



# HHS Public Access

Author manuscript

*Cancer*. Author manuscript; available in PMC 2021 July 15.

Published in final edited form as:

*Cancer*. 2020 July 15; 126(14): 3289–3296. doi:10.1002/cncr.32939.

## New mammography screening performance metrics based on the entire screening episode

Brian L. Sprague, PhD<sup>1,2</sup>, Diana L. Miglioretti, PhD<sup>3</sup>, Christoph I. Lee, MD<sup>4</sup>, Hannah Perry, MD<sup>2</sup>, Anna A.N. Tosteson, ScD<sup>5</sup>, Karla Kerlikowske, MD<sup>6</sup>

<sup>1</sup>Department of Surgery, University of Vermont, Burlington, VT.

<sup>2</sup>Department of Radiology and University of Vermont Cancer Center, University of Vermont, Burlington, VT.

<sup>3</sup>Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, Davis, CA, USA and Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, WA.

<sup>4</sup>Department of Radiology, University of Washington School of Medicine, and the Hutchinson Institute for Cancer Outcomes Research, Seattle, WA.

<sup>5</sup>The Dartmouth Institute for Health Policy and Clinical Practice and Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH.

<sup>6</sup>Departments of Medicine and Epidemiology and Biostatistics, University of California, San Francisco, CA, USA and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, CA.

### Abstract

**Background**—Established mammography screening performance metrics use the initial screening mammography assessment because they were developed for radiologist performance auditing, yet these metrics are frequently used to inform health policy and screening decision-making. We developed new performance metrics based on the final assessment that consider the entire screening episode, including diagnostic work-up.

**Methods**—We used data from 2,512,577 screening episodes during 2005–2017 at 146 facilities in the United States participating in the Breast Cancer Surveillance Consortium. We compared screening performance metrics based on the final assessment of the screening episode to

---

CORRESPONDENCE Brian L. Sprague, PhD, 1 S. Prospect St, UHC Room 4425, University of Vermont, Burlington, VT 05401, (t) 802-656-4112; (f) 802-656-8826; brian.sprague@uvm.edu.

#### AUTHOR CONTRIBUTIONS

Brian L. Sprague: Conceptualization, data curation, funding acquisition, investigation, project administration, visualization, writing - original draft

Diana L. Miglioretti: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, writing - review and editing

Christoph I. Lee: Conceptualization, funding acquisition, project administration, writing - review and editing

Hannah Perry: Conceptualization, writing - review and editing

Anna A.N. Tosteson: Conceptualization, funding acquisition, project administration, writing - review and editing

Karla Kerlikowske: Conceptualization, funding acquisition, investigation, data curation, project administration, supervision, writing - review and editing

conventional metrics defined using the initial assessment. We also stratified results by breast density and breast cancer risk.

**Results**—The cancer detection rate was similar for final (4.1 per 1000; 95% CI: 3.8–4.3) vs. initial assessment (4.1 per 1000; 95% CI: 3.9–4.3). The interval cancer rate was 12% higher based on final (0.77 per 1000; 95% CI: 0.71–0.83) vs. initial assessment (0.69 per 1000; 95% CI: 0.64–0.74), resulting in a modest difference in sensitivity (84.1% [95% CI: 83.0–85.1] vs. 85.7% [95% CI: 84.8–86.6%], respectively). Absolute differences in interval cancer rate between final and initial assessment increased with breast density and breast cancer risk (e.g., difference of 0.29 per 1000 for women with extremely dense breasts and 5-year risk >2.49%).

**Conclusions**—Established screening performance metrics underestimate the interval cancer rate of a mammography screening episode, particularly for women with dense breasts or elevated breast cancer risk. Women, clinicians, policymakers, and researchers should use final assessment performance metrics to support informed screening decisions.

## PRECIS

Existing mammography screening performance metrics based on the initial assessment of the screening examination underestimate the interval cancer rate of a screening episode. Metrics based on the final assessment of the screening episode should be used to inform routine screening and supplemental screening guidelines.

## Keywords

breast neoplasms; mass screening; mammography; breast density; outcome assessment

## INTRODUCTION

Existing screening mammography performance metrics, including cancer detection rate, sensitivity, and specificity, are based on the well-established definitions developed by the American College of Radiology (ACR).<sup>1, 2</sup> These metrics were designed to evaluate radiologist performance in breast imaging interpretation, yet they are also widely used to inform women, healthcare providers, and policymakers regarding the potential benefits, harms, and limitations of mammography screening.<sup>3–5</sup>

A breast cancer screening episode starts with an initial screening mammogram, and if that screening exam is positive, is followed by diagnostic mammography or other imaging tests to either resolve any suspicious findings on the screening mammogram or to refer the woman for either short-interval follow-up imaging or a biopsy. The ACR recommends separate audits of screening and diagnostic mammography such that the performance of each type of breast imaging exam can be evaluated specifically.<sup>1</sup> The ACR screening mammography performance metrics consider only the assessment made by the radiologist interpreting the initial screening mammogram (i.e., the “initial assessment”). Subsequent diagnostic imaging work-up is interpreted separately and results in a final assessment. Audits of diagnostic imaging include these exams following abnormal screening, combined with diagnostic exams for women presenting with clinical signs and symptoms.<sup>6</sup> Thus, no existing metrics describe the performance of the entire screening episode.

This approach is useful for evaluating radiologist performance, particularly since the diagnostic examinations following an abnormal screen may be interpreted by a different radiologist than the one who interpreted the initial screening exam.<sup>7</sup> However, metrics for screening performance that are based only on the initial assessment are not entirely informative for women trying to understand the likely outcomes of mammography screening or for researchers evaluating or comparing woman-level outcomes with different screening modalities. One example is a woman who has an abnormal screening mammogram followed by diagnostic mammography that is interpreted as normal. A breast cancer that presents symptomatically in the ipsilateral breast within the following 12 months would be considered a true-positive (based on the positive initial assessment), despite the fact that the screening episode did not detect her cancer. The diagnostic mammogram would be considered a false-negative and the performance statistics for diagnostic mammography would reflect that. Yet, the performance statistics for diagnostic mammography are rarely considered when evaluating screening strategies.

Screening outcomes that inform clinical decision-making are impacted by the interpretative performance of the entire episode, which includes interpretation of both screening mammography and diagnostic imaging performed to work-up abnormal screens. The debate around breast density notification laws and supplemental screening for women with dense breasts<sup>8–10</sup> calls further attention to the need for screening metrics that accurately identify women at risk of poor screening outcomes, aside from evaluating radiologist performance. We sought to calculate mammography screening performance metrics based on the final assessment of an entire screening episode. We compared these new metrics to conventional screening performance metrics based on initial assessment and illustrated the impact of these definitions by determining screening performance and outcomes measures according to breast density and breast cancer risk.

## METHODS

### Study Design and Setting

We used observational data prospectively collected from the six active breast imaging registries of the Breast Cancer Surveillance Consortium (BCSC): the Carolina Mammography Registry, Kaiser Permanente Washington Registry, New Hampshire Mammography Network, Vermont Breast Cancer Surveillance System, San Francisco Mammography Registry, and Metropolitan Chicago Breast Cancer Registry.<sup>2, 11</sup> Each registry collects clinical data on women undergoing breast imaging exams at participating facilities within its catchment area. Each registry and the Statistical Coordinating Center received institutional review board approval for either active or passive consenting processes or a waiver of consent.

### Study Participants

Women aged 40–79 with a screening mammogram (digital mammography or digital breast tomosynthesis) during 2005–2017 were eligible for the study. To reflect regular participation in screening, all analyses were limited to women undergoing a screening mammogram within 30 months after a prior screening mammogram. Thus, women undergoing their first

screening mammogram and women undergoing a screening mammogram more than 30 months since their most recent screen were excluded. We also excluded screening exams among women with a personal history of breast cancer, mastectomy, or breast augmentation (among whom screening performance statistics are known to differ greatly) and exams occurring within 9 months of a prior mammogram or breast ultrasound examination (to avoid inclusion of misclassified screening exams).

### Data Collection

BCSC registries capture imaging modality, exam indication, breast density, and assessment data from participating radiology facilities using standard nomenclature defined by Breast Imaging Reporting and Data System (BI-RADS).<sup>1</sup> Breast density was recorded by the interpreting radiologist using the BI-RADS categories (almost entirely fatty, scattered areas of fibroglandular density, heterogeneously dense, and extremely dense).<sup>1</sup> Demographic, risk factor, and medical history information for women was obtained via a questionnaire completed at each mammogram or by extraction from the electronic medical record. Data on breast cancer diagnoses (invasive cancer or DCIS) diagnosed within 1 year after each screening mammogram and prior to the next screening mammogram were obtained by linking to pathology databases; regional Surveillance, Epidemiology, and End Results programs; and state tumor registries.

### Key Measures and Definitions

Radiologists interpreting screening mammograms in clinical practice assign one of six assessment categories: 0) Incomplete: need additional imaging evaluation; 1) Negative; 2) Benign; 3) Probably benign; 4) Suspicious; 5) Highly suggestive of malignancy. Based on ACR guidelines<sup>1</sup>, screening mammograms should receive an initial assessment of 0, 1, or 2, though in practice radiologists occasionally record an initial assessment of 3, 4, or 5.<sup>12</sup> Screening exams with an initial assessment of 0 should be followed by diagnostic imaging (mammography or ultrasound), which after interpretation should be assigned a final assessment of 1, 2, 3, 4, or 5. Recommendations for follow-up are linked to the assessment as follows: 1,2) return to routine screening; 3) short-interval follow-up imaging at 6 months; 4,5) recommendation for tissue biopsy.

In our analyses a screening episode was defined by identifying screening mammography exams and all subsequent diagnostic imaging occurring within the 90 days following screening exams with an initial BI-RADS assessment of 0 (needs additional evaluation) and prior to a breast biopsy. Screening performance metrics were defined separately based on a) the initial screening assessment as per established American College of Radiology BI-RADS definitions<sup>1</sup>; and b) the final assessment after diagnostic workup.

For metrics based on initial assessment, we used the standard BI-RADS definitions of a positive screen (initial BI-RADS assessment category 0, 3, 4, or 5), cancer detection rate, interval cancer (false negative) rate, sensitivity, and specificity.<sup>1</sup> For metrics based on final assessment, we defined a positive screening episode as final BI-RADS assessment categories 3, 4, or 5. While the ACR does not consider a category 3 assessment to be positive in diagnostic mammographic audits,<sup>1</sup> we observed that the majority of cancers diagnosed after

category 3 final assessments were detected within 7–8 months after the screening exam (eFigure S1, Supplemental Material), consistent with the premise that the category 3 assessment led to the cancer diagnosis via further work-up at 6 months short-interval follow-up. Sensitivity analyses examined the impact of classifying category 3 final assessments as negative instead of positive. Cancer status for all metrics was determined based on the standard 365 day follow-up stipulated by BI-RADS.<sup>1</sup>

Additional screening outcome measures defined based on initial vs. final assessment included rates of early stage (American Joint Committee on Cancer anatomic stage I and IIa) screen-detected invasive breast cancers and interval- and screen-detected advanced stage (American Joint Committee on Cancer anatomic stage IIb or higher) breast cancers. We also calculated the false-positive rate of a) additional imaging recommendation on the initial screen; b) biopsy recommendation at the conclusion of screening episode; and c) short interval follow-up recommendation at the conclusion of the screening episode.

### Statistical Analyses

The screening episode was the unit of analysis for all statistics. We calculated screening performance metrics measures separately based on the screening episode's initial assessment and final assessment. We estimated 95% confidence intervals using generalized estimating equations with a working independence correlation structure to account for correlation among mammograms from the same woman, radiologist, or facility.<sup>13, 14</sup> Selected screening performance measures were calculated for subgroups of women defined by breast density and estimated 5-year BCSC breast cancer risk. We estimated each woman's 5-year invasive breast cancer risk using the BCSC version 2.0 risk model,<sup>15</sup> and categorized as low (0 to <1.00%), average (1.00–1.66%), intermediate (1.67%–2.49%), high (2.5%–3.99%) and very high (>3.99%).<sup>5, 16</sup> SAS statistical software version 9.4 (SAS Institute Inc.) was used for all analyses.

## RESULTS

Over 2.5 million mammography screening episodes were identified among 791,347 individual women, with exams interpreted by 705 radiologists at 146 facilities (Table 1). A total of 12,131 cancers occurred during the 1 year period following a screening mammogram. A higher proportion of women diagnosed with breast cancer were older, postmenopausal, had a first degree family history of breast cancer, had heterogeneously or extremely dense breasts, and had an elevated BCSC 5-year risk compared to women not diagnosed with breast cancer.

Of the 2,512,577 screening episodes, 8.6% had a positive (abnormal) initial assessment (Table 2). Among the initial positive screens, 64.8% had a negative final assessment, 19.5% had a category 3 short-interval follow-up final assessment, and 15.7% had a category 4/5 biopsy recommendation final assessment. Among the 12,131 cancers, 85.7% were diagnosed among women after a positive initial assessment (Table 2). 1.6% of cancers occurred after a positive initial assessment that was resolved to a category 1 or 2 final assessment, and 4.5% of cancers occurred after a positive initial assessment that was resolved to a category 3 final assessment.

The cancer detection rate was comparable based on the final vs. initial assessment (Table 3). The rate of interval cancers was 12% higher based on the final assessment (0.77 per 1000; 95% CI: 0.71–0.83 per 1000) compared to the initial assessment (0.69 per 1000; 95% CI: 0.61–0.79 per 1000). Sensitivity was higher based on initial assessment (85.7%; 95% CI: 84.8–86.6%) compared to final assessment (84.1%; 95% CI: 83.0–85.1), whereas specificity was lower based on initial assessment (91.8%; 95% CI: 91.0–92.4%) compared to final assessment (97.4%; 95% CI: 97.0–97.6%).

There were no large absolute differences in rates of screen-detected early stage breast cancer, or advanced stage interval- and screen-detected breast cancers according to final vs. initial assessment (Table 3). The false-positive rate based on initial assessment was 82.1 per 1000, whereas the false-positive rate of short interval follow-up (category 3) and biopsy recommendation (categories 4, 5) were 16.5 per 1000 and 9.8 per 1000, respectively.

The absolute differences in screening performance metrics based on final vs. initial assessment were greatest for women with dense breasts (Table 4). The interval cancer rate for women with extremely dense breasts was 1.57 per 1000 (95% CI: 1.38–1.80 per 1000) based on final assessment, compared to 1.42 (95% CI: 1.23–1.63 per 1000) based on initial assessment. The absolute difference in interval cancer rates according to final vs. initial assessment also increased with increasing BCSC 5-year risk (Figure 1). For example, among women with almost entirely fat breast density and low breast cancer risk there was no difference in interval cancer rate based on final vs. initial assessment, whereas among women with extremely dense breasts and risk >2.49% the interval cancer rate was higher by 0.29 per 1000 based on final vs. initial assessment.

In sensitivity analyses, we found that classification of category 3 final assessments as negative (instead of positive) resulted in a slightly lower cancer detection rate; increases in the interval cancer rate and specificity; and a moderate decrease in sensitivity (eTable S1, Supplemental Material).

## DISCUSSION

Determination of screening performance metrics based on the final assessment of the screening episode rather than the initial assessment of the screening exam results in re-classification of approximately 1.9% of screen-detected cancers. This has a modest influence on the cancer detection rate and screening sensitivity, but corresponds to a 12% increase in the interval cancer rate. This phenomenon is largest among women with dense breasts or high risk levels. These differences have consequences for women, healthcare providers, and policymakers considering primary and supplemental breast cancer screening strategies.

Based on the initial assessment, it appears that the interval cancer rate on mammography screening is 0.69 per 1000. However, this increased to 0.77 per 1000 when using the final assessment, which more accurately reflects the clinical outcome of the screening episode. The difference reflects cancers that had a positive initial screening mammogram but were resolved to a negative final assessment upon additional imaging. These results indicate that the mammography screening process has a higher failure rate than previously appreciated

based on established screening mammography performance metrics that are based on the initial assessment alone.

Our findings are particularly relevant to women with dense breasts, who in most US states are now informed of the limitations of mammography and advised to discuss supplemental screening options with their healthcare providers.<sup>17</sup> While a federal US law is pending,<sup>18</sup> there remains widespread debate and uncertainty regarding the appropriateness of supplemental screening for women with dense breasts.<sup>8, 10, 19</sup> Supplemental screening would likely benefit women who have high rates of interval cancers after screening mammography. Two recent studies have shown that supplemental screening with magnetic resonance imaging (MRI) of women with extremely dense breasts yields a substantially higher rate of detected cancers,<sup>20, 21</sup> with one of the studies also demonstrating a decrease in the interval cancer rate.<sup>20</sup> We have previously demonstrated that the risk of interval and advanced stage cancers among women with dense breasts varies considerably according to their BCSC 5-year risk.<sup>5, 16</sup> Our current results demonstrate that interval cancer rates based on the final assessment of the screening episode are higher than previously recognized. Furthermore, the degree of underestimation is largest for women with dense breasts or high BCSC 5-year risk. Healthcare providers and policy makers should consider these new estimates based on the complete screening episode when considering supplemental screening recommendations. Researchers should use these new definitions based on the final assessment of the screening episode in future studies designed to identify women at high risk of mammography screening failures. We recently used this approach in a study examining risk of advanced-stage breast cancer as a screening outcome.<sup>16</sup> Studies comparing screening outcomes according to modality (e.g., digital breast tomosynthesis vs. digital mammography) or evaluating the benefits of supplemental screening modalities should also consider this approach.

The specificity of screening mammography is higher when based on the final assessment vs. the initial assessment. This occurs because the majority of positive initial assessments are resolved on diagnostic imaging with a negative final assessment, and in most instances no cancer is diagnosed within the following year. The false-positive rate based on the final assessment is only 1.0%, compared to 9.1% based on the initial assessment. Notably, these false-positives have different consequences, though both are important. A false-positive on the initial screen in most instances involves only additional imaging, whereas a false-positive final positive assessment involves a recommendation for short interval follow-up imaging and/or biopsy. For this reason, recent policy-level evaluations of mammography screening have considered additional imaging and biopsy recommendations separately.<sup>4</sup>

The BI-RADS manual stipulates that category 3 final assessments should be classified as negative when evaluating the performance of diagnostic mammography.<sup>1</sup> For our evaluation of screening episodes, we classified category 3 final assessments as positive because for most cancers diagnosed after a category 3 final assessment it appeared that the screening episode led to an asymptomatic cancer diagnosis. A category 3 final assessment typically comes with a recommendation for short interval follow-up at six months.<sup>1</sup> In some instances, biopsy may be performed immediately based on patient preference. We observed that the majority of cancers occurring after a category 3 final assessment were diagnosed

within 7–8 months following the screening exam. This suggests diagnosis via short interval follow-up as a consequence of the category 3 assessment.

The BCSC includes a large, geographically diverse sample of academic and community practice breast imaging facilities in the US that collectively serve a racially diverse population of women. Our results are representative of the wide range of diagnostic work-up pathways used in clinical practice in the US.<sup>22</sup> Limitations of the study include the inability to definitively determine whether a cancer diagnosis was directly attributable to the screening episode. This limitation is common to existing screening performance metrics used for radiologist and facility performance audits<sup>1</sup> and research studies in which individual chart review is not feasible.<sup>2, 5</sup> In our study, 97.3% of the cancers occurring within 1 year of a category 4 or 5 final assessment were diagnosed within 90 days of the screening exam, suggesting that the impact of this limitation is small.

In summary, conventional mammography screening performance metrics based on the initial assessment of the screening examination underestimate the interval cancer rate of a screening episode, particularly for women with dense breasts or high breast cancer risk level. Women, clinicians, policymakers, and researchers should consider screening outcome measures based on the final assessment in order to support informed decisions about routine screening and the need for supplemental breast cancer screening.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS

We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study. You can learn more about the BCSC at: <http://www.bsc-research.org/>.

### FUNDING

This work was supported by the National Cancer Institute (grant P01CA154292). Data collection for this research was additionally supported by grant U54CA163303 from the National Cancer Institute, grant R01 HS018366-01A1 from the Agency for Healthcare Research and Quality, and award PCS-1504-30370 from the Patient-Centered Outcomes Research Institute (PCORI). The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: <http://www.bsc-research.org/work/acknowledgement.html>. The statements presented in this work are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health, or PCORI, its Board of Governors or Methodology Committee.

### CONFLICT OF INTEREST STATEMENT

Diana Miglioretti received an honorarium and travel support to serve on a scientific advisory board for Hologic in May and July 2017. Christoph Lee was a co-investigator of a grant from GE Healthcare that was paid to his institution (University of Washington School of Medicine). Karla Kerlikowske received travel funds from the Breast Cancer Research Foundation and the American Cancer Society to attend the 2017 International Working Group on Risk Assessment and Strategies for Breast Cancer Screening and Prevention. All other authors report no potential conflicts of interest.



## REFERENCES

1. American College of Radiology. ACR BI-RADS® - Mammography 5th Edition. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology, 2013.
2. 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: 49–58. [PubMed: 27918707]
3. Siu AL. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med*. 2016;164: 279–296. [PubMed: 26757170]
4. Nelson HD, O’Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data. *Ann Intern Med*. 2016;164: 226–235. [PubMed: 26756902]
5. Kerlikowske K, Zhu W, Tosteson AN, et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study. *Ann Intern Med*. 2015;162: 673–681. [PubMed: 25984843]
6. Sprague BL, Arao RF, Miglioretti DL, et al. National Performance Benchmarks for Modern Diagnostic Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology*. 2017;283: 59–69. [PubMed: 28244803]
7. Buist DS, Anderson ML, Smith RA, et al. Effect of radiologists’ diagnostic work-up volume on interpretive performance. *Radiology*. 2014;273: 351–364. [PubMed: 24960110]
8. Smetana GW, Elmore JG, Lee CI, Burns RB. Should This Woman With Dense Breasts Receive Supplemental Breast Cancer Screening?: Grand Rounds Discussion From Beth Israel Deaconess Medical Center. *Ann Intern Med*. 2018;169: 474–484. [PubMed: 30285208]
9. Haas JS, Kaplan CP. The Divide Between Breast Density Notification Laws and Evidence-Based Guidelines for Breast Cancer Screening: Legislating Practice. *JAMA Intern Med*. 2015;175: 1439–1440. [PubMed: 26147642]
10. Melnikow J, Fenton JJ, Whitlock EP, et al. Supplemental Screening for Breast Cancer in Women With Dense Breasts: A Systematic Review for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2016;164: 268–278. [PubMed: 26757021]
11. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. *AJR. American journal of roentgenology*. 1997;169: 1001–1008. [PubMed: 9308451]
12. Taplin SH, Ichikawa LE, Kerlikowske K, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. *Radiology*. 2002;222: 529–535. [PubMed: 11818624]
13. Miglioretti DL, Heagerty PJ. Marginal modeling of nonnested multilevel data using standard software. *Am J Epidemiol*. 2007;165: 453–463. [PubMed: 17121864]
14. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics*. 2004;5: 381–398. [PubMed: 15208201]
15. Tice JA, Miglioretti DL, Li CS, Vachon CM, Gard CC, Kerlikowske K. Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer. *J Clin Oncol*. 2015;33: 3137–3143. [PubMed: 26282663]
16. Kerlikowske K, Sprague BL, Tosteson ANA, et al. Identifying Women Having Routine Screening at High Risk of Advanced Breast Cancer for Discussion of Supplemental Imaging *JAMA Intern Med*. (in press).
17. Are You Dense Inc. State Density Reporting Efforts - because your life matters®. Available from URL: <https://www.areyoudenseadvocacy.org/dense> [accessed April 10, 2019, 2019].
18. US Food and Drug Administration. Mammography Quality Standards Act: A Proposed Rule by the Food and Drug Administration on 03/28/2019. Available from URL: <https://federalregister.gov/d/2019-05803> [accessed April 5, 2019].
19. Keating NL, Pace LE. New Federal Requirements to Inform Patients About Breast Density: Will They Help Patients? *JAMA*. 2019;321: 2275–2276. [PubMed: 31070674]
20. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. *N Engl J Med*. 2019;381: 2091–2102. [PubMed: 31774954]

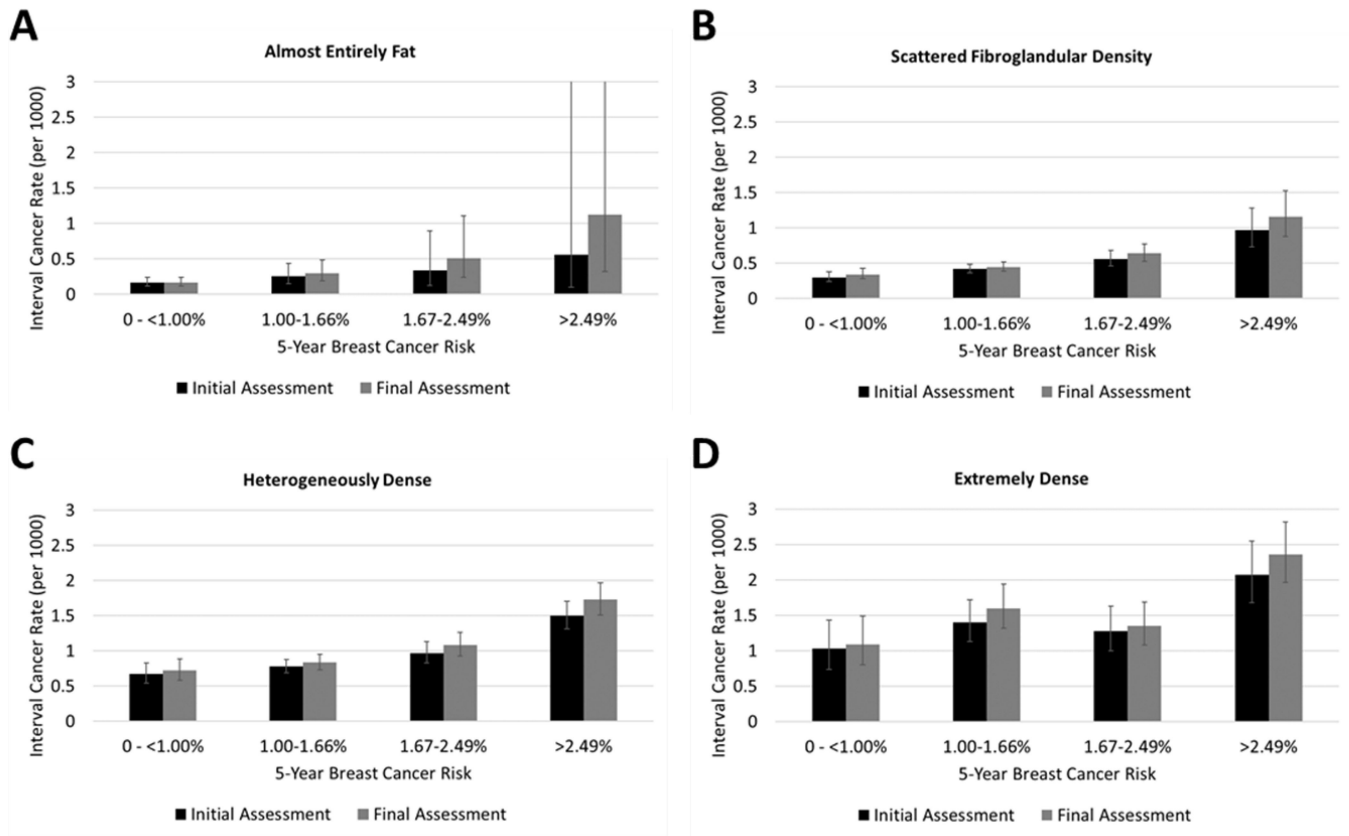
21. Comstock CE, Gatsonis C, Newstead GM, et al. Comparison of Abbreviated Breast MRI vs Digital Breast Tomosynthesis for Breast Cancer Detection Among Women With Dense Breasts Undergoing Screening. *JAMA*. 2020;323: 746–756. [PubMed: 32096852]
22. Hubbard RA, Zhu W, Horblyuk R, et al. Diagnostic imaging and biopsy pathways following abnormal screen-film and digital screening mammography. *Breast Cancer Res Treat*. 2013;138: 879–887. [PubMed: 23471650]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 1.** Interval cancer rate based on initial vs. final assessment, according to breast density and five-year breast cancer risk, among women with a recent screening mammogram in the Breast Cancer Surveillance Consortium, 2005–2017. Error bars depict 95% confidence intervals.

**Table 1.**

Exam-level characteristics of study participants, overall and by breast cancer status within 1 year follow-up.

	All Screening Examinations		No Breast Cancer		Breast Cancer	
	N	%	N	%	N	%
Screening mammograms <sup>†</sup>	2,512,577	100%	2,500,446	100%	12,131	100%
Age, years						
40–49	623,156	25%	621,093	25%	2,063	17%
50–59	889,719	35%	885,928	35%	3,791	31%
60–69	758,883	30%	754,349	30%	4,534	37%
70–79	240,819	10%	239,076	10%	1,743	14%
Race/ethnicity						
White, non-Hispanic	1,685,099	67%	1,676,540	67%	8,559	71%
Black, non-Hispanic	244,665	10%	243,498	10%	1,167	10%
Asian/Native Hawaiian/Pacific Islander	299,151	12%	297,759	12%	1,392	11%
Hispanic	134,438	5%	133,990	5%	448	4%
Other/Mixed/Unknown	149,224	6%	148,659	6%	565	5%
Menopausal status						
Pre- or peri-menopausal	651,230	29%	648,736	29%	2,494	23%
Postmenopausal	1,440,724	65%	1,432,642	65%	8,082	73%
Surgical menopausal	118,989	5%	118,490	5%	499	5%
Unknown	301,634	(12%)	300,578	(12%)	1,056	(9%)
First degree family history of breast cancer						
No	2,009,458	83%	2,000,707	83%	8,751	75%
Yes	417,981	17%	415,039	17%	2,942	25%
Unknown	85,138	(3%)	84,700	(3%)	438	(4%)
BI-RADS breast density <sup>‡</sup>						
Almost entirely fat	234,897	10%	234,233	10%	664	6%
Scattered fibroglandular density	987,913	43%	983,390	43%	4,523	41%
Heterogeneously dense	910,134	39%	905,306	39%	4,828	44%
Extremely dense	186,699	8%	185,731	8%	968	9%
Unknown	192,934	(8%)	191,786	(8%)	1,148	(9%)
BCSC 5-year risk <sup>//</sup>						
0 – <1.00%	646,582	28%	644,926	28%	1,656	15%
1.00–1.66%	931,062	40%	926,976	40%	4,086	37%
1.67–2.49%	499,586	22%	496,521	22%	3,065	28%
2.50–3.99%	211,302	9%	209,512	9%	1,790	16%
4.00%	31,111	1%	30,725	1%	386	4%
Unknown	192,934	(8%)	191,786	(8%)	1,148	(9%)

\* Breast cancer cases within 12 months of screening mammography

<sup>†</sup> Subsequent screening examinations after first screening mammography and within 30 months of prior mammography<sup>‡</sup> Breast Imaging Reporting and Data System (BI-RADS)

//Breast Cancer Surveillance Consortium (BCSC) 5-year risk calculated using age, race, first degree family history of breast cancer, history of breast biopsy, BI-RADS density.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2.**

Initial vs. final assessment among women with a recent screening examination in the Breast Cancer Surveillance Consortium, 2005–2017.

Initial Assessment	Final Assessment			Total
	Negative (BI-RADS 1,2)	Short-interval follow-up (BI-RADS 3)	Biopsy recommendation (BI-RADS 4,5)	
<i>All screens</i>				
Negative (BI-RADS 1,2)	2,295,935 (91.4%)	NA	NA	2,295,935 (91.4%)
Positive (BI-RADS 0,3,4,5)	140,438 (5.6%)	41,889 (1.7%)	34,315 (1.4%)	216,642 (8.6%)
Total	2,436,373 (97.0%)	41889 (1.7%)	34,315 (1.4%)	2,512,577 (100.0%)
<i>Screens followed by a cancer diagnosis</i>				
Negative (BI-RADS 1,2)	1,736 (14.3%)	NA	NA	1,736 (14.3%)
Positive (BI-RADS 0,3,4,5)	195 (1.6%)	551 (4.5%)	9,649 (79.5%)	10,395 (85.7%)
Total	1,931 (15.9%)	551 (4.5%)	9,649 (79.5%)	12,131 (100.0%)

Screening mammography performance metrics based on initial vs. final assessment, among women with a recent screening mammogram in the Breast Cancer Surveillance Consortium, 2005–2017.

**Table 3.**

Screening Performance Metric	Initial Assessment		Final Assessment	
	Estimate	95% CI	Estimate	95% CI
<i>Standard performance measures</i>				
Cancer detection rate, per 1000	4.1	3.9, 4.3	4.1	3.8, 4.3
Interval cancer rate, per 1000	0.69	0.64, 0.74	0.77	0.71, 0.83
Sensitivity, %	85.7	84.8, 86.6	84.1	83.0, 85.1
Specificity, %	91.8	91.0, 92.4	97.4	97.0, 97.6
<i>Screening benefit</i>				
Stage I or IIa screen-detected invasive cancers, per 1000	2.7	2.5, 2.8	2.6	2.5, 2.8
<i>Screening false-alarms</i>				
False-positive recall for additional imaging, per 1000	82.1	75.3, 89.4	NA	NA
False-positive short interval follow-up recommendation, per 1000	NA	NA	16.5	13.9, 19.5
False-positive biopsy recommendation, per 1000	NA	NA	9.8	8.5, 11.3
<i>Screening failures</i>				
Stage IIb or higher interval invasive cancers, per 1000	0.16	0.15, 0.18	0.18	0.16, 0.20
Stage IIb or higher screen-detected invasive cancers, per 1000	0.29	0.26, 0.33	0.27	0.24, 0.31

CI, confidence interval; NA, not applicable.

Breast cancer screening performance measures based on initial vs. final assessment, according to breast density, among women with a recent screening mammogram in the Breast Cancer Surveillance Consortium, 2005–2017.

**Table 4.**

Screening Performance Metric	BI-RADS Breast Density			
	Almost entirely fat	Scattered fibroglandular density	Heterogeneously dense	Extremely dense
<i>Based on initial assessment</i>				
Cancer detection rate, per 1000	2.6 (2.3, 2.9)	4.1 (3.9, 4.4)	4.4 (4.1, 4.7)	3.8 (3.4, 4.2)
Interval cancer rate, per 1000	0.21 (0.14, 0.30)	0.44 (0.39, 0.49)	0.93 (0.85, 1.03)	1.42 (1.23, 1.63)
Sensitivity, %	92.6 (89.7, 94.8)	90.4 (89.2, 91.4)	82.4 (80.8, 84.0)	72.6 (69.0, 76.0)
Specificity, %	95.5 (95.0, 95.9)	92.4 (91.6, 93.1)	90.1 (89.1, 91.0)	91.0 (90.0, 91.9)
<i>Based on final assessment</i>				
Cancer detection rate, per 1000	2.6 (2.3, 2.9)	4.1 (3.8, 4.3)	4.3 (4.0, 4.6)	3.6 (3.3, 4.0)
Interval cancer rate, per 1000	0.23 (0.17, 0.33)	0.49 (0.44, 0.55)	1.03 (0.93, 1.14)	1.57 (1.38, 1.80)
Sensitivity, %	91.7 (88.7, 94.0)	89.3 (87.9, 90.5)	80.5 (78.6, 82.3)	69.6 (66.0, 73.1)
Specificity, %	98.4 (98.1, 98.6)	97.6 (97.3, 97.9)	96.9 (96.5, 97.2)	96.9 (96.5, 97.3)

BI-RADS, Breast Imaging Reporting and Data System; CI, confidence interval