





# Comparative Effectiveness of Digital Breast Tomosynthesis for Breast Cancer Screening Among Women 40-64 Years Old

Ilana B. Richman , MD, MHS,<sup>1,2,\*</sup> Jessica B. Long, MPH,<sup>1,2</sup> Jessica R. Hoag, PhD,<sup>3</sup> Akhil Upneja , BS,<sup>4</sup> Regina Hooley, MD,<sup>5</sup> Xiao Xu , PhD,<sup>2,6</sup> Natalia Kunst , MSc,<sup>2,7</sup> Jenerius A. Aminawung, MD, MPH,<sup>1,2</sup> Kelly A. Kyanko, MD, MHS,<sup>8</sup> Susan H. Busch, PhD,<sup>2,9</sup> Cary P. Gross, MD<sup>1,2</sup>

<sup>1</sup>Department of Medicine, Yale School of Medicine, New Haven, CT, USA; <sup>2</sup>Cancer Outcomes, Public Policy, and Effectiveness Research (COPPER) Center, Yale School of Medicine, New Haven, CT, USA; <sup>3</sup>CATO SMS, Cary, NC, USA; <sup>4</sup>Yale School of Medicine, New Haven, CT, USA; <sup>5</sup>Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA; <sup>6</sup>Department of Obstetrics, Gynecology and Reproductive Sciences, Yale School of Medicine, New Haven, CT, USA; <sup>7</sup>Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA; <sup>8</sup>Department of Population Health, New York University School of Medicine, New York, NY, USA and <sup>9</sup>Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA

\*Correspondence to: Ilana Richman, MD, Department of Medicine, Yale School of Medicine, 367 Cedar St, Harkness A, Room 301a, New Haven, CT 06510, USA (e-mail: ilana.richman@yale.edu).

## Abstract

**Background:** Digital breast tomosynthesis (DBT) may have a higher cancer detection rate and lower recall compared with 2-dimensional (2D) mammography for breast cancer screening. The goal of this study was to evaluate screening outcomes with DBT in a real-world cohort and to characterize the population health impact of DBT as it is widely adopted. **Methods:** This observational study evaluated breast cancer screening outcomes among women screened with 2D mammography vs DBT. We used deidentified administrative data from a large private health insurer and included women aged 40-64 years screened between January 2015 and December 2017. Outcomes included recall, biopsy, and incident cancers detected. We used 2 complementary techniques: a patient-level analysis using multivariable logistic regression and an area-level analysis evaluating the relationship between population-level adoption of DBT use and outcomes. All statistical tests were 2-sided. **Results:** Our sample included 7 602 869 mammograms in 4 580 698 women, 27.5% of whom received DBT. DBT was associated with modestly lower recall compared with 2D mammography (113.6 recalls per 1000 screens, 99% confidence interval [CI] = 113.0 to 114.2 vs 115.4, 99% CI = 115.0 to 115.8,  $P < .001$ ), although younger women aged 40-44 years had a larger reduction in recall (153 recalls per 1000 screens, 99% CI = 151 to 155 vs 164 recalls per 1000 screens, 99% CI = 163 to 166,  $P < .001$ ). DBT was associated with higher biopsy rates than 2D mammography (19.6 biopsies per 1000 screens, 99% CI = 19.3 to 19.8 vs 15.2, 99% CI = 15.1 to 15.4,  $P < .001$ ) and a higher cancer detection rate (4.9 incident cancers per 1000 screens, 99% CI = 4.7 to 5.0 vs 3.8, 99% CI = 3.7 to 3.9,  $P < .001$ ). Point estimates from the area-level analysis generally supported these findings. **Conclusions:** In a large population of privately insured women, DBT was associated with a slightly lower recall rate than 2D mammography and a higher cancer detection rate. Whether this increased cancer detection improves clinical outcomes remains unknown.

Two-dimensional (2D) mammography has been the standard of care for breast cancer screening for nearly 4 decades, but it has important limitations that result in imperfect sensitivity and specificity (1). Digital breast tomosynthesis (DBT) is a newer breast imaging modality that uses multiple x-ray exposures to create a quasi 3-dimensional image of breast tissue (2). Initial studies of DBT have generally found that women screened with DBT are less likely to need follow-up imaging for an abnormal finding (known as recall) compared with women screened with 2D mammography (3-5). Several studies have also reported a

higher cancer detection rate among women screened with DBT (4-7). This early experience with tomosynthesis has generated enthusiasm, resulting in growing use of DBT in the United States. By the end of 2017, more than 40% of privately insured women were screened with tomosynthesis, and by 2021 77% of mammographic facilities reported ownership of a tomosynthesis unit (8,9).

Although early studies suggest that DBT has some advantages compared with 2D mammography, the current body of evidence has important limitations. First, many studies reflect

Received: June 24, 2020; Revised: November 27, 2020; Accepted: March 30, 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

the experience of high-volume centers, early adopters, or limited geographies (4,5,7,10). Results from these selected settings may not be broadly applicable to diverse patients and practices and do not capture the potential population health impact of changes in screening technology use. Second, published studies have reported discrepant findings regarding DBT performance. Among US studies, a recent meta-analysis of tomosynthesis studies demonstrated a wide range in the reduction in recall with statistical heterogeneity among studies (11). European studies, including some randomized trials, generally have reported lower recall rates overall and have not consistently shown reductions in recall with DBT (11-13). Last, studies have reported conflicting results with respect to cancer detection, with many reporting no increase in cancer detection with DBT, albeit with limited power to detect small differences (11,14). Thus, important questions remain about the effectiveness of tomosynthesis for breast cancer screening, particularly in real-world clinical practice.

Our study aims to address these gaps in knowledge in 2 ways. Our primary approach was to evaluate the comparative effectiveness of DBT and 2D mammography using cross-sectional data from a large real-world cohort of privately insured women undergoing screening. Specifically, we compared proximal screening outcomes, including recall and cancer detection, among privately insured women undergoing DBT compared with 2D mammography. Our second approach was to evaluate population-level changes in recall and cancer detection over time with the introduction of DBT. This approach both mitigates confounding by using a longitudinal design and evaluates the population-level impact of the introduction of DBT on screening outcomes.

## Methods

### Data Source and Study Sample

We used data from the Blue Cross Blue Shield Axis, a large database of deidentified commercial insurance claims, accessed through a secure data environment. We included women aged 40-64 years who had at least 1 screening mammogram between January 1, 2015, and December 31, 2017, and who were continuously enrolled during the 2 years before the index mammogram and the year following the index mammogram. We excluded women with a history of breast cancer, as defined by a breast cancer-related diagnosis code on any claim in the 2 years before mammography. We also excluded women who had claims indicating a genetic cancer syndrome or prophylactic mastectomy. Additional details about sample selection are provided in the [Supplementary Methods](#) and [Supplementary Table 1](#) (available online).

### Exposure

We identified screening 2D mammograms using a validated algorithm that distinguishes screening from diagnostic mammograms ([Supplementary Methods](#); [Supplementary Table 1](#), available online) (15). Because DBT must always be billed with a 2D mammogram, we identified screening DBT based on the presence of a claim for DBT on the same day as a screening 2D mammogram, as defined by a previously validated algorithm (16).

## Outcomes

We evaluated 2 types of outcomes following the index screening mammogram: subsequent workup and incident breast cancers. For subsequent workup, we measured recall, magnetic resonance imaging (MRI), and biopsies. We defined “recall” as any diagnostic mammogram (either 2D or DBT) or unilateral or limited ultrasound in the 4 months following the index screening mammogram. We separately evaluated ultrasounds that were likely screening exams (bilateral, whole breast) and did not consider them to constitute recall. We also evaluated breast MRI use, but did not consider MRI to indicate recall because MRI is typically either used for high-risk screening or for workup of a newly diagnosed malignancy. Lastly, we measured biopsy use, identified by current procedural terminology (CPT) code, following screening mammography.

We identified incident breast cancers in the 4 months following mammography using a validated algorithm designed to identify screen-detected cancers (17). As an exploratory analysis, we also identified potential interval cancers by applying a modified version of this algorithm to the period between 5 and 12 months after the index screening test. Interval cancers are those identified clinically before the next scheduled mammogram (18). Details of these definitions are provided in the [Supplementary Methods](#) and [Supplementary Table 1](#) (available online).

Finally, we calculated the positive predictive value (PPV) among those recalled as the proportion of women diagnosed with cancer among those who had recall imaging (diagnostic mammogram, diagnostic DBT, or diagnostic ultrasound). We also calculated the positive predictive value among those who underwent biopsy, a value that is similar to the “PPV3” metric commonly used in mammography benchmarking (19).

### Covariates

We evaluated clinical and demographic differences between women screened with 2D and DBT, including age, use of supplemental screening ultrasound, family history of breast cancer, screening time period, residence in a metropolitan region, time since last breast cancer screening, and hospital referral region (HRR) of residence. Women were assigned to HRRs based on zip code of residence (20). Supplemental ultrasound was defined by a claim for bilateral, whole-breast ultrasound. Additional details of covariates are included in the [Supplementary Methods](#) and [Supplementary Table 1](#) (available online).

### Mammogram-Level Analysis

We evaluated demographic and clinical differences among women screened with 2D and DBT using  $\chi^2$  testing. For the mammogram-level analysis, we used both unadjusted and multivariable logistic regression to model the relationship between screening type (DBT or 2D mammography) and screening outcome (recall, subsequent diagnostic testing, and cancer detection). Multivariable models were adjusted for age, metro location, family history of breast cancer, time since last screening, use of supplemental screening ultrasound, time period, and HRR modeled as fixed effects. We included cluster-robust standard errors at the person level to account for the correlation of mammogram results in the same woman. We also performed

**Table 1.** Sample selection

No.	Remaining, %	Description
57 859 780	—	Female and aged 40+ y in 2015-2017
9 153 381	15.8	Had a claim for potential screening mammogram in 2015-2017 (primary procedures)
9 123 786	15.8	Did not have a claim for any type of mammogram in the 9 mo before potential screening mammogram previously identified
9 092 718	15.7	Are 40+ y on day of screening mammogram
4 861 039	8.4	Had continuous coverage 24 mo before through 12 mo after potential screening mammogram previously identified
4 821 596	8.3	Identified potential screening mammogram did not have any breast cancer diagnosis in the 24 mo before that screening mammogram on first diagnosis
4 781 592	8.3	Identified potential screening mammogram did not have any breast cancer diagnosis in the 24 mo before that screening mammogram on any diagnosis
4 781 560	8.3	Did not have claim for prophylactic removal of breast in the 24 mo before potential screening mammogram
4 778 342	8.3	Did not have claim for genetic susceptibility to breast cancer in the 24 mo before potential screening mammogram
4 760 968	8.2	Screening mammogram allows woman to be assigned to single hospital referral region
4 580 698	7.9	Were aged 40-64 y at mammogram

analyses for main outcome measures stratified by time period and by age category. For all multivariable regression analyses, we reported the results as risk-adjusted rates per 1000 screening mammograms based on regression model-predicted probabilities of the outcome evaluated at observed values for each patient. We also report regression results as odds ratios in the [Supplementary Table 2](#) (available online). As an alternate specification, we fit a multilevel mixed-effects model with patients clustered within HRRs (see the [Supplementary Methods](#), available online). Because we evaluated multiple screening outcomes in multiple models, we used a more conservative *P* value of less than .01 to suggest statistical significance. All statistical tests were 2-sided.

### Area-Level Analysis

We also performed a longitudinal, area-level analysis to quantify the relationship between population-level DBT use and population-level changes in screening outcomes among HRRs during the study period. Here, the unit of analysis was the HRR, and DBT use and screening outcomes were observed for each HRR in 6-month intervals over a 3-year period. This area-level approach was designed to be relatively robust to confounding by indication ([Supplementary Figure 1](#), available online). Likewise, the longitudinal design accounts for factors that vary across regions such as differences in population risk, referral patterns, or radiologist practice style that may confound cross-sectional analyses (21).

For this area-level analysis, we used the same exposure and outcome definitions as in the mammogram-level analysis, but both the exposure (DBT use) and outcomes were calculated as the rate per 1000 women screened in each HRR at 6-month intervals. For example, we defined DBT use as the number of women who received DBT per 1000 women screened in an HRR

in a 6-month interval. Likewise, we measured the number of recalls, biopsies, invasive cancers, and potential interval cancers per 1000 women screened over a 6-month period in each HRR.

The area-level analysis used linear regression to evaluate the relationship between DBT use and screening outcomes among HRRs. Models were adjusted for area-level use of screening ultrasound, HRR, and time period fixed effects. We weighted models screened population size in each HRR and clustered standard errors by HRR. We expressed model results using marginal effects assuming 1% DBT use in a population or 99% DBT use in a population. Models were evaluated using standard methods, described in the [Supplementary Methods](#) (available online). In a sensitivity analysis, we excluded HRR-level observations with fewer than 5 instances of the outcome or covariates.

We used Stata versions 16.0 SE and 14.2 SE for statistical analyses. Statistical code is available from the authors on request and as permitted by the data use agreement.

### Results

Our final sample included 7 602 869 screening mammograms performed in 4 580 698 women ([Table 1](#)). During the study period, DBT was used in 27.5% of screening exams overall. DBT use increased from 12.0% of screening exams in early 2015 to 43.2% of screening exams in late 2017. DBT screens were more common among women who lived in a metropolitan area (89.9% vs 84.7%, *P* < .001) and among women who had received a mammogram in the past 24 months (59.9% vs 55.1%, *P* < .001; [Table 3](#)).

In the mammogram-level analysis, after adjusting for covariates, the recall rate (diagnostic 2D mammography, diagnostic DBT, and/or diagnostic ultrasound) was slightly lower among

**Table 2.** Mammograms per patient

Mammograms per screened woman, No.	Women, No.	Total mammograms, No.
1	2 649 990	2 649 990
2	1 379 355	2 758 710
3	731 243	2 193 729
4	110	440
Total	4 760 698	7 602 869

**Table 3.** Characteristics of study sample by screening type

Characteristic	2 D, No. (%)	DBT, No. (%)	P <sup>a</sup>	
Total	5 378 213 (72.5)	2 044 656 (27.5)		
Age, y				
40-44	681 652 (12.7)	276 124 (13.5)	<.001	
45-49	974 359 (18.1)	388 929 (19.0)		
50-54	1 190 903 (22.1)	459 060 (22.5)		
55-59	1 340 285 (24.9)	496 736 (24.3)		
60-64	1 191 014 (22.1)	423 807 (20.7)		
Metro status				
Nonmetro	810 063 (15.1)	204 497 (10.0)	<.001	
Metro	4 557 281 (84.7)	1 837 285 (89.9)		
Unknown metro status	10 869 (0.2)	2874 (0.1)		
Timing of index mammogram				
January 1, 2015-June 30, 2015	972 514 (18.1)	132 436 (6.5)	<.001	
July 1, 2015-December 31, 2015	1 073 886 (20.0)	216 883 (10.6)		
January 1, 2016-June 30, 2016	865 146 (16.1)	267 548 (13.1)		
July 1, 2016-December 31, 2016	932 066 (17.3)	394 516 (19.3)		
January 1, 2017-June 30, 2017	759 036 (14.1)	443 101 (21.7)		
July 1, 2017-December 31, 2017	775 565 (14.4)	590 172 (28.9)		
Months since last mammogram				
9-12	107 654 (2.0)	44 056 (2.2)		<.001
12-24	2 855 130 (53.1)	1 179 740 (57.7)		
>24	462 672 (8.6)	189 810 (9.3)		
Not observed	1 952 757 (36.3)	631 050 (30.9)		
Family history of breast cancer				
No	5 046 692 (93.8)	1 862 544 (91.1)	<.001	
Yes	331 521 (6.2)	182 112 (8.9)		
Observed mammogram order				
First	3 542 287 (65.9)	1 038 411 (50.8)	<.001	
Second	1 415 961 (26.3)	694 747 (34.0)		
Third	419 915 (7.8)	311 438 (15.2)		
Fourth	50 (0.001)	60 (0.003)		

<sup>a</sup>P value calculated using a  $\chi^2$  test. 2 D = 2-dimensional; DBT = digital breast tomosynthesis.

women screened with DBT compared with those screened with 2D mammography (113.6 recalls per 1000 screens, 99% CI = 113.8 to 114.2 vs 115.4, 99% CI = 115.0 to 115.8,  $P < .001$ ; [Table 4](#)). In age-stratified analyses, women aged 40-44 years screened with DBT had the largest reduction in recall (-11.5 recalls per 1000 screens, 99% CI = -18.8 to -9.1,  $P < .001$ ; [Table 6](#)). In adjusted analyses, women screened with DBT were also more likely to undergo biopsy than those screened with 2D mammography (19.6 biopsies per 1000 screens, 99% CI = 19.3 to 19.8 vs 15.2, 99% CI = 15.1 to 15.4,  $P < .001$ ; [Table 4](#)). Among those who had a biopsy, the proportion of women who were ultimately diagnosed with breast cancer was similar (PPV3 = 24.0%, 99% CI = 23.4% to 24.6% among those screened with DBT vs 23.6%, 99% CI = 23.2% to 24.0% for 2D,  $P = .24$ ) ([Table 4](#)).

In adjusted analyses, DBT was associated with a higher rate of cancer detection compared with 2D mammography (4.9 incident cancers per 1000 screens, 99% CI = 4.7 to 5.0 vs 3.8, 99% CI = 3.7 to 3.9,  $P < .001$ ; [Table 4](#)). DBT was associated with higher cancer detection rates across all age groups ([Table 6](#)). Cancers diagnosed between 5 and 12 months after mammography, which may include interval cancers, were slightly higher, with DBT vs 2D (0.52 incident cancers per 1000 screens, 99% CI = 0.47 to 0.56 vs 0.46, 99% CI = 0.43 to 0.48,  $P < .001$ ; [Table 4](#)). Unadjusted results are reported in [Table 4](#) and were generally similar. Multilevel models with HRR-level random effects also gave similar results ([Table 5](#)).

In analyses stratified by time period, reductions in recall varied by time period, with smaller reductions in recall earlier on (reduction in recall of -0.2 per 1000 screens, 99% CI = -2.8 to 2.4 January-June 2015 vs a reduction in recall of -1.3 per 1000, 99% CI = -2.9 to 0.3 in July-December 2017,  $P_{\text{interaction}} = .007$ ; [Table 7](#)). In contrast, differences in the risk-adjusted rate of cancer detection did not statistically significantly change over time (difference of 1.3 per 1000 screens, 99% CI = 0.7 to 1.9 in January-June 2015 and 0.9 per 1000, 99% CI = 0.7 to 1.4 in July-December 2017,  $P_{\text{interaction}} = .08$ ; [Table 7](#)).

The area-level analysis modeled the relationship between changes in population-level DBT use and population-level screening outcomes over time at the regional level. We found that at the population level, adoption of screening DBT was not associated with a statistically significant change in recall rate. Specifically, we found that if DBT were used in 1% of screening exams, recall would be 117.2 per 1000 screens (99% CI = 114.1 to 120.4) vs a recall rate of 109.1 per 1000 (99% CI = 100.4 to 117.7) if DBT were used in 99% of exams ([Table 8](#)). Using DBT for 1% of screening exams would be associated with a cancer detection rate of 4.0 per 1000 screens (99% CI = 3.7 to 4.2) vs a cancer detection rate of 4.4 per 1000 (99% CI = 3.8 to 5.1) if DBT were used in 99% screening exams ([Table 8](#)). A sensitivity analysis that excluded HRRs based on a minimum volume had similar results, though estimated recall rates were higher overall and the reduction in recall with DBT was smaller.



Table 4. Screening outcomes, mammogram-level analysis

Outcome per 1000	Unadjusted			Adjusted <sup>d</sup>			P <sup>c</sup>
	2 D (99% CI)	DBT (99% CI)	Difference (99% CI)	2 D (99% CI)	DBT (99% CI)	Difference (99% CI)	
Any recall	114.8 (114.5 to 115.2)	115.2 (114.6 to 115.8)	0.4 (-0.3 to 1.1)	115.4 (115.0 to 115.8)	113.6 (113.0 to 114.2)	-1.8 (-2.6 to 1.1)	<.001
Diagnostic 2 D mammogram	90 (90 to 91)	75 (75 to 76)	-15.4 (-16.0 to -14.9)	91.6 (91.2 to 91.9)	72.7 (72.2 to 73.2)	-18.9 (-19.5 to 18.2)	<.001
Diagnostic DBT	11.2 (11.0 to 11.3)	41.2 (40.8 to 41.5)	30.0 (29.6 to 30.4)	13.0 (12.8 to 13.1)	30.3 (30.0 to 30.6)	17.3 (17.0 to 17.7)	<.001
Diagnostic ultrasound	78.1 (77.8 to 78.4)	87.9 (87.4 to 88.4)	9.8 (9.2 to 10.4)	78.1 (77.7 to 78.4)	88.1 (87.5 to 88.7)	10.0 (9.3 to 10.7)	<.001
MRI	4.0 (4.0 to 4.1)	5.6 (5.5 to 5.8)	1.6 (1.5 to 1.8)	4.0 (4.0 to 4.1)	5.6 (5.5 to 5.8)	1.6 (1.4 to 1.8)	<.001
Biopsy	15.4 (15.2 to 15.5)	19.1 (18.9 to 19.4)	3.8 (3.5 to 4.0)	15.2 (15.1 to 15.4)	19.6 (19.3 to 19.8)	4.4 (4.0 to 4.7)	<.001
Incident cancer							
0-4 mo	3.8 (3.7 to 3.9)	4.8 (4.7 to 4.9)	0.97 (0.83 to 1.12)	3.8 (3.7 to 3.9)	4.9 (4.7 to 5.0)	1.1 (0.9 to 1.3)	<.001
5-12 mo	0.46 (0.43 to 0.48)	0.5 (0.5 to 0.5)	0.04 (-0.01 to 0.09)	0.45 (0.43 to 0.48)	0.52 (0.47 to 0.56)	0.07 (0.01 to 0.12)	.002
Positive predictive value							
Recall <sup>a</sup>	3.3 (3.2 to 3.3)	4.1 (4.0 to 4.2)	0.8 (0.7 to 1.0)	3.2 (3.2 to 3.3)	4.2 (4.1 to 4.4)	1.01 (0.87 to 1.15)	<.001
Biopsy <sup>b</sup>	23.6 (23.2 to 24.1)	24.1 (23.5 to 24.6)	0.5 (-0.2 to 1.2)	23.6 (23.2 to 24.0)	24.0 (23.4 to 24.6)	0.34 (-0.41 to 1.09)	.24

<sup>a</sup>Calculated as number of women diagnosed with cancer per 1000 women who had subsequent diagnostic imaging. CI = confidence interval; 2D = 2-dimensional; DBT = digital breast tomosynthesis; HRR = hospital referral region;

MRI = magnetic resonance imaging.

<sup>b</sup>Calculated as percent of women diagnosed with cancer from among those who underwent biopsy.

<sup>c</sup>Two-sided P values given from unadjusted or adjusted logistic regression models.

<sup>d</sup>Models adjusted for use of screening ultrasound, age, time period of index mammogram, time since last mammogram, metro location, HRR, and family history of breast cancer.

## Discussion

We evaluated the comparative effectiveness of DBT and 2D mammography in a large privately insured population of women aged 40-64 years. In contrast to a number of other US studies, we observed only a small reduction in recall with DBT compared with 2D mammography. Other studies have demonstrated a typical reduction in recall of 22 per 1000, which is substantially larger than the 1.8 per 1000 observed here (8). Reduction in recall has been cited as a major advantage of DBT because it may translate into lower costs, more efficient subsequent workup, and less potential anxiety for women undergoing screening while preserving sensitivity (2). However, our results suggest that recall may not be meaningfully reduced among women screened with DBT compared with women screened with 2D mammography.

There are several possible explanations for our findings related to recall. First, performance improvements with DBT may simply be smaller than previously reported in initial studies, as this technology is applied to more heterogeneous populations and in less controlled settings. Second, time may play a role. Our study captures relatively early experience with DBT, and several studies have shown that reductions in recall may be smaller earlier after adoption. This apparent effect of time may be driven by several factors, including a patient's previous screening history and the availability of prior screens for comparison (22-25). Last, a variety of other factors, such as differences in resolution across imaging systems, may influence recall rates (26).

It is important to note that our study was observational, and selection bias may also account for the small reduction in recall noted in our study compared with other studies. For example, if higher-risk women are preferentially referred to DBT, this may attenuate any apparent reduction in recall. We designed an area-level analysis to address this possibility. Point estimates from the area-level analysis also suggested only a modest reduction in recall. Taken together, our findings suggest that when DBT use is extended to typical populations, DBT may offer only a small improvement in recall compared with 2D mammography.

Our findings also have implications for the value of DBT. A recent cost-effectiveness study found that DBT was not cost-effective compared with 2D mammography at current DBT prices but could be cost-effective at a lower price (27). This cost-effectiveness analysis assumed approximately a 5%-10% improvement in specificity with DBT. Although we did not formally evaluate specificity, our findings imply that specificity is not markedly improved with DBT and the overall cost effectiveness of DBT may be even less than originally expected.

We also found that DBT is associated with a higher cancer detection rate compared with 2D mammography. This finding was similar in magnitude to previous studies, which have also estimated an approximately 1 per 1000 increase in screen-detected cancers (11). Though not statistically significant, the point estimate from the area-level analysis was quite similar and supports findings from the patient-level analysis. Overall, our results suggest that DBT does indeed have a higher cancer detection rate than mammography and that this greater sensitivity is sustained even in the diverse practice settings studied here.

A key question is whether this increase in cancer detection will translate into an improvement in health. Although a higher cancer detection rate seems advantageous, identifying more cancers does not necessarily translate into improved health and may only contribute to longer lead times. Further, more sensitive tests can actually cause harm through overdiagnosis. Overdiagnosis and resultant overtreatment of some breast

Table 5. Multilevel model

Outcome per 1000 <sup>a</sup>	2 D (99% CI)	DBT (99% CI)	Difference (99% CI)	P <sup>†</sup>
Any recall	111.1 (107.5 to 114.7)	109.5 (106.0 to 113.1)	-1.6 (-2.2 to 1.0)	<.001
Biopsy	15.6 (15.2 to 16.0)	20.0 (19.5 to 20.6)	4.4 (4.2 to 4.7)	<.001
Incident cancer				
0-4 mo	3.8 (3.7 to 3.9)	4.9 (4.8 to 5.1)	1.1 (1.0 to 1.2)	<.001
5-12 mo	0.45 (0.43 to 0.47)	0.50 (0.47 to 0.54)	0.05 (0.01 to 0.09)	.009

<sup>a</sup>Model includes HRR-level random effects and mammograms are nested within HRRs. Models adjusted for use of screening ultrasound, age, time period of index mammogram, time since last mammogram, metro location, and family history of breast cancer as fixed effects. CI = confidence interval; 2D = 2-dimensional; DBT = digital breast tomosynthesis; HRR = hospital referral region.

<sup>†</sup>Two-sided P value is from multilevel models.

Table 6. Adjusted mammogram-level results, stratified by age group

Outcome per 1000 <sup>a</sup>	2 D (99% CI)	DBT (99% CI)	Difference (99% CI)	P <sup>b</sup>	P <sub>interaction</sub> <sup>c</sup>
Any recall	115.4 (115.0 to 115.8)	113.6 (113.0 to 114.2)	-1.8 (-2.6 to -1.1)	<.001	<.001
40-44-y age group	164 (163 to 166)	153 (151 to 155)	-11.5 (-13.8 to -9.1)	<.001	
45-49-y age group	137 (136 to 137)	135 (133 to 136)	-1.8 (-3.7 to 0.1)	.02	
50-54-y age group	114 (114 to 115)	116 (115 to 118)	1.9 (0.3 to 3.5)	.003	
55-59-y age group	96 (95 to 97)	97 (95 to 98)	0.7 (-0.7 to 2.2)	.18	
60-64-y age group	92 (91 to 93)	89 (87 to 90)	-3.2 (-4.7 to -1.7)	<.001	
Biopsy	15.2 (15.1 to 15.4)	19.6 (19.3 to 19.8)	4.4 (4.0 to 4.7)	<.001	.001
40-44-y age group	19.0 (18.6 to 19.5)	23.7 (22.9 to 24.6)	4.7 (3.7 to 5.7)	<.001	
45-49-y age group	17.0 (16.6 to 17.3)	21.6 (20.9 to 22.2)	4.6 (3.8 to 5.4)	<.001	
50-54-y age group	15.3 (15.0 to 15.6)	20.3 (19.7 to 20.9)	5.0 (4.3 to 5.7)	<.001	
55-59-y age group	13.3 (13.0 to 13.6)	17.1 (16.6 to 17.7)	3.8 (3.2 to 4.5)	<.001	
60-64-y age group	13.6 (13.3 to 13.9)	17.4 (16.8 to 18.0)	3.8 (3.1 to 4.5)	<.001	
Cancer detection, 0-4 mo	3.8 (3.7 to 3.9)	4.9 (4.7 to 5.0)	1.1 (0.9 to 1.3)	<.001	.001
40-44-y age group	2.3 (2.2 to 2.5)	3.2 (2.8 to 3.5)	0.8 (0.5 to 1.2)	<.001	
45-49-y age group	3.2 (3.0 to 3.3)	3.9 (3.6 to 4.2)	0.8 (0.4 to 1.1)	<.001	
50-54-y age group	3.6 (3.5 to 3.8)	4.8 (4.5 to 5.1)	1.2 (0.8 to 1.5)	<.001	
55-59-y age group	4.0 (3.9 to 4.2)	5.3 (5.0 to 5.6)	1.3 (1.0 to 1.7)	<.001	
60-64-y age group	5.1 (4.9 to 5.3)	6.3 (5.9 to 6.6)	1.2 (0.8 to 1.6)	<.001	
Cancer detection, 5-12 mo	0.45 (0.43 to 0.48)	0.52 (0.47 to 0.56)	0.07 (0.01 to 0.12)	.002	.54
40-44-y age group	0.5 (0.4 to 0.6)	0.6 (0.4 to 0.7)	0.04 (-0.13 to 0.21)	.51	
45-49-y age group	0.5 (0.5 to 0.6)	0.6 (0.5 to 0.7)	0.03 (-0.12 to 0.17)	.64	
50-54-y age group	0.5 (0.4 to 0.5)	0.7 (0.5 to 0.8)	0.18 (0.04 to 0.32)	.001	
55-59-y age group	0.5 (0.4 to 0.5)	0.5 (0.4 to 0.6)	0.02 (-0.09 to 0.13)	.69	
60-64-y age group	0.5 (0.5 to 0.6)	0.6 (0.5 to 0.7)	0.08 (-0.05 to 0.22)	.10	

<sup>a</sup>Models adjusted for use of screening ultrasound, time period of index mammogram, time since last mammogram, metro location, hospital referral region, and family history of breast cancer. CI = confidence interval; 2D = 2-dimensional; DBT = digital breast tomosynthesis.

<sup>b</sup>Two-sided P value is from multivariable logistic regression and indicates whether the difference in outcome is statistically significantly different from zero.

<sup>c</sup>Two-sided P value for interaction between screening type and age group from multivariable logistic regression model.

cancers are increasingly recognized as challenging consequences of mammography (28). Despite the increase in cancer detection, we did not see a decline in potential interval cancers up to a year after screening, albeit using an exploratory definition. Ongoing studies, including the Tomosynthesis Mammographic Imaging Screening trial, will help answer the question of whether DBT reduces the incidence of aggressive interval cancers or whether its increased sensitivity merely picks up small, slower-growing breast cancers (29). However, it may be years before that trial is complete. As screening becomes more sensitive, we will need improved strategies for understanding which screen-detected cancers are likely to progress and which are not.

This work has some important limitations. First, we used administrative claims that provide information about use of clinical services and outcomes but can be subject to error, misclassification, and underreporting. When possible, we used validated algorithms to identify key variables, including

screening mammography and incident breast cancers. To identify DBT use, we built on established algorithms. One limitation of this approach is that it may underidentify DBT use if providers did not submit claims because DBT was not a covered service. However, rates of DBT use observed in this study were similar to use in Medicare, which did reimburse uniformly beginning in 2015 (6). Our definition of recall also relied on observing subsequent imaging, which may undercount recall if women do not return for follow-up. However, this is likely to be a small proportion of women, and our recall rates were actually somewhat higher overall than observed elsewhere.

Lastly, this was an observational study, which has inherent limitations for assessing causality. Indeed, we observed important demographic and clinical differences between women screened with DBT and those screened with 2D mammography. However, we adjusted for observed differences in our patient-level analysis, and our longitudinal, area-level approach, which

Table 7. Adjusted mammogram-level results, stratified by time period

Outcome per 1000 <sup>a</sup>	2 D (99% CI)	DBT (99% CI)	Difference (99% CI)	P <sup>b</sup>	P <sub>interaction</sub> <sup>c</sup>
Recall	115.4 (115.0 to 115.8)	113.6 (113.0 to 114.2)	-1.8 (-2.6 to -1.1)	<.001	.007
Jan-Jun 2015	118 (117 to 119)	118 (115 to 120)	-0.2 (-2.8 to 2.4)	.84	
July-Dec 2015	113 (112 to 114)	111 (109 to 113)	-2.1 (-4.1 to 0.0)	.009	
Jan-Jun 2016	116 (115 to 117)	115 (114 to 117)	-0.8 (-2.8 to 1.2)	.30	
July-Dec 2016	114 (113 to 115)	111 (110 to 113)	-2.7 (-4.4 to 1.0)	<.001	
Jan-Jun 2017	116 (115 to 117)	115 (113 to 116)	-1.8 (-3.5 to 0.01)	.01	
July-Dec 2017	115 (114 to 116)	114 (113 to 115)	-1.3 (-2.9 to 0.3)	.04	
Biopsy <sup>c</sup>	15.2 (15.1 to 15.4)	19.6 (19.3 to 19.8)	4.4 (4.0 to 4.7)	<.001	.001
Jan-Jun 2015	16.0 (15.7 to 16.3)	20.7 (19.6 to 21.8)	4.7 (3.5 to 5.8)	<.001	
July-Dec 2015	15.1 (14.8 to 15.4)	19.5 (18.6 to 20.3)	4.4 (3.5 to 5.3)	<.001	
Jan-Jun 2016	15.5 (15.1 to 15.8)	20.1 (19.3 to 20.9)	4.6 (3.7 to 5.5)	<.001	
July-Dec 2016	15.1 (14.7 to 15.4)	19.5 (18.9 to 20.1)	4.5 (3.7 to 5.2)	<.001	
Jan-Jun 2017	15.1 (14.8 to 15.5)	19.5 (18.9 to 20.1)	4.4 (3.6 to 5.1)	<.001	
July-Dec 2017	14.7 (14.3 to 15.1)	18.6 (18.1 to 19.1)	3.9 (3.3 to 4.6)	<.001	
Cancer detection, 0-4 mo	3.8 (3.7 to 3.9)	4.9 (4.7 to 5.0)	1.1 (0.9 to 1.3)	<.001	.08
Jan-Jun 2015	3.8 (3.6 to 3.9)	5.0 (4.5 to 5.6)	1.3 (0.7 to 1.9)	<.001	
July-Dec 2015	3.7 (3.5 to 3.8)	5.0 (4.5 to 5.4)	1.3 (0.8 to 1.8)	<.001	
Jan-Jun 2016	3.9 (3.7 to 4.1)	5.0 (4.7 to 5.4)	1.1 (0.7 to 1.6)	<.001	
July-Dec 2016	3.9 (3.7 to 4.0)	5.0 (4.7 to 5.4)	1.2 (0.8 to 1.6)	<.001	
Jan-Jun 2017	3.8 (3.6 to 4.0)	4.9 (4.6 to 5.2)	1.1 (0.7 to 1.4)	<.001	
July-Dec 2017	3.8 (3.7 to 4.0)	4.7 (4.5 to 5.0)	0.9 (0.5 to 1.2)	<.001	
Cancer detection, 5-12 mo	0.45 (0.43 to 0.48)	0.52 (0.47 to 0.56)	0.07 (0.01 to 0.12)	.002	.54
Jan-Jun 2015	0.58 (0.51 to 0.64)	0.64 (0.45 to 0.84)	0.07 (-0.14 to 0.28)	.41	
July-Dec 2015	0.52 (0.46 to 0.58)	0.62 (0.46 to 0.79)	0.10 (-0.08 to 0.28)	.14	
Jan-Jun 2016	0.52 (0.45 to 0.59)	0.61 (0.47 to 0.76)	0.10 (-0.07 to 0.26)	.12	
July-Dec 2016	0.52 (0.46 to 0.59)	0.53 (0.43 to 0.63)	0.01 (-0.12 to 0.13)	.91	
Jan-Jun 2017	0.44 (0.37 to 0.51)	0.60 (0.48 to 0.72)	0.16 (0.01 to 0.30)	<.001	
July-Dec 2017	0.49 (0.42 to 0.57)	0.54 (0.44 to 0.63)	0.04 (-0.08 to 0.17)	.39	

<sup>a</sup>Models adjusted for use of screening ultrasound, age, time since last mammogram, metro location, hospital referral region, and family history of breast cancer. CI = confidence interval; 2 D = 2-dimensional; DBT = digital breast tomosynthesis.

<sup>b</sup>Two-sided P value is from multivariable logistic regression and indicates whether the difference in outcome is statistically significantly different from zero.

<sup>c</sup>Two-sided P value for interaction between screening type and age group.

Table 8. Screening outcomes in the area-level analysis

Outcome per 1000 women screened <sup>a</sup>	HRRs, No.	1% Population use of DBT (99% CI)	99% Population use of DBT (99% CI)	Difference	P <sup>b</sup>
All HRRs					
Recall	306	117.2 (114.1 to 120.4)	109.1 (100.4 to 117.7)	-8.1	.08
Biopsy	306	15.7 (15.1 to 16.2)	18.4 (16.9 to 19.9)	2.7	.001
Incident cancer					
0-4 mo	306	3.95 (3.72 to 4.18)	4.44 (3.80 to 5.07)	0.49	.15
5-12 mo	306	0.69 (0.61 to 0.77)	0.51 (0.29 to 0.72)	-0.18	.10
Excluding HRRs with N < 5 per outcome					
Recall	154	124.1 (118.7 to 129.5)	117.7 (104.5 to 130.9)	-6.4	.38
Biopsy	153	15.7 (14.9 to 16.6)	18.8 (16.7 to 20.9)	3.1	.007
Incident cancer					
0-4 mo	143	3.96 (3.66 to 4.26)	4.50 (3.76 to 5.23)	0.54	.18
5-12 mo	73	0.79 (0.66 to 0.92)	0.57 (0.28 to 0.86)	-0.22	.18

<sup>a</sup>Outcome indicates number per 1000 women screened, assuming 1% of the population uses DBT or 99% of the population uses DBT. Analyses adjusted for HRR fixed effects, time period fixed effects, and use of screening ultrasound per 1000 screened women in each time period. Models were weighted by the size of the screened population in each HRR. CI = confidence interval; DBT = digital breast tomosynthesis; HRR = hospital referral region.

<sup>b</sup>Two-sided P value for coefficient for DBT in multivariable model.

is relatively robust to confounding, produced results supportive of our main patient-level analysis.

Using administrative data from a large privately insured population, we evaluated screening outcomes after DBT compared with 2D mammography. Our results suggest that DBT is associated with a higher cancer detection rate than 2D mammography but may not substantially reduce the need for

subsequent imaging in typical practice. A critical remaining question is whether improved cancer detection translates into reductions in morbidity and mortality for women diagnosed with breast cancer. Randomized prospective studies and long-term follow-up are needed to determine whether DBT technology improves outcomes in women undergoing regular mammographic screening.

## Funding

This work was supported by National Institutes of Health/National Center for Advancing Translational Sciences KL2 TR001862 (Richman) and American Cancer Society RSGI-15-151-01 (Gross, Xu, Busch).

## Notes

**Role of the funder:** The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Disclosures:** Drs. Richman and Kyanko report salary support from the Centers for Medicare and Medicaid Services to develop health-care quality measures outside of the submitted work. Dr Gross reports research grants from NCCN/Pfizer Astra-Zeneca, Johnson & Johnson, and Genentech, and travel funding from Flatiron Health. Ms Kunst reports funding from the Research Council of Norway and LINK Medical Research (grant numbers 276146 and 304034) during the conduct of the study, and personal fees for speaking from Thermo Fisher Scientific. Professor Busch reports research funding from NIDA, NIMH, and the Robert Wood Johnson Foundation. Dr Hooley reports honoraria from Hologic. Jessica Hoag is an employee of CATO SMS and was an employee of ICON, Plc during this work.

**Author contributions:** IR: study concept and design, interpretation of results, drafting of the manuscript; JBL: study design, data analysis, editorial contributions; JRH: study design, data analysis; AU: study design, editorial contributions; XX: interpretation of results, methodologic expertise, editorial contributions; JAA: interpretation of results, editorial contributions; KAK: interpretation of results, editorial contributions; NK: interpretation of results, editorial contributions; SHB: methodologic expertise, interpretation of results, editorial contributions; CPG: study concept and design, interpretation of results, editorial contributions, contribution of materials and resources.

**Prior presentations:** A version of this work was presented at the Society of General Internal Medicine annual meeting in May 2019.

## Data Availability

The authors are unable to share primary data due to the identifiable and sensitive nature of the data used.

## References

- Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation. *Ann Intern Med*. 2016;164(4):244–255. doi:10.7326/M15-0969
- Hooley RJ, Durand MA, Philpotts LE. Advances in digital breast tomosynthesis. *AJR Am J Roentgenol*. 2017;208(2):256–266. doi:10.2214/AJR.16.17127
- Aujero MP, Gavenonis SC, Benjamin R, Zhang Z, Holt JS. Clinical Performance of synthesized two-dimensional mammography combined with tomosynthesis in a large screening population. *Radiology*. 2017;283(1):70–76. doi:10.1148/radiol.2017162674
- 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. doi:10.1001/jamanetworkopen.2020.11792
- Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA*. 2014;311(24):2499. doi:10.1001/jama.2014.6095
- Conant EF, Barlow WE, Herschorn SD, et al.; for the Population-based Research Optimizing Screening Through Personalized Regimen (PROSPR) Consortium. Association of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density. *JAMA Oncol*. 2019;5(5):635–642. doi:10.1001/jamaoncol.2018.7078
- Greenberg JS, Javitt MC, Katzen J, Michael S, Holland AE. Clinical performance metrics of 3D digital breast tomosynthesis compared with 2D digital mammography for breast cancer screening in community practice. *AJR Am J Roentgenol*. 2014;203(3):687–693. doi:10.2214/AJR.14.12642
- Richman IB, Hoag JR, Xu X, et al. Adoption of digital breast tomosynthesis in clinical practice. *JAMA Intern Med*. 2019;179(9):1292. doi:10.1001/jamainternmed.2019.1058
- MQSA National Statistics. <https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics>. Published May 1, 2021. Accessed May 5, 2021.
- Fujii MH, Herschorn SD, Sowden M, et al. Detection rates for benign and malignant diagnoses on breast cancer screening with digital breast tomosynthesis in a statewide mammography registry study. *AJR Am J Roentgenol*. 2019;212(3):706–711. doi:10.2214/AJR.18.20255
- Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. *JNCI J Natl Cancer Inst*. 2018;110(9):942–949. doi:10.1093/jnci/djy121
- Maxwell AJ, Michell M, Lim YY, et al. A randomised trial of screening with digital breast tomosynthesis plus conventional digital 2D mammography versus 2D mammography alone in younger higher risk women. *Eur J Radiol*. 2017;94:133–139. doi:10.1016/j.ejrad.2017.06.018
- Pattacini P, Nitrosi A, Giorgi Rossi P, et al.; for the RETomo Working Group. Digital mammography versus digital mammography plus tomosynthesis for breast cancer screening: the Reggio Emilia Tomosynthesis randomized trial. *Radiology*. 2018;288(2):375–385. doi:10.1148/radiol.2018172119
- Hofvind S, Hølen AS, Aase HS, et al. Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): a randomised, controlled trial. *Lancet Oncol*. 2019;20(6):795–805. doi:10.1016/S1470-2045(19)30161-5
- Fenton JJ, Zhu W, Balch S, Smith-Bindman R, Fishman P, Hubbard RA. Distinguishing screening from diagnostic mammograms using Medicare claims data. *Med Care*. 2014;52(7):e44–e51. doi:10.1097/MLR.0b013e318269e0f5
- Medicare Claims Processing Manual, Transmittal #3160. <https://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/Downloads/R3160CP.pdf>. Published January 7, 2015. Accessed May 5, 2021.
- Fenton JJ, Onega T, Zhu W, et al. Validation of a Medicare claims-based algorithm for identifying breast cancers detected at screening mammography. *Med Care*. 2016;54(3):e15–e22. doi:10.1097/MLR.0b013e3182a303d7
- Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*. 2017;3:12. doi:10.1038/s41523-017-0014-x
- Sprague BL, Aaro RF, Miglioretti DL, et al.; Breast Cancer Surveillance Consortium. National performance benchmarks for modern diagnostic digital mammography: update from the breast cancer surveillance consortium. *Radiology*. 2017;283(1):59–69. doi:10.1148/radiol.2017161519
- The Dartmouth Atlas of Health Care. <http://www.dartmouthatlas.org/tools/downloads.aspx?tab=39>. Accessed May 7, 2021.
- Angrist JD, Krueger AB. Empirical Strategies in Labor Economics. In: Ashenfelter OC, Card D, eds. *Handbook of Labor Economics* Vol. 3, Part A. North Holland: Elsevier; 1999:1277–1366.
- Hovda T, Brandal SHB, Sebuødegård S, et al. Screening outcome for consecutive examinations with digital breast tomosynthesis versus standard digital mammography in a population-based screening program. *Eur Radiol*. 2019;29(12):6991–6999.
- Conant EF, Zuckerman SP, McDonald ES, et al. Five consecutive years of screening with digital breast tomosynthesis: outcomes by screening year and round. *Radiology*. 2020;295(2):285–293. doi:10.1148/radiol.2020191751
- 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. doi:10.1001/jamanetworkopen.2020.1759
- Miglioretti DL, Abraham L, Lee CI, et al.; for the Breast Cancer Surveillance Consortium. Digital breast tomosynthesis: radiologist learning curve. *Radiology*. 2019;291(1):34–42. doi:10.1148/radiol.2019182305
- Hadjipanteli A, Kontos M, Constantinidou A. The role of digital breast tomosynthesis in breast cancer screening: a manufacturer- and metrics-specific analysis. *Cancer Manag Res*. 2019;11:9277–9296. doi:10.2147/CMAR.S210979
- Lowry KP, Trentham-Dietz A, Schechter CB, et al. Long-term outcomes and cost-effectiveness of breast cancer screening with digital breast tomosynthesis in the United States. *J Natl Cancer Inst*. 2020;112(6):582–589. doi:10.1093/jnci/djz184
- Brawley OW. Accepting the existence of breast cancer overdiagnosis. *Ann Intern Med*. 2017;166(5):364–365. doi:10.7326/m16-2850
- TMIST (Tomosynthesis Mammographic Imaging Screening Trial). <https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/tmist>. Published October 3, 2019. Accessed May 5, 2021.