both DBT and DM screening (age range, 18–111 years for DBT; 18–109 years for DM) (Table 1). Of all examinations performed during the study period, a greater proportion of DBT versus DM examinations were performed in White women (DBT, 84.6%; DM, 65.2%), women with a family history of breast cancer (DBT, 21.1%; DM, 16.9%), women with a personal history of breast cancer (DBT, 9.6%; DM, 5.8%), women with prior benign breast biopsy (DBT, 27.4%; DM, 23.4%), and women with a 5-year breast cancer risk of 2.5% or more (DBT, 12.9%; DM, 9.3%). No differences in the proportions of women undergoing DBT versus DM were observed based on breast density and time since last mammography.

Table 1: Clinical Characteristics for Screening DBT and DM Examinations from the BCSC, 2011–2018

	DBT		DM	
Characteristic	All Examinations	Examinations with Cancer	All Examinations	Examinations with Cancer
No. of examinations	458 175	3018	1 843 591	11 212
No. of women	239 889	2994	825 122	11 153
Age (y)*	58.0 (50.0, 67.0)	63.0 (54.0, 71.0)	58.0 (50.0, 67.0)	63.0 (54.0, 71.0)
Age group				
<30 years	216 (0.0)	1 (0.0)	614 (0.0)	1 (0.0)
30–39 years	5604 (1.2)	18 (0.6)	23 148 (1.3)	83 (0.7)
40–49 years	100 970 (22.0)	409 (13.6)	428 806 (23.3)	1673 (14.9)
50–59 years	141 641 (30.9)	773 (25.6)	558 559 (30.3)	2793 (24.9)
60–69 years	129 485 (28.3)	963 (31.9)	497 512 (27.0)	3546 (31.6)
70–79 years	65 153 (14.2)	683 (22.6)	253 975 (13.8)	2273 (20.3)
≥80 years	15 106 (3.3)	171 (5.7)	80 977 (4.4)	843 (7.5)
Race†				
Asian or Pacific Islander	17 805 (4.0)	103 (3.5)	260 052 (14.7)	1502 (13.7)
Black	30 069 (6.7)	188 (6.4)	203 763 (11.5)	1417 (13.0)
Hispanic or Latina	14 035 (3.1)	52 (1.8)	118 789 (6.7)	518 (4.7)
White	378 094 (84.6)	2548 (86.6)	115 4850 (65.2)	7307 (66.8)
Other	7112 (1.6)	52 (1.8)	33 181 (1.9)	192 (1.8)
Unknown	11 060 (2.4)	75 (2.5)	72 956 (4.0)	276 (2.5)

(Table 1 continues)

Table 1 *(continued)*: Clinical Characteristics for Screening DBT and DM Examinations from the BCSC, 2011–2018

Characteristic	DBT		DM	
	All Examinations	Examinations with Cancer	All Examinations	Examinations with Cancer
Family history of breast cancer				
Yes	90 285 (21.1)	833 (29.5)	302 170 (16.9)	2551 (23.7)
No	336 836 (78.9)	1988 (70.5)	1 481 919 (83.1)	8209 (76.3)
Unknown	31 054 (6.8)	197 (6.5)	59 502 (3.2)	452 (4.0)
Personal history of breast cancer				
Yes	36 820 (9.6)	551 (22.9)	95 135 (5.8)	1393 (15.6)
No	346 470 (90.4)	1850 (77.1)	1 547 891 (94.2)	7515 (84.4)
Unknown	74885 (16.3)	617 (20.4)	200 565 (10.9)	2304 (20.5)
History of prior benign breast biopsy				
Yes	125 608 (27.4)	1275 (42.2)	432 322 (23.4)	3897 (34.8)
No‡	332 567 (72.6)	1743 (57.8)	1 411 269 (76.6)	7315 (65.2)
Time since last mammogram				
No previous mammogram	15 343 (3.5)	110 (3.8)	101 618 (5.8)	633 (6.0)
Within a year (0–11 mo)	5223 (1.2)	40 (1.4)	24 462 (1.4)	161 (1.5)
1–2 years (12–35 mo)	384 784 (87.1)	2399 (82.4)	1 472 095 (83.9)	8300 (78.4)
≥3 years (≥36 mo)	36 376 (8.2)	363 (12.5)	155 500 (8.9)	1488 (14.1)
Unknown	16 449 (3.6)	106 (3.5)	89 916 (4.9)	630 (5.6)
Menopausal status				
Premenopause	103 156 (26.0)	465 (17.2)	511 070 (30.4)	2196 (21.0)
Postmenopause	274 806 (69.1)	2143 (79.2)	1 083 098 (64.5)	7832 (75.0)
Surgical menopause	19 504 (4.9)	99 (3.7)	84 588 (5.0)	414 (4.0)
Unknown	60 709 (13.3)	311 (10.3)	164 835 (8.9)	770 (6.9)
Breast density				
Almost entirely fat	46 608 (10.5)	212 (7.3)	168 140 (9.9)	764 (7.5)
Scattered fibroglandular	212 239 (47.7)	1329 (45.6)	773 909 (45.4)	4665 (45.8)
Heterogeneously dense	156 197 (35.1)	1183 (40.6)	639 401 (37.5)	4056 (39.8)
Extremely dense	30 145 (6.8)	190 (6.5)	121 971 (7.2)	699 (6.9)
Unknown	12 986 (2.8)	104 (3.4)	140 170 (7.6)	1028 (9.2)
BCSC 5-year risk*§	1.4 (0.9, 2.0)	1.7 (1.2, 2.4)	1.3 (0.9, 1.8)	1.5 (1.1, 2.1)
0% to less than 0.99%	103 551 (27.4)	322 (15.5)	473 675 (32.3)	1506 (19.6)
1.00%–1.66%	139 330 (36.9)	676 (32.5)	567 345 (38.7)	3017 (39.3)
1.67%–2.49%	86 225 (22.8)	603 (29.0)	287 470 (19.6)	1878 (24.5)
2.50%–3.99%	41 335 (10.9)	378 (18.2)	118 640 (8.1)	1064 (13.9)
≥4.00%	7403 (2.0)	101 (4.9)	17 450 (1.2)	210 (2.7)
Unknown	494 (0.1)	7 (0.3)	6548 (0.4)	31 (0.4)

Note.—Unless otherwise stated, data are presented as numbers of examinations, with percentages in parentheses. BCSC = Breast Cancer Surveillance Consortium, DBT = digital breast tomosynthesis, DM = digital mammography.

* Data are medians, and data in parentheses are IQRs

† All categories except Hispanic are non-Hispanic. The term *Other* includes Native American, Alaskan Native, Multiracial, and Other.

‡ No includes unknowns.

§ Calculated for examinations among women aged 35–74 years, without a personal history of breast cancer, and not missing breast density.

Screening Performance Measures

The DBT AIR was 8.3% (95% CI: 7.5, 9.3), while the DM AIR was 10.3% (95% CI: 9.4, 11.3) (Table 2). The DBT CDR was 5.8 per 1000 examinations (95% CI: 5.4, 6.1), while the DM CDR was 5.3 per 1000 examinations (95% CI: 5.0, 5.6). The DBT invasive CDR was 4.5 per 1000 examinations (95% CI: 4.2, 4.8), while the DM CDR was 3.9 per 10000 examinations (95% CI: 3.7, 4.1). DBT sensitivity was 87.4% (95% CI: 85.2, 89.4), which was similar to DM sensitivity (87.6% [95% CI: 86.3, 88.8]), but DBT specificity was higher than DM specificity (DBT, 92.2% [95% CI: 91.3, 93.0]; DM, 90.2% [95% CI: 89.2, 91.1]). Both the DBT FNR and the DM FNR were 0.8 per 1000 examinations. PPVs were all higher with DBT than with DM

Table 2: Performance Measures for Screening DBT and DM Examinations from the Breast Cancer Surveillance Consortium, 2011–2018

Performance Measure	DBT	DM
Abnormal interpretation (recall) rate (%)	8.3 (7.5, 9.3)	10.3 (9.4, 11.3)
No. of abnormal interpretations	38203	189552
Total no. of examinations	458175	1843591
CDR (per 1000 examinations)	5.8 (5.4, 6.1)	5.3 (5.0, 5.6)
No. of TP examinations	2638	9822
Total no. of examinations	458175	1843591
Invasive CDR (per 1000 examinations)	4.5 (4.2, 4.8)	3.9 (3.7, 4.1)
No. of invasive TP examinations	2072	7182
Total no. of examinations	458175	1843591
Sensitivity (%)	87.4 (85.2, 89.4)	87.6 (86.3, 88.8)
No. of TP examinations	2638	9822
No. of examinations with cancers	3018	11212
Specificity (%)	92.2 (91.3, 93.0)	90.2 (89.2, 91.1)
No. of TN examinations	419592	1652649
No. of examinations without cancer	455157	1832379
FNR (per 1000 examinations)	0.8 (0.7, 1.0)	0.8 (0.7, 0.9)
No. of FN examinations	380	1390
Total no. of examinations	458175	1843591
PPV_1 (abnormal interpretation) (%)	6.9 (6.3, 7.6)	5.2 (4.8, 5.6)
No. of cancers	2638	9822
Initial BI-RADS 0, 3, 4, or 5 assessment	38203	189552
PPV_2 (biopsy recommended) (%)	32.2 (29.2, 35.3)	27.9 (25.2, 30.9)
No. of cancers	2429	8996
Final BI-RADS 4 or 5 assessment	7546	32198
PPV_3 (biopsy performed) (%)	35.5 (32.3, 38.9)	31.7 (28.9, 34.8)
No. of cancers	2303	8437
Final BI-RADS 4 or 5 assessment and biopsy	6490	26587

Note.—Data in parentheses are 95% CIs and are based on binomial confidence limits. CDR = cancer detection rate, DBT = digital breast tomosynthesis, DM = digital mammography, FN = false negative, FNR = false-negative rate, PPV = positive predictive value, TN = true negative, TP = true positive.

screening. DBT PPV_1 was 6.9% (95% CI: 6.3, 7.6) versus 5.2% (95% CI: 4.8, 5.6) for DM; DBT PPV_2 was 32.2% (95% CI: 29.2, 35.3) versus 27.9% (95% CI: 25.2, 30.9) for DM; and DBT PPV_3 was 35.5% (95% CI: 32.2, 38.9) versus 31.7% (95% CI: 28.9, 34.8) for DM.

Cancer Characteristics

A total of 12460 breast cancers were screen detected (2638 with DBT, 9822 with DM) (Table 3). DBT screen-detected cancers, in comparison with DM-detected cancers, were more likely to be invasive rather than ductal carcinoma in situ (DBT, 78.5% invasive; DM, 73.1% invasive). The median size of cancers detected with DBT was 12.0 mm (IQR, 7.0–18.0 mm), and those detected with DM were similar in size (median size, 13.0 mm; IQR, 8.0–20.0 mm). The percentage of minimal cancers (ductal carcinoma in situ or invasive cancers ≤10 mm) were also similar between DBT (54.6%) and DM (55.6%). However, cancers detected with DBT versus those detected with DM were less frequently categorized by worse prognostic stage characteristics: anatomic stage IIb or higher (DBT, 7.5%, DM, 8.9%), prognostic pathologic stage IIb or higher (DBT, 2.8%, DM, 3.2%), and node-positive disease (DBT, 21.2%; DM, 27.8%).

Performance Trends over Time

Table 4 outlines the current and previously published screening performance benchmarks for BCSC (current, 2011–2018; previous, 2007–2013) and for the American College of Radiology National Mammography Database (current, 2008–2014; previous, 2008–2012). Compared with the DM BCSC screening performance benchmarks from 2007 to 2013 (12), DBT AIR is lower (8.3% [95% CI: 7.5, 9.3] vs 11.6% [95% CI: 11.5, 11.6]), CDR is higher (5.8 per 1000 examinations [95% CI: 5.4, 6.1] vs 5.1 per 1000 examinations [95% CI: 5.0, 5.2]), and specificity is higher (92.2% [95% CI: 91.3, 93.0] vs 88.9% [95% CI: 88.8, 88.9]). DBT sensitivity and FNR were similar to DM BCSC benchmarks (both 2007–2013 and 2011–2018). PPV_1, PPV_2, and PPV_3 were higher in the DBT and DM benchmark cohorts from 2011 to 2018 than in the DM benchmark cohort from 2007 to 2013 (Table 4).

Radiologists Performing within Acceptable Ranges

Of 525 total radiologists, 249 radiologists from 82 imaging facilities interpreted 1000 or more screening DM studies, and 84 radiologists from 33 facilities interpreted 1000 or more screening DBT studies during the study period. Overall, the number of radiologists performing within an acceptable range was higher for DBT screening than for DM screening, except for sensitivity, which was lower (20) (Fig 2; Figs S1, S2). For DBT, the median AIR was 8.1% (IQR, 6.4%–10.3%), with 75.0% of radiologists falling within the acceptable performance range of 5%–12%. The median CDR was 5.5 per 1000 examinations (IQR, 4.7–6.6), with 97.6% of radiologists meeting acceptable performance (CDR ≥2.5 per 1000 examinations). The median sensitivity was 89.2% (IQR, 83.3%–92.9%), with 91.8% of radiologists meeting acceptable performance (>75% sensitivity). The median specificity was 92.4% (IQR, 90.5%–93.9%), with 74.0% of radiologists meeting acceptable performance (88%–95% specificity). The median FNR was 0.7 per 1000 examinations (IQR, 0.4–1.1).

For DBT PPVs (Fig 3), the median PPV_1 was 7.2% (IQR, 5.6–9.3), with 53.6% of radiologists meeting acceptable performance (range, 3%–8%). Median PPV_2 was 30.6% (IQR, 26.2–40.0), with 63.5% of radiologists meeting

Table 3: Characteristics of Cancers Detected with Screening DBT and DM Examinations from the Breast Cancer Surveillance Consortium, 2011–2018

Characteristic	DBT (*n* = 2638)	DM (*n* = 9822)
Histologic type		
DCIS	566 (21.5)	2640 (26.9)
Invasive	2072 (78.5)	7182 (73.1)
Invasive cancer size (mm)*	12.0 (7.0–18.0)	13.0 (8.0–20.0)
1–5 mm	281 (14.1)	978 (14.0)
6–10 mm	547 (27.5)	1745 (24.9)
11–15 mm	507 (25.5)	1716 (24.5)
16–20 mm	253 (12.7)	971 (13.9)
>20 mm	399 (20.1)	1595 (22.8)
Unknown	85 (4.1)	177 (2.5)
Minimal cancer		
DCIS or invasive cancer ≤10 mm	1159 (45.4)	4282 (44.4)
Invasive cancer >10 mm	1394 (54.6)	5363 (55.6)
Unknown	85 (3.2)	177 (1.8)
Advanced cancer†		
No	2397 (93.6)	8962 (92.5)
Yes	163 (6.4)	725 (7.5)
Unknown	78 (3.0)	135 (1.4)
Invasive axillary lymph node status		
Negative	1595 (78.8)	5104 (72.2)
Positive	428 (21.2)	1970 (27.8)
Unknown	49 (2.4)	108 (1.5)
AJCC anatomic stage		
0	566 (22.4)	2640 (27.4)
I	1446 (57.2)	4868 (50.6)
IIa	324 (12.8)	1264 (13.1)
IIb	101 (4.0)	455 (4.7)
III	71 (2.8)	331 (3.4)
IV	18 (0.7)	69 (0.7)
Unknown	112 (4.2)	195 (2.0)
AJCC prognostic pathologic stage‡		
0	566 (25.5)	2640 (30.5)
I	1526 (68.7)	5467 (63.3)
IIa	66 (3.0)	256 (3.0)
IIb	20 (0.9)	74 (0.9)
III	25 (1.1)	137 (1.6)
IV	18 (0.8)	69 (0.8)
Unknown	417 (15.8)	1179 (12.0)

Note.—Unless otherwise indicated, data are presented as numbers of cancers, with percentages in parentheses. AJCC = American Joint Committee on Cancer, DBT = digital breast tomosynthesis, DCIS = ductal carcinoma in situ, DM = digital mammography.

* Data are medians, and data in parentheses are IQRs.

† Prognostic pathologic stage greater than or equal to II, if missing then anatomic stage IIb or higher.

acceptable performance (range, 20%–40%). Median PPV_3 was 35.7% (IQR, 29.4%–42.5%). For radiologists who detected 10 or more cancers with DBT, the median percentage of minimal cancer detected was 55.6% (IQR, 45.5%–62.7%), the median percentage of node-negative cancers detected was 79.3% (IQR, 69.2–87.4), and the median size of detected invasive cancers was 15.4 mm (IQR, 13.7–18.2 mm).

Among radiologists meeting volume thresholds for each modality, a larger proportion of radiologists were within acceptable performance ranges for AIR, CDR, and specificity with DBT compared with previously published proportions of radiologists within acceptable ranges with DM. For radiologists meeting the study period modality-specific examination volume criteria, a larger proportion of radiologists met acceptable ranges with DBT (34.2%) versus DM (27.9%) for all screening performance measures (Table S1).

Discussion

With DBT now the most popular breast cancer screening modality in the United States, our study objective was to evaluate screening DBT performance in community practice. We found that the AIR was lower (8.3% with DBT vs 11.6% with DM from 2007 to 2013), while the CDR was modestly higher (5.8 per 1000 examinations with DBT vs 5.1 per 1000 examinations with DM from 2007 to 2013), with no difference in the FNR. This translates to similar sensitivity with higher specificity for DBT versus DM screening. Moreover, all PPVs are higher with DBT, suggesting higher cancer yields for callbacks and women with screen-detected abnormalities recommended for and undergoing breast biopsy.

DBT screening also appears to reveal more invasive cancers than DM screening. The invasive cancers detected tended to be early stage and node negative. These findings are consistent with a recent BCSC analysis that compared DBT with DM and found that DBT was associated with reduced risk of advanced cancer at diagnosis among women with extremely dense breasts and high 5-year cancer risk (≥1.67%) compared with DM screening (21).

Of note, U.S. community screening performance appears to have steadily improved over time, with changes in both screening technology and imaging modality experience. We found that not only have performance measures improved with DBT over DM, but more recent BCSC DM performance (from 2011 to 2018) has exceeded previously reported BCSC DM performance (from 2007 to 2013) (12). For instance, screening DM AIR was lower (10.3% vs 11.6%) and specificity was higher (90.2% vs 88.9%) between the current and prior BCSC DM performance screening benchmarks (11,12).

Finally, we found that a larger proportion of radiologists are meeting or exceeding screening performance benchmarks for AIR, CDR, and specificity with DBT versus DM. This is a reassuring result, as there is minimal additional training required for DBT interpretation (8 hours of instruction or case review) (5). These promising trends are also likely a reflection

Table 4: Comparison of Current and Previous Performance Benchmarks for Screening DBT and DM Examinations

Measurement	DBT BCSC Benchmarks 2011–2018	DM BCSC Benchmarks 2011–2018	BCSC Benchmarks 2007–2013*	ACR NMD 2008–2012†	ACR NMD 2008–2014‡
AIR (recall rate) (%)	8.3 (7.5, 9.3)	10.3 (9.4, 11.3)	11.6 (11.5, 11.6)	10	9.6 (9.6, 9.7)
CDR (per 1000 examinations)	5.8 (5.4, 6.1)	5.3 (5.0, 5.6)	5.1 (5.0, 5.2)	3.43	3.7 (3.7, 3.8)
Sensitivity (%)	87.4 (85.2, 89.4)	87.6 (86.3, 88.8)	86.9 (86.3, 87.6)	NA	NA
Specificity (%)	92.2 (91.3, 93.0)	90.2 (89.2, 91.1)	88.9 (88.8, 88.9)	NA	NA
FNR (per 1000 examinations)	0.8 (0.7, 1.0)	0.8 (0.7, 0.9)	0.8 (0.7, 0.8)	NA	NA
PPV_1 (abnormal interpretation)	6.9 (6.3, 7.6)	5.2 (4.8, 5.6)	4.4 (4.3, 4.5)	NA	NA
PPV_2 (recommendation for tissue diagnosis)	32.2 (29.2, 35.3)	27.9 (25.2, 30.9)	25.6 (25.1, 26.1)	18.5	20 (20, 21)
PPV_3 (biopsy performed)	35.5 (32.2, 38.9)	31.7 (28.9, 34.8)	28.6 (28.0, 29.3)	29.2	29 (28, 29)

Note.—Data in parentheses are 95% CIs. ACR = American College of Radiology, AIR = abnormal interpretation rate, BCSC = Breast Cancer Surveillance Consortium, CDR = cancer detection rate, DBT = digital breast tomosynthesis, DM = digital mammography, FNR = false-negative rate, NA = not attainable due to lack of cancer registry linkage, NMD = National Mammography Database, PPV = positive predictive value.

* Data are from Lehman et al (12).

† Data are from the American College of Radiology (17).

‡ Data are from Lee et al (25).

of the relatively rapid learning curve for U.S. community radiologists when transitioning from DM to DBT (22). Nevertheless, there remains wide variability in individual performance across radiologists. For instance, in a prior study, we showed that 14.3% of BCSC radiologists had higher recall rates with DBT versus DM (23).

Direct comparison of all reported BCSC performance benchmarks to those published by the National Mammography Database is not possible (24,25). Since the National Mammography Database does not perform cancer registry linkage for long-term cancer outcomes, only a subset of performance metrics can be calculated from their data (AIR, CDR, and PPV). In contrast, all BCSC registries have near-complete cancer capture through linkage with state and regional tumor registries in addition to high completeness of pathology results from breast biopsies. This unique aspect of the BCSC registries allows for calculating key performance metrics, such as sensitivity, specificity, and FNR. Other strengths include a large screening population that is similar to the U.S. general screening population by age, race, and ethnicity (6,12), as well as diversity in geographic reach and types of screening facilities providing data (26,27).

Our analysis corroborates prior studies that have shown improved screening performance and outcomes with DBT in U.S. screening cohorts (6,21,23,28,29). Given improvements in screening performance over time with DBT, it may be time to reconsider the acceptable performance ranges for radiologists currently used by facilities for quality improvement efforts. These acceptable performance ranges, first established based on a mix of screen-film and digital mammograms (20), are likely outdated and are not as applicable to current screening practice. With the observed overall improvements across several screening performance metrics over time and with newer technology, the screening community could promote improved quality of care by focusing on meeting multiple concurrent benchmarks with the wider DBT adoption (eg, maintaining CDR while bringing AIR and PPV_1 into acceptable ranges, simultaneously).

Our study had some limitations. First, we limited our radiologist performance analysis to only those with sufficient modality-specific interpretive volumes. Coupled with relatively low rates of breast cancer in the general screening population, our performance measures for cancers detected excluded a substantial proportion of radiologists. Thus, radiologists with lower screening volumes may not achieve as high a performance as those with adequate DBT screening volumes reflected in these reported DBT benchmarks (22,30). Second, direct comparison of DBT with DM performance may be hampered by selection bias, with women at higher risk opting for the newer screening modality. However, accounting for population differences with inverse probability weighting derived from propensity scores, we found recall rate was lower with DBT versus DM, and early-stage invasive cancer was borderline significantly higher (21). Third, comparisons made with prior BCSC screening performance benchmarks were purely descriptive, and any noted differences could be due to changes in the study population over time. Finally, the impact of improved screening performance with DBT on long-term outcomes, such as breast cancer mortality, is currently unknown.

The data presented in this study can be used by individual screening facilities and radiologists for quality improvement efforts. It can also provide reference data and baseline metrics to

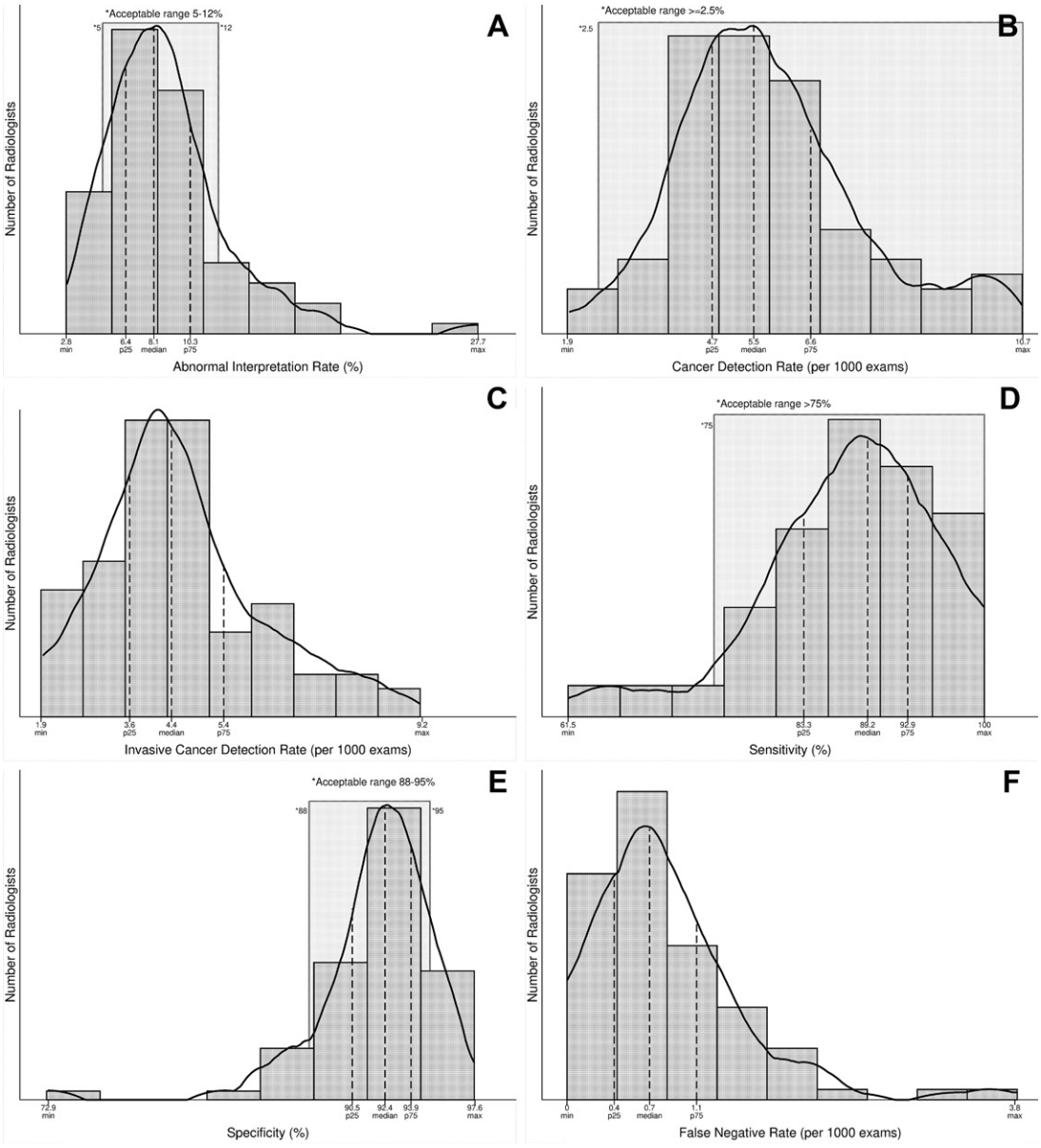

Figure 2: Radiologist digital breast tomosynthesis (DBT) screening performance and acceptable ranges. Histograms show the distribution of DBT screening performance benchmarks by radiologists including **(A)** abnormal interpretation rate (AIR), **(B)** cancer detection rate (CDR) (per 1000 examinations), **(C)** invasive CDR (per 1000 examinations), **(D)** sensitivity, **(E)** specificity, and **(F)** false-negative rate (per 1000 examinations). Lightly shaded region indicates radiologists within acceptable ranges established previously by expert opinion (if applicable). With DBT, 97.6%, 91.8%, 75.0%, and 74.0% of radiologists achieved the recommended acceptable performance ranges for CDR, sensitivity, AIR, and specificity, respectively. Only radiologists with at least 1000 DBT screening interpretations during the study period were included ($n = 84$). Sensitivity and specificity were restricted to radiologists with at least 10 DBT screening-detected cancers ($n = 73$). Max = maximum, min = minimum, p25 = 25th percentile, p75 = 75th percentile.

help guide ongoing research in the evaluation of performance of newer artificial intelligence algorithms for DBT screening (31). In conclusion, a large proportion of radiologists in U.S. community practice met acceptable performance ranges for CDR, AIR, and specificity with DBT. This report provides updated U.S. community–based DBT and DM performance data that will be highly important to all breast cancer screening stakeholders.

Acknowledgments: We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study. Information about the Breast Cancer Surveillance Consortium can be found at *http://www.bcsc-research.org/*.

Author contributions: Guarantors of integrity of entire study, **L.A., D.L.M., T.O., K.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **C.I.L., T.O., K.K.**; clinical studies, **D.L.M., T.O., K.K.**; statistical analysis, **L.A., D.L.M.**; and manuscript editing, all authors

Disclosures of conflicts of interest: C.I.L. Textbook royalty payments from McGraw Hill, Oxford University Press, and UpToDate; on the Grail DataSafety Monitoring Board; personal fees for journal editorial board work from the American College of Radiology. **L.A.** No relevant relationships. **D.L.M.** No relevant relationships. **T.O.** No

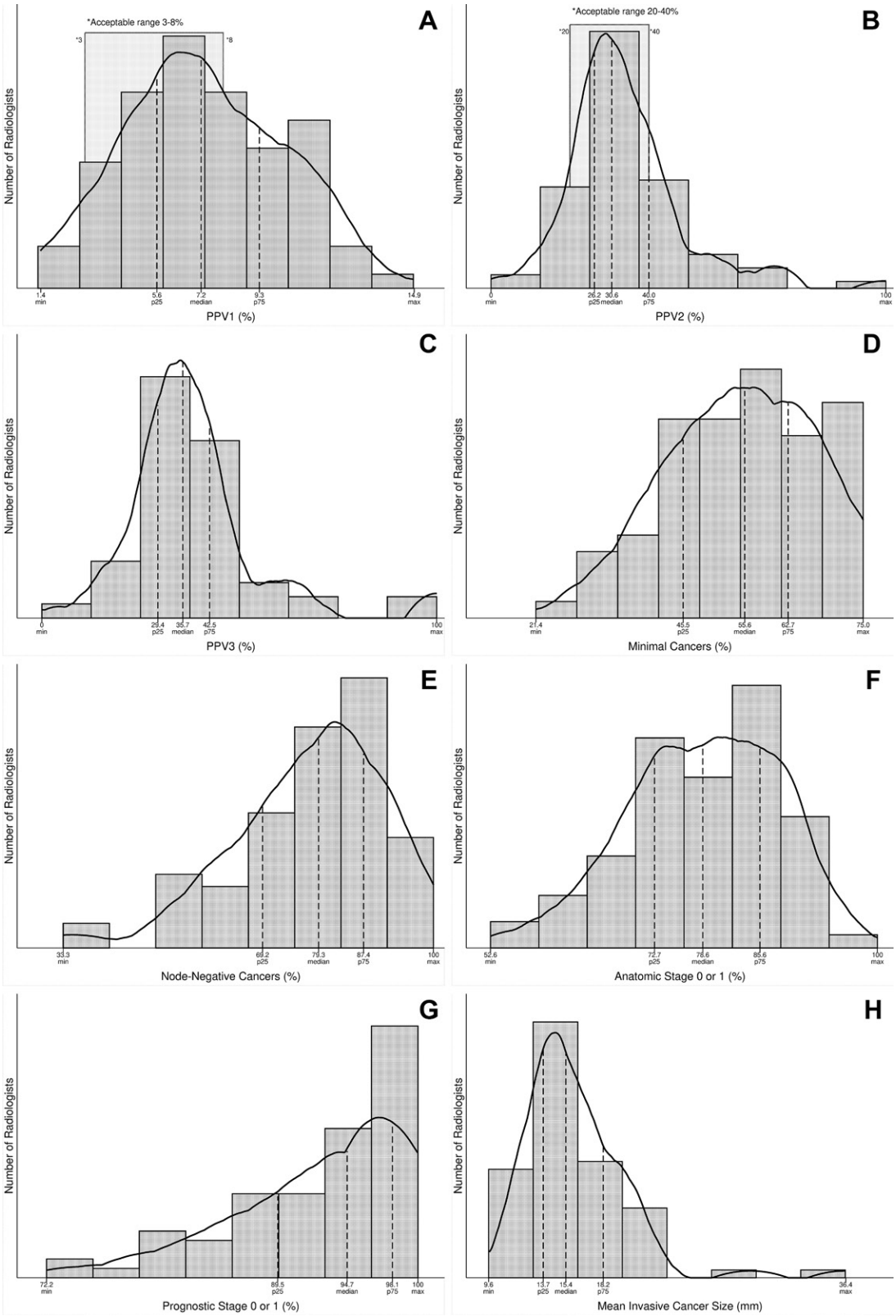

Figure 3: Additional radiologist digital breast tomosynthesis (DBT) screening performance measures and acceptable ranges. Histograms show the distribution of DBT screening performance benchmarks by radiologists including **(A)** PPV_1, **(B)** PPV_2, **(C)** PPV_3, **(D)** minimal cancers, **(E)** node-negative cancers, **(F)** anatomic stage 0 or 1, **(G)** prognostic stage 0 or 1, and **(H)** mean invasive cancer size. Lightly shaded region indicates radiologists within acceptable ranges established previously by expert opinion (if applicable). With DBT, 53.6% and 63.5% of radiologists assessed achieved the recommended acceptable performance ranges for PPV_1 and PPV_2, respectively. Only radiologists with at least 1000 DBT screening interpretations during the study period were included (n = 84). The percentage of minimal cancers and the percentage of node-negative cancers were restricted to radiologists with at least 10 DBT screening-detected cancers (n = 73). BI-RADS = Breast Imaging Reporting and Data System, Max = maximum, min = minimum, p25 = 25th percentile, p75 = 75th percentile, PPV_1 = examinations with abnormal interpretation (BI-RADS 0, 3, 4, or 5), PPV_2 = recommendation for tissue diagnosis (BI-RADS 4 or 5), PPV_3 = PPV of biopsy performed.

relevant relationships. **K.K.** No relevant relationships. **J.M.L.** No relevant relationships. **B.L.S.** No relevant relationships. **A.N.A.T.** No relevant relationships. **G.H.R.** No relevant relationships. **E.J.A.B.** No relevant relationships. **R.M.d.A.** No relevant relationships. **L.M.H.** No relevant relationships.

References

1. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380(9855):1778–1786.
2. Oeffinger KC, Fontham ET, Etzioni R, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA 2015;314(15):1599–1614. [Published correction appears in JAMA 2016;315(13):1406.]
3. Siu AL; U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med 2016;164(4):279–296. [Published correction appears in Ann Intern Med2016;164(6):448.]
4. Hofvind S, Skaane P, Elmore JG, Sebuødegård S, Hoff SR, Lee CI. Mammographic performance in a population-based screening program: before, during, and after the transition from screen-film to full-field digital mammography. Radiology 2014;272(1):52–62.
5. Lee CI, Lehman CD. Digital Breast Tomosynthesis and the Challenges of Implementing an Emerging Breast Cancer Screening Technology Into Clinical Practice. J Am Coll Radiol 2016;13(11S):R61–R66.
6. Lowry KP, Coley RY, Miglioretti DL, et al. Screening Performance of Digital Breast Tomosynthesis vs Digital Mammography in Community Practice by Patient Age, Screening Round, and Breast Density. JAMA Netw Open 2020;3(7):e2011792.
7. Lee CI, Zhu W, Onega T, et al. Comparative Access to and Use of Digital Breast Tomosynthesis Screening by Women's Race/Ethnicity and Socioeconomic Status. JAMA Netw Open 2021;4(2):e2037546.
8. U.S. Food and Drug Administration. MQSA National Statistics. https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics. Published 2022. Accessed September 2, 2022.
9. Monsees BS. The Mammography Quality Standards Act. An overview of the regulations and guidance. Radiol Clin North Am 2000;38(4):759–772.
10. Burnside ES, Sickles EA, Bassett LW, et al. The ACR BI-RADS experience: learning from history. J Am Coll Radiol 2009;6(12):851–860.
11. Rosenberg RD, Yankaskas BC, Abraham LA, et al. Performance benchmarks for screening mammography. Radiology 2006;241(1):55–66.
12. Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology 2017;283(1):49–58.
13. Lee JM, Ichikawa L, Valencia E, et al. Performance Benchmarks for Screening Breast MR Imaging in Community Practice. Radiology 2017;285(1):44–52.
14. 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.
15. Sickles EA, Miglioretti DL, Ballard-Barbash R, et al. Performance benchmarks for diagnostic mammography. Radiology 2005;235(3):775–790.
16. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med 2008;148(5):337–347.
17. American College of Radiology. American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, Va: American College of Radiology, 2013.
18. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. Biostatistics 2004;5(3):381–398.
19. Miglioretti DL, Heagerty PJ. Marginal modeling of nonnested multilevel data using standard software. Am J Epidemiol 2007;165(4):453–463.
20. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology 2010;255(2):354–361.
21. Kerlikowske K, Su YR, Sprague BL, et al. Association of Screening With Digital Breast Tomosynthesis vs Digital Mammography With Risk of Interval Invasive and Advanced Breast Cancer. JAMA 2022;327(22):2220–2230.
22. Miglioretti DL, Abraham L, Lee CI, et al. Digital Breast Tomosynthesis: Radiologist Learning Curve. Radiology 2019;291(1):34–42.
23. Sprague BL, Coley RY, Kerlikowske K, et al. Assessment of Radiologist Performance in Breast Cancer Screening Using Digital Breast Tomosynthesis vs Digital Mammography. JAMA Netw Open 2020;3(3):e201759.
24. Lee CS, Bhargavan-Chatfield M, Burnside ES, Nagy P, Sickles EA. The National Mammography Database: Preliminary Data. AJR Am J Roentgenol 2016;206(4):883–890.
25. Lee CS, Sengupta D, Bhargavan-Chatfield M, Sickles EA, Burnside ES, Zuley ML. Association of Patient Age With Outcomes of Current-Era, Large-Scale Screening Mammography: Analysis of Data From the National Mammography Database. JAMA Oncol 2017;3(8):1134–1136.
26. Lee CI, Zhu W, Onega TL, et al. The Effect of Digital Breast Tomosynthesis Adoption on Facility-Level Breast Cancer Screening Volume. AJR Am J Roentgenol 2018;211(5):957–963.
27. Lee CI, Bogart A, Hubbard RA, et al. Advanced Breast Imaging Availability by Screening Facility Characteristics. Acad Radiol 2015;22(7):846–852.
28. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of Digital Breast Tomosynthesis Compared With Digital Mammography: Outcomes Analysis From 3 Years of Breast Cancer Screening. JAMA Oncol 2016;2(6):737–743.
29. McDonald ES, McCarthy AM, Akhtar AL, Synnestvedt MB, Schnall M, Conant EF. Baseline Screening Mammography: Performance of Full-Field Digital Mammography Versus Digital Breast Tomosynthesis. AJR Am J Roentgenol 2015;205(5):1143–1148.
30. 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.
31. Anderson AW, Marinovich ML, Houssami N, et al. Independent External Validation of Artificial Intelligence Algorithms for Automated Interpretation of Screening Mammography: A Systematic Review. J Am Coll Radiol 2022;19(2 2 Pt A):259–273.