JNCI J Natl Cancer Inst (2016) 108(4): djv367

doi:10.1093/jnci/djv367
First published online December 28, 2015
Article

ARTICLE

Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666

Wendie A. Berg, Andriy I. Bandos, Ellen B. Mendelson, Daniel Lehrer, Roberta A. Jong, Etta D. Pisano

Affiliations of authors: Magee-Womens Hospital of UPMC (WAB) and Department of Biostatistics (AIB), University of Pittsburgh School of Medicine, Pittsburgh, PA; Department of Radiology, Feinberg School of Medicine, Northwestern University, Chicago, IL (EBM); CERIM, Buenos Aires, Argentina (DL); Department of Medical Imaging, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada (RAJ); Department of Radiology, Medical University of South Carolina, Charleston, SC (EDP).

Correspondence to: Wendie A. Berg, MD, PhD, Department of Radiology, Magee-Womens Hospital, 300 Halket St., Pittsburgh, PA 15213 (e-mail: wendieberg@gmail.com).

Abstract

Background: Mammography is not widely available in all countries, and breast cancer incidence is increasing. We considered performance characteristics using ultrasound (US) instead of mammography to screen for breast cancer.

Methods: Two thousand eight hundred nine participants were enrolled at 20 sites in the United States, Canada, and Argentina in American College of Radiology Imaging 6666. Two thousand six hundred sixty-two participants completed three annual screens (7473 examinations) with US and film-screen (n = 4351) or digital (n = 3122) mammography and had biopsy or 12-month follow-up. Cancer detection, recall, and positive predictive values were determined. All statistical tests were two-sided.

Results: One hundred ten women had 111 breast cancer events: 89 (80.2%) invasive cancers, median size 12 mm. The number of US screens to detect one cancer was 129 (95% bootstrap confidence interval [CI] = 110 to 156), and for mammography 127 (95% CI = 109 to 152). Cancer detection was comparable for each of US and mammography at 58 of 111 (52.3%) vs 59 of 111 (53.2%, $P = .90$), with US-detected cancers more likely invasive (53/58, 91.4%, median size 12 mm, range = 2–40 mm), vs mammography at 41 of 59 (69.5%, median size 13 mm, range = 1–55 mm, $P < .001$). Invasive cancers detected by US were more frequently node-negative, 34 of 53 (64.2%) vs 18 of 41 (43.9%) by mammography ($P = .003$). For 4814 incidence screens (years 2 and 3), US had higher recall and biopsy rates and lower PPV of biopsy (PPV3) than mammography: The recall rate was 10.7% (n = 515) vs 9.4% (n = 453, $P = .03$), the biopsy rate was 5.5% (n = 266) vs 2.0% (n = 97, $P < .001$), and PPV3 was 11.7% (31/266) vs 38.1% (37/97, $P < .001$).

Conclusions: Cancer detection rate with US is comparable with mammography, with a greater proportion of invasive and node-negative cancers among US detections. False positives are more common with US screening.

Worldwide, the number of breast cancer cases is increasing, with 1.4 million new cases globally in 2008 (1), over 1.6 million cases in 2010 (2), and projections of 2.1 million by 2030 (1). Nearly half of this burden is observed in developed countries, many of which have organized screening. Fully 23% of global breast cancer cases are seen in women age 15 to 49 years in developing countries (2). More importantly, even after correcting for the increasing number of cases, deaths from breast cancer are increasing worldwide, with 425 000 deaths in 2010, including 68 000 in women age 49 years or younger in developing countries (2).

While advances in treatment have improved outcomes from breast cancer, axillary lymph node status remains the most important prognostic factor. Clinically detected cancers are larger, with median size of 2.6 cm, compared with those found with screening mammography, with median size of 1.5 cm (3).

Received: February 11, 2015; **Revised:** May 27, 2015; **Accepted:** October 28, 2015

Cancers found clinically are more likely to show axillary nodal metastases: 38% to 45% are node positive compared with 18% to 25% of cancers detected by mammography screening (4,5).

The mortality benefit from mammographic screening is because of identification of node-negative invasive cancers (6). In randomized controlled trials, a 15% reduction in breast cancer mortality has been observed in women age 40 to 49 years at entry, and a 22% reduction in women age 50 to 74 years (7). The reduced benefit from mammography in younger women is because of several factors, including the masking effects of dense parenchyma, which is more common in younger women (8), and also because cancers are more often rapidly growing invasive cancers, which may present clinically in the interval between screens (9).

Mammography is not widely available in developing countries. A test that detects node-negative invasive cancers as well as or better than mammography, with cancer detection not limited by breast density, portable, less expensive, and not using ionizing radiation, could contribute to breast cancer mortality reduction worldwide.

In the prospective international multicenter American College of Radiology Imaging Network (ACRIN) protocol 6666, a statistically significant increase in cancer detection was observed when physician-performed whole-breast screening ultrasound (US) was added to mammography (10,11). In ACRIN 6666, screening US was performed and interpreted independently of mammographic results. The study affords the opportunity to consider performance characteristics of programmatic breast cancer screening using US alone, while also comparing results with those observed with screening mammography in the same participants.

Methods

Participants

Participants were asymptomatic women with heterogeneously or extremely dense breast tissue (12) in at least one quadrant and at least one other risk factor for breast cancer (prioritized as in [11] and detailed in the Supplementary Materials, available online). Participants were at least age 25 years (median = 55 years, range = 25–91 years) at study entry and provided written, informed consent at their initial visit. Two thousand eight hundred nine women were recruited from 20 sites (18 in the United States, one in Buenos Aires, Argentina, and one in Toronto, Ontario, Canada; one other site qualified but did not enroll participants) between April 2004 and February 2006, of whom 2725 were eligible. Two thousand six hundred fifty-nine women completed the initial screen with suitable reference standard of follow-up or biopsy, as did 2493 women in year 2 and 2321 women in year 3 (for a total of 7473 paired screening examinations in 2662 unique participants, as detailed in the Supplementary Materials, available online, and in [11]). Based on self-assigned race/ethnicity, 2467 of 2659 (92.8%) women at first screen were Caucasian, 265 (10%) were Hispanic, 91 (3.4%) were African American, and 90 (3.4%) were Asian, with accrual at each site paralleling local population demographics.

Web-based data capture and quality monitoring were conducted by ACRIN's Biostatistics and Data Management Center (Center for Statistical Sciences, Brown University, Providence, RI, and ACRIN Headquarters, Philadelphia, PA, respectively). The study was Health Insurance Portability and Accountability Act–compliant and received institutional review board approval from all participating sites, ACRIN and National Cancer Institute Cancer Imaging Program approval, and data and safety monitoring committee review every six months.

Screening Methods

Each participant underwent three rounds of mammographic and physician-performed ultrasonographic screening examinations at 0 months (screen 1), 12 months (screen 2), and 24 months (screen 3) in a randomized order assigned prior to initial study imaging. Reference standard was available for 2662 unique participants: 2659 women screened initially, 2493 women at screen 2, and 2321 women at screen 3. At least two-view mammography was performed using either screen-film (n = 4351) or digital (n = 3122) technique. Ultrasound was physician performed using handheld high-resolution linear array broad bandwidth transducers with maximum frequency of at least 12 MHz using standard technique and documentation (13). The radiologist performing and interpreting the screening US and a different radiologist interpreting the study mammogram were not permitted to know the results of the other current screening examination until their interpretations had been recorded, although prior breast imaging (if any) was available together with risk factor and biopsy/surgical history.

Assessments for each breast and lesion were recorded using Breast Imaging–Reporting and Data System (BI-RADS) (12,14,15). An expanded seven-point BI-RADS assessment scale was used at the lesion and breast level: 1, negative; 2, benign; 3, probably benign; 4a, low suspicion; 4b, intermediate suspicion; 4c, moderate suspicion; and 5, highly suggestive of malignancy. Further details of screening methods are included in the Supplementary Materials (available online).

Reference Standard

Reference standard was cancer based on biopsy within 365 days of mammographic screening or no cancer based on a minimum clinical follow-up of one year. Each mammographic and US screen was targeted to occur 365 days after the previous annual screening. A complete examination of all study breasts performed more than 11 full months after the previous screen was considered the next annual screen. A diagnosis of invasive or intraductal breast cancer was considered disease positive. Further details of reference standard are included in the Supplementary Materials (available online).

Statistical Considerations

The primary unit of analysis was the participant (evaluated at an annual screening session). As in the original report of the primary analysis for ACRIN 6666 (10), a participant's BI-RADs score and recommendation were derived as the BI-RADS score from the breast with cancer, or, for participants without cancer, the maximum breast-level BI-RADS score. For a participant with verified cancer diagnosed during the study, the breast with cancer was excluded from analysis for the next annual screen. As per the 5th edition of BI-RADS (16), a recommendation for additional diagnostic testing or biopsy prior to the next screening round (ie, a "recall") was considered test positive, including all BI-RADS assessments of 4a, 4b, 4c, or 5; an assessment of BI-RADS 3 was also considered test positive provided the recommendation was for short-interval (usually 6 months) follow-up, additional imaging, or biopsy.

The sensitivity, specificity, recall rate, positive predictive value of recall (PPV1), and positive predictive values of participant-level biopsies recommended (PPV2) and biopsies performed (PPV3) were estimated. Invasive tumor size was recorded. Results of nodal staging were reported when performed, but nodal staging was not performed in women with a personal history of axillary nodal dissection prior to study entry.

Data were analyzed using SAS v. 9.3 (SAS Institute Inc., Cary, NC). Sensitivity levels and specificity levels estimated for individual years were compared using exact McNemar's test; 95% Wilson confidence intervals (CIs) were provided for individual proportions (proc freq, SAS v. 9.3). Comparisons of proportions estimated from the data spanning multiple years (eg, sensitivity or specificity for year 2 and 3 screenings combined), as well as tests for trend, were performed using generalized estimating equation (GEE) model for binary data (proc genmod, SAS, v. 9.3) accounting for possible correlation between assessments of the same patient. Cluster bootstrap (17) with participants as resampling units, based on 10 000 resamples, was used for estimating 95% CIs for the difference in rates as well as for individual proportions estimated from the data combining multiple years. To evaluate the sensitivity of the primary results to clustering within the 20 centers, we additionally performed cluster bootstrap with sites as resampling units. All presented analyses are exploratory, following the primary analysis comparing the combination of mammography and ultrasound to mammography alone (10,11). The reported *P* values and 95% confidence intervals are two-sided, with .05 threshold used for statistical significance assessments.

Results

Cancer Detection

Across 7473 screens in 2662 unique participants, 110 women were diagnosed with 111 breast cancer events; one participant was diagnosed initially in year 1, then with contralateral cancer in year 3. Of the 111 cancers, 89 (80.2%) were invasive, with median size of 12 mm (range 1 to 55 mm) and 57 of 70 (81%) with nodal staging were node negative.

The number of US screens needed to detect one cancer was 129 (95% CI = 110 to 156), and for mammography 127 (95% CI = 109 to 152). The total number of cancers detected was comparable across modalities, at 58 of 111 (52.3%) for US and 59 of 111 (53.2%) for mammography (*P* = .90, with 95% CI for the difference from -14.4% to 12.6 %) (Table 1). The cancer detection rate with US was 9 per 1000 in year 1 (95% CI = 6.1 to 13.4) and 7.1 per 1000 in years 2 and 3 (95% CI = 5.2 to 9.1); these rates were statistically nonsignificantly different (*P* = .54 and .53) from those of mammography at 7.5 per 1000 in year 1 (95% CI = 4.9 to 11.6) and 8.1 per 1000 in years 2 and 3 (95% CI = 6.2 to 10.2). Of 89 invasive cancers, 53 (59.6%) were detected by US and 41 (46.1%) by mammography (*P* = .11) (Table 1).

There were no statistically significant differences in proportions of detected cancers in participants of different breast density or age groups (Table 2). The rate of cancer detection only by US and not mammography appeared to increase with increasing breast density, at three of 17 (17.6%) cancers in breasts that were visually 26% to 40% dense mammographically, and six of 16 (37.5%) of cancers in breasts that were visually more than 80% dense, though the trend was not statistically significant (*P* = .11) and estimate of density was subjective. For invasive cancers, two of 10, 20.0%, were seen only on US in breasts that were visually 26% to 40% dense and six of 12, 50.0%, were seen only on US in breasts that were visually more than 80% dense (P_{trend} = .06) (Table 3).

Despite close similarity in total number of detected cancers, the vast majority of cancers detected by US were invasive (53/58, 91.4%, median size = 12 mm, range = 2–40 mm), compared with 41 of 59 (69.5%, median size = 13 mm, range = 1–55 mm) by mammography (bootstrap *P* < .001), and invasive cancers depicted by US were statistically significantly more frequently node negative (34/53, 64.2%) compared with those seen on mammography (18/41, 43.9%, bootstrap *P* = .003). The differences remained statistically significant in the cluster bootstrap with sites as resampling units. Cancers seen only on screening US were all stage IIA or lower (Table 4). Among 89 invasive cancers, 30 (33.7%) were seen only with US and 18 (20.2%) only with mammography.

DCIS was much more likely to be detected by mammography, with 18 of 22 (81.8%) seen on mammography compared with only five of 22 (22.7%) by US (*P* = .002, Table 1). Two of 22 (9.1%) DCIS were seen only on US (one high nuclear grade and one intermediate nuclear grade).

False Positives

For 2659 first study screens with reference standard, of which over 98% were incidence screens for mammography (ie, a prior screening mammogram was available) and just over 11% were incidence screens for ultrasound (10), US prompted recall of more women than mammography (555, 20.9%, vs 306, 11.5%, *P* < .001) (Table 1). When comparing incidence screens in years 2 and 3 for both modalities, the recall rate of US at 515 of 4814 (10.7%) was comparable with, although still slightly higher than, mammography, at 453 of 4814 (9.4%, *P* = .03). Of 7362 screens in women without cancer, 1012 were recommended for further testing prior to the next screening US (overall specificity = 86.3%). Overall, 810 of 2552 (31.7%, 95% CI = 30.0% to 33.6%) unique women without cancer were recalled at least once during the three screening rounds because of US compared with 591 of 2552 (23.2%, 95% CI = 21.6% to 24.8%) prompted by mammography (*P* < .001) (Table 5). When results of mammography and US were integrated, 294 recalls were avoided: 1107 of 2552 (43.2%) participants without cancer were actually recalled.

At the per-screen level, the overall specificity of US was lower than that of mammography: 6350 of 7362 (86.3%) vs 6662 of 7362 (90.5%, *P* < .001) (Table 1). The difference remained highly statistically significant in the cluster bootstrap analysis with sites as resampling units. The age-based and breast density–based distributions of the false-positive results for individual screens are summarized in Table 6. False-positive recalls from US decreased with increasing patient age (*P* = .002) (Table 6), a tendency observed for mammography as well but not statistically significant (*P* = .13). More false positives were seen with US with increasing breast density (*P* = .03) (Table 6), but no consistent trend was observed for mammography (*P* = .25). The greatest increases in false positives with US compared with mammography (*P* < .001) were in women age 40 to 69 years and for women with density visually estimated from mammograms at greater than 40%.

Likelihood of cancer was lower with US than mammography for each of recall (PPV1), biopsies recommended (PPV2), and biopsies performed (PPV3) (Table 1). For incidence screening (years 2 and 3), short-term follow-up rates were previously reported (11) as 3.9% (190/4814) for US vs 1.6% for mammography (76/4814, and this difference was statistically significant, *P* < .001); biopsy rate was 5.5% (266/4814) vs 2.0% (97/4814, *P* < .001), and PPV3 of biopsies performed was 11.7% (31/266) vs 38% (37/97) (*P* < .001).

Table 1. Breast cancer detection by ultrasound or mammography in 2662 participants screened for three annual rounds (7473 screens)*

Screening performance characteristic	US alone		Mammography alone		Differences US vs mammography		US but not mammography	
	No./total exams	Rate (95% CI†), %	No./total exams	Rate (95% CI†), %	Diff (95% CI†)	P‡	No./total exams	Rate (95% CI†), %
Cancer detection Rate per 1000 screens								
Year 1	24/2659	9.0 (6.1 to 13.4)	20/2659	7.5 (4.9 to 11.6)	1.5 (-4.8 to 1.9)	.54	14/2659	5.3 (3.1 to 8.8)
Years 2&3	34/4814	7.1 (5.2 to 9.1)	39/4814	8.1 (6.2 to 10.2)	-1.0 (-3.6 to 1.5)	.53	18/4814	3.7 (2.3 to 5.4)
Sensitivity								
Overall	58/111	52.3 (43.2 to 61.3)	59/111	53.2 (44.1 to 62.2)	-0.90 (-14.4 to 12.6)	.90	32/111	28.8 (20.7 to 36.9)
Invasive cancers	53/89	59.6 (49.2 to 69.2)	41/89	46.1 (36.1 to 56.4)	13.5 (-2.2 to 28.1)	.11	30/89	33.7 (24.7 to 44.0)
DCIS	5/22	22.7 (10.1 to 43.4)	18/22	81.8 (61.5 to 92.7)	-59.1 (-86.4 to -31.8)	.002	2/22	9.1 (2.5 to 27.8)
Year 1	24/36	66.7 (50.3 to 79.8)	20/36	55.6 (39.6 to 70.5)	11.1 (-15.0 to 36.4)	.54	14/36	38.9 (24.8 to 55.1)
Years 2&3	34/75	45.3 (34.2 to 56.4)	39/75	52.0 (41.0 to 63.0)	-6.7 (-23.0 to 9.7)	.53	18/75	24.0 (14.6 to 33.8)
Specificity								
Overall	6350/7362	86.3 (86.1 to 87.8)	6662/7362	90.5 (90.5 to 91.9)	-4.2 (-5.3 to -3.2)	<.001	552/7362	7.5 (6.9 to 8.2)
Year 1	2092/2623	79.8 (78.2 to 81.2)	2337/2623	89.1 (87.8 to 90.2)	-9.3 (-11.3 to -7.5)	<.001	207/2623	7.9 (6.9 to 9.0)
Years 2&3	4258/4739	89.9 (89.8 to 91.6)	4325/4739	91.3 (91.2 to 92.9)	-1.4 (-2.6 to -0.1)	.02	345/4739	7.3 (6.5 to 8.1)
Recall rate								
Year 1	555/2659	20.9 (19.4 to 22.5)	306/2659	11.5 (10.4 to 12.8)	9.4 (7.5 to 11.2)	<.001	466/2659	17.5 (16.1 to 19.0)
Years 2&3	515/4814	10.7 (9.8 to 11.6)	453/4814	9.4 (8.6 to 10.3)	1.3 (0.1 to 2.5)	.03	430/4814	8.9 (8.1 to 9.8)
PPV1								
Year 1	24/555	4.3 (2.9 to 6.4)	20/306	6.5 (4.3 to 9.9)	-2.2 (-4.7 to 0.2)	.06	14/466	3.0 (1.8 to 5.0)
Years 2&3	34/515	6.6 (4.9 to 8.5)	39/453	8.6 (6.6 to 10.8)	-2.0 (-4.6 to 0.5)	.12	18/430	4.2 (2.5 to 6.0)
PPV2								
Year 1	22/257	8.6 (5.7 to 12.6)	19/83	22.9 (15.2 to 33.0)	-14.3 (-22.4 to -6.6)	<.001	12/215	5.6 (3.2 to 9.5)
Years 2&3	33/302	10.9 (8.0 to 14.0)	38/126	30.2 (23.4 to 37.5)	-19.2 (-26.5 to -12.4)	<.001	17/246	6.9 (4.1 to 9.9)
Biopsy rate, %								
Year 1	233/2659	8.8 (7.7 to 9.9)	65/2659	2.4 (1.9 to 3.1)	6.3 (5.2 to 7.5)	<.001	208/2659	7.8 (6.9 to 8.9)
Years 2&3	266/4814	5.5 (4.9 to 6.2)	97/4814	2.0 (1.7 to 2.4)	3.5 (2.8 to 4.2)	<.001	239/4814	5.0 (4.3 to 5.6)
PPV3§								
Year 1	21/233	9.0 (6.0 to 13.4)	19/65	29.2 (19.6 to 41.2)	-20.2 (-30.2 to 10.7)	<.001	12/208	5.8 (3.3 to 9.8)
Years 2&3	31/266	11.7 (8.4 to 15.1)	37/97	38.1 (29.7 to 47.3)	-26.5 (-35.8 to -17.7)	<.001	18/239	7.5 (4.5 to 10.7)

* Some of this information can be found in Table 3 of Berg et al. (11), but this table represents a reanalysis of the data. CI = confidence interval; DCIS = ductal carcinoma in situ; PPV1 = positive predictive value of recall; PPV2 = positive predictive value of participant-level biopsies recommended; PPV3 = positive predictive value of biopsies actually performed; US = ultrasound.

† Patient-level 95% bootstrap CIs (based on 10 000 resamples) for proportions computed from data combining multiple years. Wilson 95% CIs are reported for simple proportions (proc freq SAS v. 9.3, Cary, NC). (The cluster bootstrap CIs after accounting for the possible correlation within sites differ only fractionally.)

‡ P value for two-sided McNemar's test for comparison of simple correlated proportions (proc freq SAS v. 9.3); P value from the two-sided Wald test for the imaging modality (US/mammography) coefficient of generalized estimating equation models for comparison of proportions over multiple years (proc genmod, SAS v. 9.3).

§ PPV3 = rate of malignancies among biopsies actually performed.

Table 2. Breast cancer detection by ultrasound or mammography for categories of visually estimated breast density and participant age at time of screening across three annual screening rounds

	Screens with cancer		US sensitivity		Mammography sensitivity		Difference in sensitivity of US vs mammography		US but not mammography detections	
Screen characteristic	No. cancers/ No. screens	(Incidence, %)	No. detected/ No. Cancers	(%)	No. detected/ No. Cancers	(%)	Estimate	P*	No. detected/ No. cancers	(%)
Density, %										
≤25	1/128	(0.8)	0/1	(0.0)	0/1	(0.0)	0.0	1.00	0/1	(0.0)
26–40	17/710	(2.4)	9/17	(52.9)	11/17	(64.7)	-11.8	.73	3/17	(17.6)
41–60	36/2390	(1.5)	18/36	(50.0)	21/36	(58.3)	-8.3	.66	9/36	(25.0)
61–80	41/2890	(1.4)	22/41	(53.7)	18/41	(43.9)	9.8	.54	14/41	(34.1)
>80	16/1352	(1.2)	9/16	(56.3)	9/16	(56.3)	0.0	1.00	6/16	(37.5)
P_{trend}†	---	---	---	.65	---	.38	.39	---	---	.11‡
Unknown	0/3	(0)	0/0	(NA)	0/0	(NA)	NA	NA	0/0	(NA)
Age, y										
<40	2/289	(0.7)	1/2	(50.0)	2/2	(100)	-50.0	1.00	0/2	(0.0)
40–49	16/1538	(1.0)	8/16	(50.0)	7/16	(43.8)	6.3	1.00	6/16	(37.5)
50–69	79/4916	(1.6)	39/79	(49.4)	42/79	(53.2)	-3.8	.76	20/79	(25.3)
>69	14/730	(1.9)	10/14	(71.4)	8/14	(57.1)	14.3	.75	6/14	(42.9)
P_{trend}†	---	---	---	.27	---	.69	.68	---	---	.68‡

* Two-sided Exact McNemar's test. NA = not applicable; US = ultrasound.
† Using two-sided Wald test for the factor's coefficient of the generalized estimating equation model accounting for possible correlation between assessments of the same patients (proc genmod, SAS, v. 9.3, Cary, NC). The test for trend was performed for the two lowest categories grouped together; conclusions remain the same with for the test for trend with presented categories.
‡ Care must be taken in interpreting P values for "US but not mammography detections" because of post hoc nature of the analyses and sparse data.

Table 3. Invasive breast cancer detection by ultrasound or mammography for categories of visually estimated breast density and participant age at time of screening across three annual screening rounds

	Screens with cancer		US		Mammography		Difference in US vs mammography		US, but not mammography detections	
Screen characteristic	No. cancers/ No. screens	(Incidence, %)	No. detected/ No. cancers	(%)	No. detected/ No. cancers	(%)	Estimate, %	P*	No. detected/ No. cancers	(%)
Density, %										
≤25	1/128	(0.8)	0/1	(0.0)	0/1	(0)	0.0	1.00	0/1	(0.0)
26–40	10/710	(1.4)	6/10	(60.0)	6/10	(60)	0.0	1.00	2/10	(20.0)
41–60	30/2390	(1.3)	16/30	(53.3)	17/30	(56.7)	-3.3	1.00	8/30	(26.7)
61–80	36/2890	(1.2)	22/36	(61.1)	13/36	(36.1)	25.0	0.06	14/36	(38.9)
>80	12/1352	(0.9)	9/12	(75.0)	5/12	(41.7)	33.3	0.29	6/12	(50.0)
P_{trend}†	---	---	.23	---	.19	---	.08		.06‡	---
Unknown	0/3	(0)	0/0	(NA)	0/0	(NA)	NA		0/0	(NA)
Age, y										
<40	1/289	(0.3)	1/1	(100.0)	1/1	(100.0)	0.0	1.00	0/1	(0.0)
40–49	14/1538	(0.9)	8/14	(57.1)	5/14	(35.7)	21.4	0.51	6/14	(42.9)
50–69	61/4916	(1.2)	34/61	(55.7)	28/61	(45.9)	9.8	0.36	18/61	(29.5)
>69	13/730	(1.8)	10/13	(76.9)	7/13	(53.8)	23.1	0.51	6/13	(46.2)
P_{trend}†	---	---	.38	---	.47	---	.94		.80‡	---

* Two-sided exact McNemar's test. NA = not applicable; US = ultrasound.
† Using two-sided Wald test for the factor's coefficient of the generalized estimating equation model accounting for possible correlation between assessments of the same patients (proc genmod, SAS, v. 9.3, Cary, NC). The test for trend was performed for the two lowest categories grouped together; conclusions remain the same with for the test for trend with presented categories.
‡ Care must be taken in interpreting P values for "US but not mammography detections" because of the post hoc nature of the analyses and sparse data.

Discussion

Many developing countries lack any screening for breast cancer. US is an important test for evaluating palpable breast lumps as it affords direct correlation of clinical and imaging findings and its use has begun in developing countries (18,19). Even low-cost (approximately $15 000), portable US systems are now equipped with high-resolution linear transducers (12 MHz or higher) and are effective at distinguishing simple cysts from suspicious masses (20). The equipment used in this study, between 2004 and 2008, is comparable with what is now available on low-cost devices. We found that, despite a higher rate of false positives,

screening US depicted a similar number of cancers as did mammography but with statistically significantly higher proportions of invasive and node-negative invasive cancers.

We are not the first to suggest that US could replace mammography in some women, though prior reports are limited to women with symptoms. In 3129 symptomatic women in Thailand, US showed an area under the curve of 0.962, which was better than mammography at 0.954 (P = .015), and adding mammography to US produced statistically insignificant improvement (21). In 1208 focally symptomatic women age 30 to 39 years, Lehman et al. (22) found higher sensitivity for US than mammography among the 23 (1.9%) women with cancer, with 22 of 23 (95.6%) cancers seen by US and only 14 of 23 (60.9%) with mammography (P = .0098), albeit with a higher false-positive rate for US. They suggest that mammography may have been unnecessary in such women, with only one second malignancy seen only on mammography (22). Mistry et al. (23) reported that no cancers would have been missed in women age 35 to 39 years if United Kingdom best practice guidelines recommending that only US be performed in symptomatic women under age 40 years had been followed. Houssami et al. (24) found higher sensitivity of US than mammography among symptomatic women age 45 years and younger at the same specificity. In ACRIN 6666, we found no difference in cancer detection rates by US or mammography in categories of age or breast density.

One barrier to implementing any screening program is the harm of false positives. As has been observed in some (25,26), but not all (27,28), prior studies of mammography, we found false positives more common in younger women on US but not mammography. In our study, with increasing breast density, false positives increased for US but not mammography, although increasing false positives have been observed in prior studies of mammography (27,29). Availability of prior comparison examinations reduces false-positive recalls for all breast imaging modalities to date (30–32). In this study, recall rates decreased from 20.9% for the first screening US to 10.7% in years 2 and 3. Weigert recently reported (33) that by year 3 technologist-performed screening US across multiple practices in Connecticut prompted false-positive recalls for only 7.7% of women in year 3, compared with 13.8% in year 1 (34). In a separate analysis from ACRIN 6666 (35), we showed that probably benign masses seen only on US, assessed as BI-RADS 3, can be followed at one year (obviating initial 6-month follow-up or biopsy), which could greatly reduce additional testing prompted by screening US. Similarly, we found no malignancies among multiple bilateral circumscribed benign-appearing masses identified on screening US and now recommend BI-RADS 2 assessment (36).

There are several limitations to our analysis. We observed more invasive cancers detected by US than by mammography; however, a larger study is needed to statistically support greater sensitivity of US to invasive cancers. Only 41.8% (3122/7473) of mammograms in this study were performed with digital technique, which may slightly underestimate cancer detection by mammography (25), particularly because our study was enriched in women with dense breasts. Importantly, we reported no difference in supplemental yield of US after digital vs film screen mammography (11). All of our study participants had at least one risk factor in addition to breast density; cancer detection rates are expected to be lower in lower-prevalence populations, and biopsy rates may have been artificially high in this population because of both patient and radiologist concerns in this elevated-risk

Table 4. Stage distribution of 111 breast cancer events in 2662 women screened with ultrasound and mammography for three years by method of cancer detection

Stage*	US only	Mammography only	Both mammography and US	MRI†	Clinically detected‡
0	2	15	3	1	0
I	25	11	9	7	5
IIA	5	3	10	1	1
IIB	0	1	1	0	0
IIIA	0	1	0	0	0
IIIB	0	0	2	0	3
IIIC	0	1	0	0	0
IV	0	1	1	0	0

* According to American Joint Committee on Cancer 7th edition (46). US = ultrasound.

† Among a subset of 612 women who had a single screening MRI examination after the third round of screening mammography and US, nine women were diagnosed with cancer seen only on MRI, including one woman diagnosed on MRI in year 3 after an initial contralateral diagnosis by mammography only in year 1.

‡ There were two other cancers detected that were not seen on study imaging or clinically: One woman was diagnosed with ductal carcinoma in situ (stage 0) because of computer-assisted detection applied to mammography after study results were recorded, and another woman with a pathogenic BRCA1 mutation was found to have a 7 mm grade 3 invasive ductal carcinoma (stage I) on off-study MRI six months after the third screening round.

Table 5. Cumulative unique participants recalled or biopsied because of ultrasound or mammography for 2662 women during the three-year period

Performance characteristic	US		Mammography		Absolute percent difference US vs mammography	
	No./total participants	Rate (95% CI*)	No./total participants	Rate (95% CI*)	Estimate	P†
Overall recall rate	877/2662	32.9 (31.2 to 34.7)	657/2662	24.7 (23.1 to 26.3)	8.26	<.001
Cancer patients recalls	58/110‡	52.7 (43.5 to 61.8)	59/110	53.6 (44.3 to 62.7)	-0.91	1.00
Cancer patients recalls for wrong reason§	9/110	8.2 (4.4 to 14.8)	7/110	6.4 (3.1 to 12.6)	1.82	.79
Noncancer patients recalls	810/2552	31.7 (30.0 to 33.6)	591/2552	23.2 (21.6 to 24.8)	8.58	<.001
Overall biopsy rate	447/2662	16.8 (15.4 to 18.3)	157/2662	5.9 (5.1 to 6.9)	10.89	<.001
Noncancer patients biopsy (at least one)	390/2552	15.3 (13.9 to 16.7)	100/2552	3.9 (3.2 to 4.74)	11.36	<.001

* 95% Wilson confidence limits for simple proportions. CI = confidence interval; US = ultrasound.

† Two-sided exact McNemar's test.

‡ One hundred ten women were diagnosed with 111 cancer events (one woman diagnosed in year 1 was diagnosed with contralateral cancer in year 3).

§ Women recalled prior to the appearance of the confirmed cancer or because of finding in a cancer-free location.

ARTICLE

Table 6. False positives* by ultrasound or mammography as a function of visually estimated breast density or participant age

	Screens without cancer		US		Mammography		Difference in US vs mammography		US, but not mammography false positives	
Screen characteristic	No. noncancers/ No. Screens	Prevalence (%)	No. recalls/ No. Noncancers	(%)	No. recalls/ No. Noncancers	(%)	Estimate (%)	P†	No. recalls/ No. Noncancers	(%)
Density, %										
≤25	127/128	(99.2)	13/127	(10.2)	13/127	(10.2)	0.0	1.00	9/127	(7.1)
26–40	693/710	(97.6)	84/693	(12.1)	72/693	(10.4)	1.7	0.31	70/693	(10.1)
41–60	2354/2390	(98.5)	301/2354	(12.8)	200/2354	(8.5)	4.3	<0.001	268/2354	(11.4)
61–80	2849/2890	(98.6)	422/2849	(14.8)	264/2849	(9.3)	5.5	<0.001	361/2849	(12.7)
>80	1336/1352	(98.8)	192/1336	(14.4)	151/1336	(11.3)	3.1	0.02	156/1336	(11.7)
P_{trend}‡	---	---	---	.03	---	.25	.16	---	---	.09§
Unknown	3/3	(100.0)	0/3	(0.0)	0/3	(0.0)	0.0	1.00	0/3	(0.0)
Age, y										
<40	287/289	(99.3)	43/287	(15.0)	33/287	(11.5)	3.5	0.28	36/287	(12.5)
40–49	1522/1538	(99.0)	245/1522	(16.1)	158/1522	(10.4)	5.7	<0.001	203/1522	(13.3)
50–69	4837/4916	(98.4)	643/4837	(13.3)	443/4837	(9.2)	4.1	<0.001	554/4837	(11.5)
>69	716/730	(98.1)	81/716	(11.3)	66/716	(9.2)	2.1	0.21	71/716	(9.9)
P_{trend}‡	---	---	---	.002	---	.13	.48	---	---	.02§

* Positive test was defined as Breast Imaging–Reporting and Data System 3 or higher; false positive had no diagnosis of cancer within 365 days of the screening exam. US = ultrasound.

† Using the two-sided Wald test for the differences between US and mammography in each group (of density or age) within the GEE model accounting for correlation between examinations of the same patients (proc genmod, SAS v. 9.3).

‡ Using the two-sided Wald test for the factor coefficient of the generalized estimating equation (GEE) model accounting for correlation between examinations of the same patients (proc genmod, SAS v. 9.3, Cary, NC).

§ Care must be taken in interpreting *P* values for "US but not mammography false positives" because of the post hoc nature of the analyses.

population. In an average-risk population using an automated arm for screening US, a 3.6 per 1000 cancer detection rate was maintained, but only 3% of women were recommended for biopsy and 31% of biopsies showed cancer (37). Results from additional screening US series are discussed in the Supplementary Materials (available online). Most participants in this study were Caucasian, and 94% had breasts less than 4 cm thick (10). High-frequency US image quality degrades with deep lesions (>3 cm), and our results would not be generalizable to women with very large breasts. All ACRIN 6666 radiologist investigators had interpreted at least 500 breast US examinations in the preceding two years and successfully completed phantom scanning (38), training in BI-RADS:US (39), and interpretive skills tasks (40). Using the same scanning and documentation approach, results with technologist-performed prevalence screening US to date show slightly lower cancer detection rates, PPV1, and PPV3, as summarized by Berg and Mendelson (13), possibly reflecting lower cancer prevalence in the populations screened. Importantly, Tohno et al. (41) reported that technologists in Japan performed better than physicians in detecting cancer during a two-day training course for handheld US screening. Training would be necessary for any facility planning to offer screening US (42), also true for developing countries. With appropriate training, US is no more operator dependent than interpreting mammography (43,44). Finally, while we had previously shown that invasive lobular cancer and low-grade invasive ductal carcinoma are overrepresented among cancers seen only on US (11), we do not have detailed molecular subtype results for the cancers in this study.

In summary, cancer detection by US was shown to be very similar to mammography, and the vast majority of cancers seen with US were invasive and node negative. While the false-positive rate of US exceeds that of mammography, the number of women recalled for extra testing becomes comparable on incidence screening rounds. Although further validation is warranted, these results suggest that screening US could be a viable alternative to mammography in countries lacking organized screening, particularly with availability of low-cost, portable US systems. Where mammography is available, US should be seen as a supplemental test for women with dense breasts who do not meet high-risk criteria for screening MRI and for high-risk women with dense breasts who are unable to tolerate MRI (45).

Funding

Funded by The Avon Foundation for Women and the National Cancer Institute at the National Institutes of Health, grants U01CA80098, U01CA79778, and R01CA187593.

Notes

Presented at the 2012 Radiological Society of North America Scientific Assembly.

Trial registration: Clinicaltrials.gov identifier NCT 00072501.

The study sponsors were not involved in the design of the study; the collection, analysis, or interpretation of data; the writing of the manuscript; or decision to submit the manuscript for publication.

We are most grateful to the 100 physician investigators; the many research assistants, ACRIN support staff, Jean Cormack (formerly at the Brown Center for Statistical Sciences), and especially to the participants for making this study possible. We also thank Marc Hurlbert (formerly of The Avon Foundation for Women) for believing in this work and helping to make it happen. Jeremy Berg deserves special mention at many levels, including support with statistical analysis.

References

1. Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008-2030): a population-based study. *Lancet Oncol*. 2012;13:790–801.

2. Forouzanfar MH, Foreman KJ, Delossantos AM, et al. Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet*. 2011;378(9801):1461–1484.
3. Mathis KL, Hoskin TL, Boughey JC, et al. Palpable presentation of breast cancer persists in the era of screening mammography. *J Am Coll Surg*. 2010;210(3):314–318.
4. Vanier A, Leux C, Allioux C, et al. Are prognostic factors more favorable for breast cancer detected by organized screening than by opportunistic screening or clinical diagnosis? A study in Loire-Atlantique (France). *Cancer Epidemiol*. 2013;37(5):683–687.
5. Tabar L, Vitak B, Chen HH, et al. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am*. 2000;38(4):625–651.
6. Smith RA, Duffy SW, Gabe R, et al. The randomized trials of breast cancer screening: what have we learned? *Radiol Clin North Am*. 2004;42(5):793–806, v.
7. Nelson HD, Tyne K, Naik A, et al. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2009;151(10):727–737, W237-W242.
8. Kerlikowske K, Ichikawa L, Miglioretti DL, et al. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk. *J Natl Cancer Inst*. 2007;99(5):386–395.
9. Buist DS, Porter PL, Lehman C, et al. Factors contributing to mammography failure in women aged 40–49 years. *J Natl Cancer Inst*. 2004;96(19):1432–1440.
10. Berg WA, Blume JD, Cormack JB, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA*. 2008;299(18):2151–2163.
11. Berg WA, Zhang Z, Lehrer D, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*. 2012;307(13):1394–1404.
12. D'Orsi CJ, Bassett LW, Berg WA, et al. *Breast Imaging Reporting and Data System, BI-RADS: Mammography*, 4th *edition*. Reston: American College of Radiology; 2003.
13. Berg WA, Mendelson EB. Technologist-performed handheld screening breast US imaging: how is it performed and what are the outcomes to date? *Radiology*. 2014;272(1):12–27.
14. Mendelson EB, Baum JK, Berg WA, et al. *Breast Imaging Reporting and Data System, BI-RADS: Ultrasound*, 1st *Ed*. Reston: American College of Radiology; 2003.
15. Ikeda DM, Hylton NM, Kuhl CK, et al. *Breast Imaging Reporting and Data System, BI-RADS: Magnetic Resonance Imaging*. Reston: American College of Radiology; 2003.
16. Sickles EA, D'Orsi CJ. Follow-up and outcome monitoring. In: *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology; 2013.
17. Field CA, Welch AH. Bootstrapping clustered data. *JR Statist Soc B*. 2007;69(3):369–390.
18. Gonzaga MA. How accurate is ultrasound in evaluating palpable breast masses? *Pan Afr Med J*. 2010;7:1.
19. Irurhe NK, Adekola OO, Awosanya GO, et al. The accuracy of ultrasonography in the diagnosis of breast pathology in symptomatic women. *Nig Q J Hosp Med*. 2012;22(4):236–239.
20. Hilton SV, Leopold GR, Olson LK, et al. Real-time breast sonography: application in 300 consecutive patients. *AJR Am J Roentgenol*. 1986;147(3):479–486.
21. Chairat R, Puttisri A, Pamarapa A, et al. Are both ultrasonography and mammography necessary for cancer investigation of breast lumps in resource-limited countries? *ISRN Oncol*. 2013;2013:257942.
22. Lehman CD, Lee CI, Loving VA, et al. Accuracy and value of breast ultrasound for primary imaging evaluation of symptomatic women 30–39 years of age. *AJR Am J Roentgenol*. 2012;199(5):1169–1177.
23. Mistry SG, Barnes N, Ooi J. Will adherence to new guidance lead to missed cancer diagnoses? Evaluation of limiting symptomatic mammograms to over forties. *Breast J*. 2013;19(2):142–148.
24. Houssami N, Irwig L, Simpson JM, et al. Sydney Breast Imaging Accuracy Study: Comparative sensitivity and specificity of mammography and sonography in young women with symptoms. *AJR Am J Roentgenol*. 2003;180(4):935–940.
25. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005;353(17):1773–1783.
26. Pace LE, Keating NL. A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*. 2014;311(13):1327–1335.
27. Lehman CD, White E, Peacock S, et al. Effect of age and breast density on screening mammograms with false-positive findings. *AJR Am J Roentgenol*. 1999;173(6):1651–1655.
28. Johns LE, Moss SM. False-positive results in the randomized controlled trial of mammographic screening from age 40 ("Age" trial). *Cancer Epidemiol Biomarkers Prev*. 2010;19(11):2758–2764.
29. Kerlikowske K, Zhu W, Hubbard RA, et al. Outcomes of screening mammography by frequency, breast density, and postmenopausal hormone therapy. *JAMA Intern Med*. 2013;173(9):807–816.
30. Schell MJ, Yankaskas BC, Ballard-Barbash R, et al. Evidence-based target recall rates for screening mammography. *Radiology*. 2007;243(3):681–689.
31. Kriege M, Brekelmans CT, Boetes C, et al. Differences between first and subsequent rounds of the MRISC breast cancer screening program for women with a familial or genetic predisposition. *Cancer*. 2006;106(11):2318–2326.
32. Warner E, Plewes DB, Hill KA, et al. Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. *JAMA*. 2004;292(11):1317–1325.
33. Weigert JM. The Connecticut experiment continues: Ultrasound in the screening of women with dense breasts years 3 and 4. In: *Radiologic Society of North America*. Chicago, IL; 2014.
34. Weigert J, Steenbergen S. The Connecticut experiment: the role of ultrasound in the screening of women with dense breasts. *Breast J*. 2012;18(6):517–522.
35. Barr RG, Zhang Z, Cormack JB, et al. Probably Benign Lesions at Screening Breast US in a Population with Elevated Risk: Prevalence and Rate of Malignancy in the ACRIN 6666 Trial. *Radiology*. 2013;269(3):701–712.
36. Berg WA, Zhang Z, Cormack JB, et al. Multiple bilateral circumscribed masses at screening breast US: Consider annual follow-up. *Radiology*. 2013;268(3):673–683.
37. Kelly KM, Dean J, Comulada WS, et al. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *Eur Radiol*. 2010;20(3):734–742.
38. Berg WA, Blume JD, Cormack JB, et al. Lesion detection and characterization in a breast US phantom: results of the ACRIN 6666 Investigators. *Radiology*. 2006;239(3):693–702.
39. Mendelson EB, Böhm-Vélez M, Berg WA, et al. ACR BI-RADS Ultrasound. In. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology; 2013.
40. Berg WA, Blume JD, Cormack JB, et al. Training the ACRIN 6666 Investigators and Effects of Feedback on Breast Ultrasound Interpretive Performance and Agreement in BI-RADS Ultrasound Feature Analysis. *AJR Am J Roentgenol*. 2012;199(1):224–235.
41. Tohno E, Takahashi H, Tamada T, et al. Educational program and testing using images for the standardization of breast cancer screening by ultrasonography. *Breast Cancer*. 2012;19(2):138–146.
42. Mendelson EB, Berg WA. Training and standards for performance, interpretation, and structured reporting for supplemental breast cancer screening. *AJR Am J Roentgenol*. 2015;204(2):265–268.
43. Berg WA, Blume JD, Cormack JB, et al. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. *Radiology*. 2006;241(2):355–365.
44. Bosch AM, Kessels AG, Beets GL, et al. Interexamination variation of whole breast ultrasound. *Br J Radiol*. 2003;76(905):328–331.
45. Berg WA. Tailored supplemental screening for breast cancer: what now and what next? *AJR Am J Roentgenol*. 2009;192(2):390–399.
46. Edge SB, Byrd DR, Compton CC, et al. *AJCC Cancer Staging Handbook*, 7th *ed*. New York: Springer; 2011.